Classes are not Clusters: Improving Label-based Evaluation of Dimensionality Reduction

Hyeon Jeon, Yun-Hsin Kuo, Michaël Aupetit, Kwan-Liu Ma, and Jinwook Seo

Abstract— A common way to evaluate the reliability of dimensionality reduction (DR) embeddings is to quantify how well labeled classes form compact, mutually separated clusters in the embeddings. This approach is based on the assumption that the classes stay as clear clusters in the original high-dimensional space. However, in reality, this assumption can be violated; a single class can be fragmented into multiple separated clusters, and multiple classes can be merged into a single cluster. We thus cannot always assure the credibility of the evaluation using class labels. In this paper, we introduce two novel quality measures—*Label-Trustworthiness* and *Label-Continuity* (Label-T&C)—advancing the process of DR evaluation based on class labels. Instead of assuming that classes are well-clustered in the original space, Label-T&C work by (1) estimating the extent to which classes form clusters in the original and embedded spaces and (2) evaluating the difference between the two. A quantitative evaluation showed that Label-T&C outperform widely used DR evaluation measures (e.g., Trustworthiness and Continuity, Kullback-Leibler divergence) in terms of the accuracy in assessing how well DR embeddings preserve the cluster structure, and are also scalable. Moreover, we present case studies demonstrating that Label-T&C can be successfully used for revealing the intrinsic characteristics of DR techniques and their hyperparameters.

Index Terms—Dimensionality Reduction, Reliability, Clustering, Clustering Validation Measures, Dimensionality Reduction Evaluation

1 Introduction

Dimensionality reduction (DR) is one of the most widely used tools in conducting the visual cluster analysis of high-dimensional data [27, 34, 52–54, 70]. Using DR for cluster analysis relies on the assumption that the cluster structure of the original data is accurately represented in the low-dimensional DR embeddings. However, DR inherently generates distortions, i.e., the original cluster structure is imprecisely represented in the embeddings [2, 7, 26, 40, 41]. As distortions can make visual cluster analysis performed with DR unreliable [27, 29], it is important to evaluate how well the original cluster structure is preserved in the DR embeddings [29, 32, 44, 70], prior to the analysis. There exist ways to evaluate the reliability of cluster structures in DR embeddings, in either a perceptual [20, 57, 70] or computational [29, 37, 47, 62] manner.

A general process to evaluate the preservation of cluster structure in DR embeddings is to utilize class labels. This is done by assessing *cluster-label matching* (CLM), that is, the extent to which class labels form clusters in the embedded space [8, 22, 32, 69, 71, 73]. CLM is mostly evaluated by using *clustering validation measures* (CVMs) [42, 68], such as the Silhouette Coefficient [55]. CVMs inform how well the groups in the given label-based data partition form clear position-based clusters. The partitions that contain mutually separated and individually condensed groups are preferred. For the label-based evaluation of DR, data embeddings and class labels are used as an input dataset and partition, respectively. Embeddings with good CLM are considered to have good quality, assuming that the original data also have good CLM.

However, such an assumption can hardly be guaranteed [3, 23, 28, 70]. There is no constraint on labels' sources. Labels can come from an external source (e.g., human annotation), possibly unrelated to the features of the data space. Labels can also result from clustering techniques, which may not align with the actual clusters. Therefore, we do not know how well labels make up the clusters in the original data; a single class can consist of multiple separated clusters, and multiple classes can be in close proximity or even overlapped [3] in a single cluster. These possibilities cast doubt on the conclusions derived from the general process of label-based DR evaluation. For instance, an embedding that accurately represents overlapping classes in the original space might be considered to have low quality as it has bad CLM.

In this work, we revisit the process of evaluating DR using class labels. We introduce two measures—*Label-Trustworthiness* (Label-T) and *Label-Continuity* (Label-C)—which examine CLM in an alternative way to assess the reliability of cluster structures in DR embeddings. In contrast to the general label-based evaluation process, Label-T&C use CVM to quantify CLM *distortions* as the difference between CLM estimated in both original and embedded spaces. Label-T quantifies the distortion due to the degradation of CLM: the score is lower when the points of two different classes get closer in the embedding than in the original space. Conversely, Label-C evaluates the distortion regarding the exaggeration of CLM: the score is lower when the points of two different classes get farther apart in the embedding than in the original space. The rationale behind our measures is that in visual cluster analysis, it is important to investigate how class labels span the original cluster structure as seen through the embedding [3–5, 12, 67] (e.g., examine the individual density of a class or the pairwise proximity between classes). Since CLM distortions reduce the reliability of cluster structures represented by the embeddings, Label-T&C scores can be interpreted as proxies for the credibility of DR-based cluster analysis.

We conduct a series of quantitative experiments to validate the effectiveness of Label-T&C. The results show that Label-T&C can better capture the distortions of cluster structures than the existing measures (e.g., Steadiness & Cohesiveness [29] and Trustworthiness & Continuity [62]) and the general process of label-based DR evaluation (i.e., naive application of CVMs). From the scalability analysis, we validate that the runtime of using Label-T&C is competitive with that of the existing methods. Finally, we demonstrate two case studies showing that Label-T&C can be used to reveal how different DR techniques or hyperparameter settings affect embedding results.

2 Background and Related Works

We discuss the state-of-the-art in interpreting and measuring the reliability of DR embeddings. We then describe works about the common assumption that high-dimensional labeled data have good CLM.

2.1 Reliability of Dimensionality Reduction

2.1.1 Dimensionality Reduction

Dimensionality reduction (DR), e.g., *t*-SNE [61], UMAP [45], aims to produce the low-dimensional embedding preserving the structure of

- *Hyeon Jeon and Jinwook Seo are with Seoul National University. E-mail: hj@hcil.snu.ac.kr, jseo@snu.ac.kr*
- *Yun-Hsin Kuo and Kwan-Liu Ma are with the University of California, Davis. E-mail: {yskuo, klma}@ucdavis.edu*
- *Michaël Aupetit is with Qatar Computing Research Institute, Hamad Bin Khalifa University. E-mail: maupetit@hbku.edu.qa*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

the input high-dimensional data. DR plays an important role in many visual analytics tasks, including cluster identification [27,69] or neighborhood search [20,21,40]. This research provides reliable measures for evaluating DR embeddings regarding the matching between clusters and classes in both input and embedding spaces, establishing a basis for more trustworthy DR-based visual analysis.

2.1.2 Distortions in Dimensionality Reduction

While transferring the data from broad high-dimensional space to narrow low-dimensional space, DR unavoidably introduces distortions [2,50]. As distortions make embeddings less reliable in representing the original data, informing distortions is important in utilizing DR for data analysis [29,50].

Several distortion types were defined to formally explain DR distortions. Aupetit [2] initially defined *stretching* and *compression*. Stretching describes the situation in which the pairwise distances in the embedded space are increased compared to the ones of the original space; conversely, compression indicates the case that the pairwise distances decreased. *Missing Neighbors* and *False Neighbors* [37,40,41,63] were introduced as an interpretation of stretching and compression in terms of the neighborhood structure. Given a high-dimensional point x and its corresponding low-dimensional point z, Missing Neighbors are defined as the k-nearest neighbors of x that are not among the ones of z. Conversely, False Neighbors are defined as the k-nearest neighbors of z that are not among the ones of x. However, Missing and False Neighbors are insufficient to explain the distortions of complex, intertwined cluster structures. For example, the relative increase of cluster density in the embedding does not incur Missing and False Neighbors distortions, because it does not alter the k-nearest neighbor structure for small k values.

As alternatives, *cluster-level* distortions, named *Missing Groups* and *False Groups*, were proposed by Jeon et al. [29]. Missing Groups occur when a cluster in the original space splits into multiple separated clusters in the embedding, and False Groups occur when a cluster in the embedding consists of multiple separated clusters in the original space. In the seminal work [29], Missing and False Groups distortions are examined based on the groups obtained by clustering techniques.

In this work, we focus on evaluating the reliability of the cluster structure of DR embeddings by quantifying both Missing and False Groups. However, instead of extracting groups using clustering techniques, we focus on the groups given by the classes of labeled data.

2.1.3 Distortion Measurement without Labels

We discuss distortion measures that do not rely on class labels. These measures take the original and embedded data as input and quantify their structural difference. Aligned with the aforementioned distortion types, they focus on three different levels of structural granularity: *global*, *local*, and *cluster*. *Global measures*, such as Kullback-Liebler divergence (KL Divergence) and Distance to Measure (DTM) [15,16], quantify how well the embeddings preserve the global structure of the original data against stretching and compression. Meanwhile, *local measures* focus on neighborhood preservation. Trustworthiness and Continuity (T&C) [62] measure how Missing and False Neighbors affected the distance-based ranking of the nearest neighbor for every data point in both spaces. Mean Relative Rank Errors (MRREs) [37] extends T&C by additionally regarding the ranking of True Neighbors: the points that are neighbors in both the original and embedded spaces. Still, despite local and global measures' wide usage in practice [29,30,35,46,50,69], they do not properly capture cluster-level distortions [29].

This leads to the necessity of measures that capture distortions on cluster structures (i.e., *cluster-level measures*). Steadiness and Cohesiveness (S&C) [29] assess how much Missing and False Groups distortions have occurred by (1) extracting clusters from one space and (2) evaluating their dispersion in the other space. However, S&C require users to specify the way of extracting and investigating clusters in both spaces, e.g., using clustering techniques, making the results of the cluster-level distortion measures sensitive to the clustering technique and hyperparameters used. S&C also suffers from a scalability issue as it requires the iterative execution of a clustering technique [25,29].

Label-T&C is a pair of cluster-level measures that aim to tackle these limitations. At first, the measures require a CVM as the sole hyperparameter, which is used to evaluate CLM in the original and embedded spaces. Thanks to the low complexity of CVM [28,42], our measures are very scalable (Sect. 5.2). Furthermore, Label-T&C are more sensitive in distinguishing Missing and False Groups distortions compared to previous measures, including S&C (Sect. 5.1).

2.1.4 Distortion Measurement with Labels

Exploiting labels is a common scheme in evaluating DR embeddings [8,17,22,32,69,71,73]. A general process to do so is to utilize CVM to measure the CLM of embeddings [8, 22, 32, 73]. However, the approach is prone to producing errors while examining the quality of DR embedding. For example, if the CLM of the original data is bad (e.g., some classes overlap), embeddings that have good CLM for bad reasons (e.g., DR artificially separates each class into a distinct cluster) will be considered high-quality embeddings. As non-expert users typically assume that DR techniques generate reliable embeddings of the original data, they may incorrectly conclude that CLM is also good in the high-dimensional space, while it is not actually true [3,28].

A sole pair of measures that relies on class labels but is independent of CVM is Class-Aware Trustworthiness and Continuity (CA-T&C) [17]. CA-T&C is a variant of T&C that assess the degradation of CLM (i.e., False Groups distortions) by estimating the extent to which Missing and False Neighbors occurred within and between classes, respectively. However, CA-T&C hardly captures the Missing Groups distortions as they do not consider the increase of CLM as distortions. The measures also mainly focus on local structures and thus cannot comprehensively examine CLM distortions.

In this work, we propose Label-T&C as novel measures utilizing class labels to evaluate DR embeddings. As with the general process of label-based DR evaluation (i.e., the process of naively applying CVM in the embedded space), our measures utilize CVMs to evaluate CLM; however, by applying CVM to both the original and embedded spaces and assessing their difference, our measures precisely capture cluster-level distortions.

2.2 The Cluster-Label Matching assumption

The assumption that the CLM is good in the high-dimensional data is used as a basis not only for the label-based evaluation of DR embeddings but also for other applications. For example, the labels are often utilized as the ground truth partition in clustering validations, where clustering techniques that generate a similar partition to that of labels obtain higher scores (i.e., external clustering validation; refer to Sect. 3.1 for details). Another application is the perception-based evaluation of DR techniques [20,21,57,69], where techniques that produce embeddings in which the visual clustering results of human subjects better match class labels are preferred. However, the assumption can be easily broken [3,23], which casts doubt on the applications' reliability.

Despite such a threat, only a few solutions have emerged. A trivial solution is to modify datasets to make them better satisfy the assumption. Aupetit [3,6] proposed to check the linear or nonlinear separability of classes and then merge overlapped classes or preserve one of them while removing the others [3]. However, classes can be separable but adjacent, not forming proper clusters (no low density or wide empty space between them). Such a strategy also does not take into account whether each class forms a single, compact cluster. Another solution is to use synthetic datasets [29,30,46], where good CLM is guaranteed by design. Still, this makes the evaluation hardly generalizable to real data. Alternatively, Jeon et al. [28] suggested a systematic way to evaluate CLM; their purpose was to verify the validity of labeled datasets for use as clustering validation benchmarks. Still, they suggested only utilizing datasets with good CLM, which reduces the number of available datasets for evaluating DR embeddings.

In this work, we neither verify the CLM of datasets in advance nor attempt to modify datasets to enhance CLM. Instead, we *acknowledge* that datasets may not satisfy the CLM, and rather assess whether the degree of CLM, either high or low, in the original dataset is well preserved in the embedding.

3 General Label-based DR Evaluation Process

The general process of label-based DR evaluation mostly relies on CVMs. We describe what CVMs are and the process of using them to evaluate CLM. We then discuss the pitfalls of the process.

Notations We define a high-dimensional data $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^D, i = 1, 2, \cdots, N\}$. We denote the low-dimensional embedding of $\mathbf{X}$ as $\mathbf{Z} = \{\mathbf{z}_i \in \mathbb{R}^d \mid i = 1, 2, \cdots, N\}$, where $D > d$. For any set $\mathbf{S} \in \{\mathbf{X}, \mathbf{Z}\}$, the distance function δ satisfies $\delta(x,y) \geq 0$, $\delta(x,y) = \delta(y,x)$ and $\delta(x,y) = 0$ if $x = y$ $\forall x, y \in \mathbf{S}$. A partition of $\mathbf{S}$ is defined as $\mathbf{P} = \{P_1, P_2, \cdots, P_k\}$ satisfying $P_i \subseteq \mathbf{S}$, $P_i \cap P_j = \emptyset$ and $\cup_{i=1}^{k} P_i = \mathbf{S}$. If a partition is defined by class labels, we denote the partition as $\mathbf{P}_L$. A clustering technique C takes $\mathbf{S}$ and δ as input and returns a partition $\mathbf{P}_C$ of $\mathbf{S}$.

3.1 Clustering Validation Measures

Clustering validation measures (CVMs) evaluate how well-clustered the given partition (i.e., clustering) is in the given data. We use CVMs to find the optimal clustering technique or hyperparameter setting that produces the partition of the data that best matches its cluster structure. CVMs are largely divided into two types: **internal CVM (IVM)** [42, 43] and **external CVM (EVM)** [68]. IVMs evaluate a partition based on the internal structure of data. Formally, the IVM score $m_I(\mathbf{P}, \mathbf{X}, \delta)$ quantifies how well the groups within the partition $\mathbf{P}$ of $\mathbf{X}$ are individually condensed and mutually separated in $\mathbf{X}$ based on distance δ. For example, the Silhouette Coefficient [55] examines how the within-group and between-group distances differ on average while using Euclidean distance as δ. Alternatively, EVMs, such as the adjusted rand index [64], rely on a ground truth partition $\mathbf{P}_{GT}$. Here, the EVM score $m_E(\mathbf{P}, \mathbf{P}_{GT})$ simply quantifies the degree of matching between the given partition $\mathbf{P}$ and $\mathbf{P}_{GT}$, regardless of the internal cluster structure of $\mathbf{S}$. A higher score is assigned if $\mathbf{P}$ better matches with $\mathbf{P}_{GT}$. Data class labels $\mathbf{P}_L$ are typically used as ground truth $\mathbf{P}_{GT}$ [23, 28].

3.2 Using CVM to Evaluate CLM

We use CVMs to quantify the CLM of a DR embedding as a proxy for its reliability [8, 32, 69, 73]. The process depends on the type of CVM:

IVM-based evaluation For a given embedding $\mathbf{Z}$, distance function δ, and class labels $\mathbf{P}_L$, $m_I(\mathbf{P}_L, \mathbf{Z}, \delta)$ represents the CLM between $\mathbf{P}_L$ and $\mathbf{Z}$. The Silhouette Coefficient is widely adopted in the visualization community [20, 22, 32, 65, 69]. The Davies-Bouldin index [18] is preferable in the context of star coordinates and Radviz [1, 14]. Notably, while Distance Consistency (DSC) [59] was designed for DR visual quality evaluation [19, 56, 58], it can also be viewed as a CVM since it considers only the separation of class labels in the embeddings.

EVM-based evaluation Given $\mathbf{Z}$, δ, $\mathbf{P}_L$, and a clustering technique C providing a partition $\mathbf{P}_C = C(\mathbf{Z}, \delta)$ of the embedded data, $m_E(\mathbf{P}_C, \mathbf{P}_L)$ represents CLM between $\mathbf{P}_L$ and $\mathbf{Z}$. K-Means and the adjusted rand index are commonly used for C and m_E, respectively [31, 71, 74].

Notice that CVMs cannot account for the internal compactness of each class in isolation, but the CVM of a class partition will get worse if some of these classes lack compactness or split across several clusters.

3.3 Pitfalls

The general process of label-based DR evaluation promotes embeddings with good CLM regardless of the CLM of the original data (Sect. 1). In other words, the process examines the extent to which CLM is harmed in embeddings while assuming that the original data has good CLM. Thus, if the assumption is broken, the process will frame embeddings that correctly represent overlapped classes to have False Groups distortions. As the process considers good CLM embeddings as high-quality ones, it is also incapable of detecting Missing Groups distortions that may arise from CLM amplification. These pitfalls were identified for the first time by Aupetit [3]. Our preliminary experiment confirms such a threat (Appendix D). The general process of label-based evaluation erroneously prefers DR techniques that maximize the separation among classes, instead of the ones that aim to preserve the original structure of data if the datasets have bad CLM. Here, we aim to introduce a new way of using class labels for DR evaluation that mitigates such a bias.

4 Label-Trustworthiness & Label-Continuity

We introduce two distortion measures—Label-Trustworthiness and Label-Continuity (Label-T&C)—as an alternative way of using class labels for DR evaluation. Our measures examine how CLM differs in *both* the original and embedded spaces where CVM is used to quantify CLM. Label-T and Label-C capture the False and Missing Groups distortions, respectively. The measures are named after Trustworthiness and Continuity, two local distortion measures that focus on capturing False and Missing Neighbors [62].

4.1 Design Rationale

Inputs, output, and hyperparameters Label-T&C take (1) the high-dimensional data $\mathbf{X}$, (2) its DR embedding $\mathbf{Z}$, and (3) class labels $\mathbf{P_L} = \{P_{L,1}, P_{L,2}, \cdots P_{L,k}\}$ as inputs. Both Label-T and Label-C output a number between 0 and 1; a higher value indicates lower distortions and a better embedding. For hyperparameters, a CVM m with distance function δ is given. If m is an EVM, we need to additionally select the clustering technique C as a hyperparameter (Sect. 3.2). The m should assign higher scores to better clusterings and range from 0 to 1 (refer to Sect. 4.2.1 for a detailed explanation about this requirement).

Step 1. Measuring CLM in the original and embedded spaces We apply CVM to both the original and embedded spaces to examine CLM. Here, unlike the general process of label-based DR evaluation that applies CVM to all classes at once, we apply CVM to every *pair* of classes, so that we can take account of the relationships of classes in more detail. Formally, we construct the class-pairwise CLM matrices $M(\mathbf{X})$ and $M(\mathbf{Z})$, where the (i,j)-th cell of the matrices $M(\mathbf{S})_{i,j}$ ($\mathbf{S} \in \{\mathbf{X}, \mathbf{Z}\}$) is defined as:

$$\begin{cases} m(\{P_{L,i}, P_{L,j}\}, \mathbf{S}, \delta) & \text{if} \quad i \neq j \text{ and } m \text{ is an IVM} \\ m(C(P_{L,i} \cup P_{L,j}, \delta), \{P_{L,i}, P_{L,j}\}) & \text{if} \quad i \neq j \text{ and } m \text{ is an EVM} \\ 0 & \text{if} \quad i = j \end{cases}.$$

Step 2. Computing distortion matrices We construct a matrix $M^* = M(\mathbf{X}) - M(\mathbf{Z})$ representing CLM distortions. We then compute M^{FG} and M^{MG}, where $M^{FG}_{i,j} = (M^*_{i,j}$ if $M^*_{i,j} > 0$, else 0), and $M^{MG}_{i,j} = (-M^*_{i,j}$ if $M^*_{i,j} < 0$, else 0). M^{FG} and M^{MG} abstract the CLM decrement (False Groups) and increment (Missing Groups), respectively.

Step 3. Aggregating distortions Finally, we average the upper-diagonal elements of M^{FG} and M^{MG} into final scores:

$$\text{Label-Trustworthiness: } 1 - \text{avg}_{i,j} M^{FG}_{i>j}$$

$$\text{Label-Continuity: } 1 - \text{avg}_{i,j} M^{MG}_{i>j}.$$

Note that we subtract the average from 1 to make embeddings with fewer distortions receive higher quality scores.

4.2 Selecting CVM for Label-T&C

We establish the requirements for CVM to get proper Label-T&C scores and present suitable CVM options. In this section, we denote $m(\mathbf{P}, \mathbf{S}, \delta)$ as a CVM score with respect to $\mathbf{P}$, $\mathbf{S}$, and δ. If m is an IVM m_I, we set $m(\mathbf{P}, \mathbf{S}, \delta) = m_I(\mathbf{P}, \mathbf{S}, \delta)$. If m is an EVM m_E, we set $m(\mathbf{P}, \mathbf{S}, \delta) = m_E(C(\mathbf{S}, \delta), \mathbf{P})$ with C, the chosen clustering technique.

4.2.1 Requirements

We set the first three requirements based on the following proposition: to be used for Label-T&C, a proper CVM should be comparable across $\mathbf{X}$ and $\mathbf{Z}$. In other words, m shall consider only *how well the given partition is clustered in the given data* and be invariant to the characteristics that differ between $\mathbf{X}$ and $\mathbf{Z}$ but are not related to the cluster structure. For example, the scaling of the pairwise distances should not alter the score. Otherwise, the evaluation will be unreliable; for example, we can simply manipulate Label-T&C scores by scaling the original or embedded data while there is no change in the cluster structure.

Previous works [10, 28] set axioms defining how a CVM can be independent of such features. They require CVMs to be invariant to the change of scale, dimensionality, and the number of points and classes and to have a fixed range. As $\mathbf{X}$ and $\mathbf{Z}$ already share the number of points and classes, we require CVMs to ensure the other three axioms.

The first axiom requires CVMs to be invariant against the scaling of distances between points, which can be inherently different in $\mathbf{X}$ and $\mathbf{Z}$:

Scale Invariance [10] *A CVM m is scale invariant if $\forall$ partition $\mathbf{P}$, data $\mathbf{S}$, and distance function δ, $m(\mathbf{P},\mathbf{S},\delta) = m(\mathbf{P},\mathbf{S},\alpha\delta)$ $\forall \alpha > 0$ (where $\alpha\delta$ is a distance function satisfying $\alpha\delta(x,y) = \alpha \cdot \delta(x,y)$, $\forall x, y \in \mathbf{S}$.).*

(R1) A CVM should satisfy scale invariance.

The second axiom focuses on the effect of the data dimension on the distance δ, due to the so-called curse of dimensionality [9]. The growing dimensions increase the average of pairwise distances while the variances remain constant [11, 24, 38], thus the differences between distances become negligible. To be used for Label-T&C, CVM should be shift invariant [38, 39] to cancel the shift of the average distances due to the different dimensions of $\mathbf{X}$ and $\mathbf{Z}$.

Shift Invariance [28] *A CVM m is shift invariant if $\forall \mathbf{P}, \mathbf{S}, \delta$, $m(\mathbf{P},\mathbf{S},\delta) = m(\mathbf{P},\mathbf{S},\delta+\beta)$ $\forall \beta > 0$ (where $\delta+\beta$ is a distance function satisfying $(\delta+\beta)(x,y) = \delta(x,y)+\beta$, $\forall x, y \in \mathbf{S}$).*

(R2) A CVM should satisfy shift invariance.

The final axiom is about requiring CVMs to produce scores that conform to a fixed range of values, which is designed to capture the remaining subtle factors that are not influenced by the cluster structure.

Range Invariance [28] *A CVM m is range invariant if $\forall \mathbf{S}, \delta$, $\min_{\mathbf{P}} m(\mathbf{P},\mathbf{S},\delta) = \alpha$ and $\max_{\mathbf{P}} m(\mathbf{P},\mathbf{X},\delta) = \beta$, where α, β are constants satisfying $\alpha < \beta$ (arbitrarily set to 0 and 1, respectively).*

(R3) A CVM should satisfy range invariance.

Additionally, we want CVMs to be stable against the change of hyperparameters. This is because the alteration of CVM scores due to the hyperparameter change can induce uncertainty in utilizing Label-T&C. This leads to the last axiom:

(R4) A CVM should have no hyperparameter or should produce similar scores on the same input regardless of the hyperparameter settings.

4.2.2 CVM Candidates

We examine CVMs commonly used for DR evaluation (Sect. 3.2) as potential candidates to be used for Label-T&C. For EVMs, we find that the combination of K-Means and adjusted rand index cannot be used. This is because the parameter K (i.e., number of clusters) in K-Means leads to the violation of **R4**. Indeed, as clustering techniques commonly require hyperparameters, EVMs hardly satisfy the aforementioned requirements. Studying how EVMs and clustering techniques can satisfy R4 is beyond the scope of this work.

For IVMs, neither the Silhouette Coefficient [55] nor the Davies-Bouldin index [18] satisfies shift invariance (**R2**; refer to Appendix A for the proof). However, we found that DSC satisfies all requirements, setting it as a proper CVM for Label-T&C (Appendix A).

We additionally found that the between-dataset Calinski-Harabasz index (CH_{btwn}) [28], a variant of Calinski-Harabasz index [60], satisfies the four requirements: satisfaction of **R1**, **R2**, and **R3** has been demonstrated earlier [28]; it also satisfies **R4** as its unique hyperparameter is the number of Monte-Carlo simulations for normalizing the measure, which barely affects the output if the number is sufficiently high. We give a brief description of these two CVMs usable for Label-T&C:

Distance Consistency (DSC) [59] DSC is defined as the number of data points closer to the centroid of another class than their own in the data, normalized by the total number of data points. As DSC ranges from 0.5 to 1 if the number of classes is two and assigns a lower score for a better CLM, we use the value $2(1-\text{DSC})$ to make it satisfy **R3** (Sect. 4.1 (Step 1)).

Between-dataset Calinski-Harabasz index (CH_{btwn}) [28] CH_{btwn} is defined as the ratio of compactness to separability. Compactness is defined as the distance between data points and the class centroids to which each point belongs, and separability is determined by the distances between class centroids and the centroid of the entire data.

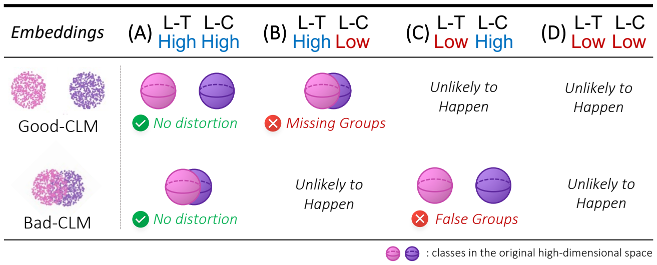

Fig. 1. Guidelines to infer the CLM of the high-dimensional data based on the CLM of the embedded data (left column) and the scores given by Label-T (L-T) and Label-C (L-C) (first row) (see Sect. 4.3 for details).

4.3 Guidelines to Interpret Label-T&C

We present the guidelines to interpret embeddings based on Label-T&C. If Label-T and Label-C are both high, the CLM of the embedding accurately depicts the CLM in the original space (Fig. 1A). High Label-T and low Label-C (Fig. 1B) mean that Missing Groups distortions have occurred, i.e., the CLM of the original data is worse than it appears in the embedding (first row); some pairs of classes appear more separated than they actually are in the data space. When the CLM of the embedding is already low (*e.g.* due to overlapping classes), Missing Groups distortions are more unlikely to happen as the CLM in the data would have to be even worse (second row). In contrast, low Label-T and high Label-C (Fig. 1C) inform that False Groups distortions have occurred; the CLM in the original data is better than in the embedding (second row). As False Groups distortions deteriorate the CLM of the embedding, the situation is unlikely to occur if the embedding has a good CLM, and thus can hardly become better (first row). Due to such a tradeoff between False and Missing Groups (i.e., more Missing Groups lead to fewer False Groups, and vice versa), it is unlikely to get low Label-T and Label-C at the same time (Fig. 1D). Our sensitivity analysis (Sect. 5.1; Fig. 4) confirms the existence of the tradeoff.

4.4 Time Complexity

The complexity of Label-T&C depends on the CVM. As DSC is $O(|\mathbf{S}||\mathbf{P}_L|\Delta_{\mathbf{S}})$, where $\Delta_{\mathbf{S}}$ denotes the dimensionality of $\mathbf{S}$, applying it to a pair of classes $P_{L,i}, P_{L,j}$ requires $O(|P_{L,i} \cup P_{L,j}|\Delta_{\mathbf{S}})$. As each class is considered $|\mathbf{P}_L|$ times, Label-T&C with DSC is $O(|\mathbf{S}||\mathbf{P}_L|\Delta_{\mathbf{S}})$. Similarly, as CH_{btwn} is $O(|\mathbf{S}||\mathbf{P}_L|^2\Delta_{\mathbf{S}})$, applying it to a pair of classes $P_{L,i}, P_{L,j}$ requires $O(|P_{L,i} \cup P_{L,j}|\Delta_{\mathbf{S}})$. Therefore, Label-T&C with CH_{btwn} is $O(|\mathbf{S}||\mathbf{P}_L|\Delta_{\mathbf{S}})$. In both cases, the time complexity is linear in all variables. We evaluate the scalability of Label-T&C in Sect. 5.2.

4.5 Implementation & Deployment

We deploy Label-T&C as a Python library. We provided an interface that allows users to implement and test custom CVM as a hyperparameter. The source code is available at `github.com/hj-n/ltnc`.

5 Quantitative Evaluations and Discussions

We conduct quantitative experiments to evaluate Label-T&C with DSC and CH_{btwn}, i.e., Label-T&C [DSC] and Label-T&C [CH_{btwn}], respectively. In the sensitivity analysis (Sect. 5.1), we check the accuracy of Label-T&C and competitors in quantifying distortions. We also evaluate the runtime of the measures (Sect. 5.2).

Competitors. We first consider all distortion measures without labels (Sect. 2.1.3) as competitors. For global measures, we use KL divergence and DTM. T&C and MRRE are used as representative local measures. MRRE [Missing] and MRRE [False] target Missing and False Neighbors, respectively. We select S&C as the sole pair of measures targeting cluster-level distortions. For the measures using labels (Sect. 2.1.4), we first add CA-T&C. We then select Silhouette and DSC as representative CVMs used in the general label-based evaluation. For T&C, MRRE, and CA-T&C, we average their score across k-nearest neighbor values: $k = [5, 10, 15, 20, 25]$, following Jeon et al. [29]. For KL divergence and DTM, we average the scores across different standard deviation

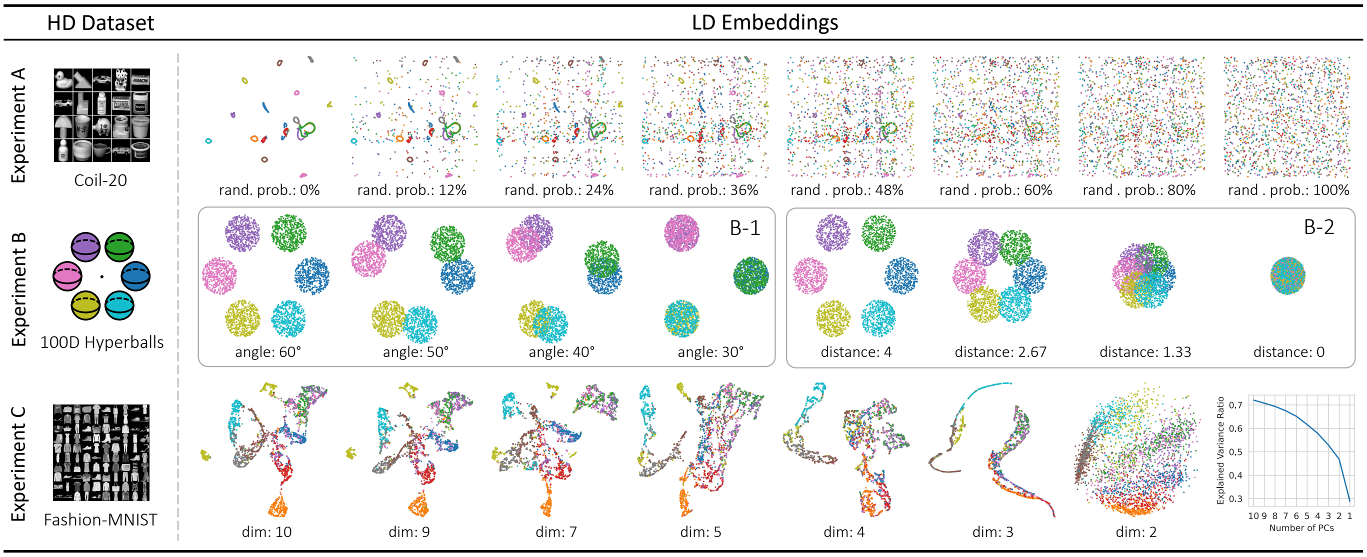

Fig. 2. The high-dimensional (HD) datasets and low-dimensional (LD) embeddings used in experiments A, B, and C of sensitivity analysis (Sect. 5.1). The experiments aim to check the distortion measures' ability to capture False Groups distortions. Class labels are mapped to colors. (A) The Coil-20 [49] dataset and the embeddings generated by randomizing the positions of the embedded points with a certain probability. (B) A HD dataset consists of six well-separated hyperballs (left) and its synthetic embeddings (right) made by initializing the embedding with six well-separated discs and gradually overlapping the discs in two different manners (B-1, 2). (C) The Fashion-MNIST [72] dataset and the PCA embeddings with different numbers of principal components (PC); here we depict the UMAP projection of PCA embeddings if it has more than two PCs (i.e., dimensionality is higher than two). We depict the relation between explained variance ratio and the number of PC in the line chart next to the embeddings.

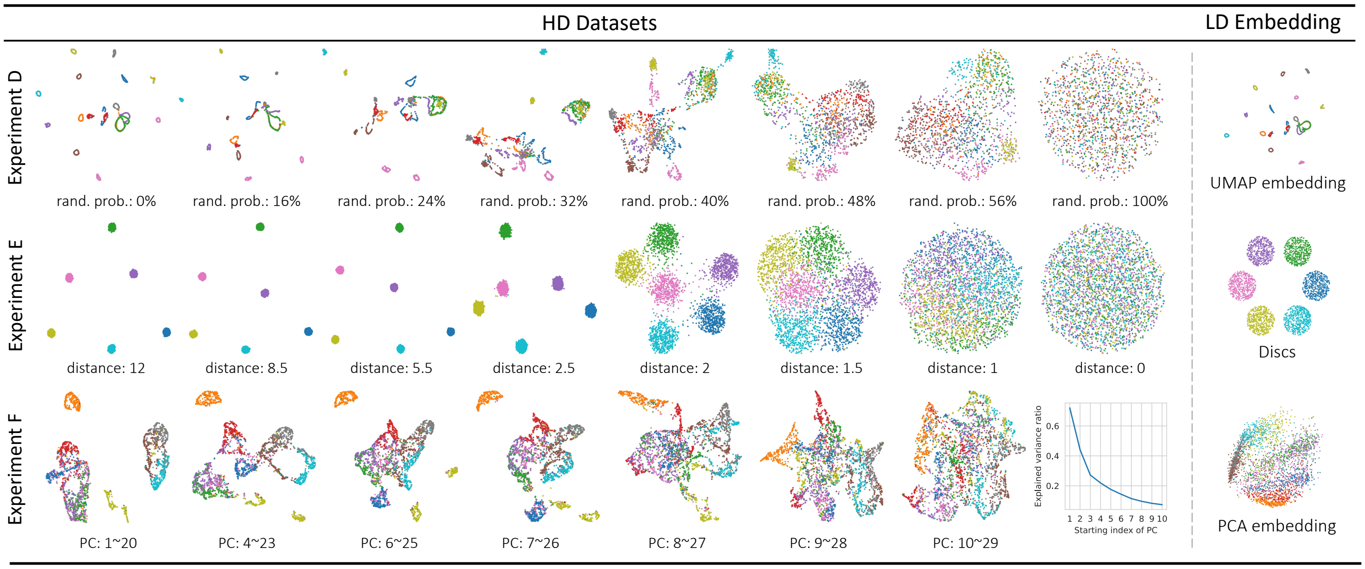

Fig. 3. The low-dimensional (LD) embeddings and corresponding high-dimensional (HD) datasets represented as UMAP embeddings, used in experiments D, E, and F of sensitivity analysis (Sect. 5.1) to examine distortion measures' ability to capture Missing Groups distortions. (D) An UMAP embedding of the Coil-20 [49] dataset (right), and the variants of the Coil-20 dataset made by randomizing the coordinates of data points in HD space with a certain probability. (E) A 2D embedding with six well-separated discs and synthetic HD datasets. We create the datasets by generating six 100D hyperballs and gradually overlapping them. (F) A 2D PCA embedding of the Fashion-MNIST dataset and corresponding HD datasets variants, created by slicing 20 principal components (PC) with different rankings. The line chart shows their corresponding explained variance ratio.

values of Gaussian kernels σ: $[0.01, 0.1, 1]$, following Moor et al. [46]. For S&C, we use the default hyperparameter setting [29].

5.1 Sensitivity Analysis

We conduct six experiments (A-E) to examine Label-T&C's sensitivity in quantifying False Groups (Fixed data and variable embeddings in experiments A, B, and C) or Missing Groups (Variable data and fixed embeddings in experiments D, E, and F) distortions. The labeled data and embeddings used in the experiments can be found in Fig. 2 (A, B, and C) and Fig. 3 (D, E, and F). In all of them, we run Label-T&C and competitors to evaluate the embeddings.

5.1.1 Objectives and Design

Experiment A: Randomizing embeddings We examine whether Label-T&C and competitors can accurately quantify False Groups distortions. We generate a 2D UMAP embedding of the Coil-20 [49] dataset. We then create variants of the embedding with different levels of False Groups distortions by randomizing the location of the points. We create 21 variants, ranging the replacement probability from 0% (same as the original embedding) to 100% (totally randomized) with an interval of 5%. The original class assignments of Coil-20 are used as labels. We hypothesize that Label-T will decrease as the replacement probability grows, properly capturing False Groups distortions, while

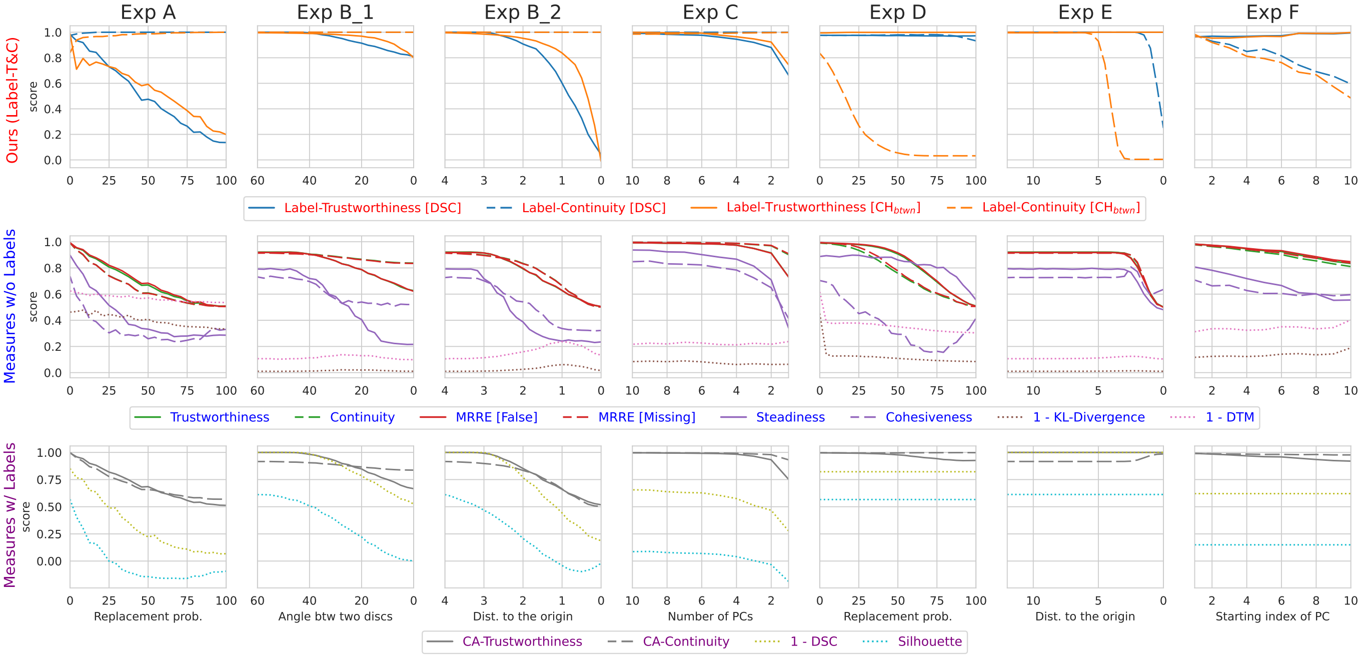

Fig. 4. The results of the sensitivity analysis (Sect. 5.1; experiments A-F). Solid lines and dashed lines represent the measure that focuses on compression (e.g., False Groups, False Neighbors) and stretching (e.g., Missing Groups, Missing Neighbors), respectively. Dotted lines represent global measures and CVMs. A pair of compression and stretching measures is represented with the same line color. Measure names in red, blue, and purple correspond to our approach, the measures without labels (Sect. 2.1.3), and the measures with labels (Sect. 2.1.4), respectively. In summary, Label-T (blue and orange bold line) and Label-C (blue and orange dotted line) accurately detect Missing and False Groups distortions, respectively. Meanwhile, all other measures, including general label-based DR evaluation (i.e. DSC and Silhouette), fail to capture these distortions.

Label-C will ignore the distortions.

Experiment B: Overlapping discs We aim to check distortion measures' ability to precisely capture False Groups distortions, as with experiment A. We create a high-dimensional dataset consisting of six hyperballs with a radius of 5 lying in 100 dimensions. We set the hyperballs to be equidistant (= 10) from the origin. We then create an artificial 2D embedding consisting of six discs (radius of 1.5) evenly and equidistantly (= 4) distributed around the origin O. Data points and labels within each disc correspond to those of each hyperball. The positions of each point within the disc and hyperball are determined randomly. The label is also set based on the disc each point belongs to. We gradually overlap the discs to artificially generate distortions. Here, we use two overlapping schemes to evaluate the sensitivity of Label-T&C in detail, resulting in two separate subexperiments (B-1, B-2). In B-1, three independent pairs of adjacent discs are overlapped; for each pair of discs (A,B) with centers C_A, C_B, we adjusted $\angle C_A O C_B$ from 60° to 0° with an interval of 2.4° (25 embedding variants in total). In B-2, we overlap all discs at once by moving them toward the origin; for each disc A, we gradually decrease $C_A O$ from 4 to 0 with an interval of 0.16 (25 embedding variants). We hypothesize that the Label-T score will go down as False Groups distortions increase due to the overlap of the discs, while Label-C will stay still. We also hypothesize that Label-T will decrease more in B-2 than in B-1, as the overlap is larger.

Experiment C: Decreasing the dimension of the embedded space We generate False Groups distortions by decreasing the dimensionality of embedded space and check whether the measures can detect the distortions. We prepare the Fashion-MNIST [72] as a high-dimensional dataset. We generate PCA embeddings with a decreasing number of top principal components (10 to 1 with an interval of 1; 10 embeddings in total). We expect the embeddings with a smaller number of principal components (i.e., embeddings lying in the space with fewer dimensions) to have more False Groups distortions as they have a smaller explained variance ratio (line chart in Fig. 2). We use the class assignments of the Fashion-MNIST dataset as labels. Our hypothesis is that Label-T will decrease as the dimensionality decreases, while Label-C will stay still.

Experiment D: Randomizing the original data We want to evaluate Label-T&C and competitors' capability in accurately quantifying Missing Groups distortions. We first generate a fixed 2D UMAP embedding of the Coil-20 [49] dataset. We then generate the variants of the original data by mixing the points in the high-dimensional space with a fixed probability, producing Missing Groups distortions. We control the replacement probability from 0% to 100% with an interval of 5%, resulting in 21 variants. The class assignments of the original data are used as labels. We hypothesize that Label-C will decrease as Missing Groups distortions increase (i.e., replacement probability increase), and that Label-T will ignore the distortions.

Experiment E: Overlapping hyperballs We want to evaluate whether Label-T&C and competitors can precisely capture Missing Groups distortions. We prepare variants of high-dimensional data and fixed low-dimensional embedding consisting of six 100D hyperballs and corresponding 2D discs, respectively. The points within the same disc have the same label. All discs are well separated from each other. We artificially overlap hyperballs to generate Missing Groups distortions. For each hyperball A_H, we gradually decrease $C_{A_H} O$ from 4 to 0 with an interval of 0.16 (25 variants in total). We hypothesize that Label-C will decrease as hyperballs overlap, while Label-T will stay still.

Experiment F: Decreasing the dimension of the original data space We examine whether the distortion measures can detect the Missing Groups distortions made by the decrease in the dimensionality of the original data. We prepare a 2D PCA embedding of the Fashion-MNIST dataset. We then select ten 20D PCA embeddings with different sets of principal components as high-dimensional datasets; the i-th dataset variant consists of the (i)-th to $(i+19)$-th principal components, where $1 \leq i \leq 10$. We expect the dataset with a higher order to have more Missing Groups distortions over the embedding as they have a smaller explained variance ratio (line chart in Fig. 3). We used the class assignments of the Fashion-MNIST dataset as labels. We hypothesize that Label-C will decrease as the starting index of principal components increases, while Label-T will stay still.

5.1.2 Results

Fig. 4 shows the results of our experiments that we comment on below.

Experiment A As the randomization probability grows, both Label-T [DSC] and Label-T [CH_{btwn}] similarly decrease linearly while Label-C

[DSC] and Label-C [CH_{btwn}] slightly increase, confirming our hypothesis. Meanwhile, S&C and local measures decrease regardless of the distortion type, while global measures slightly increase. In the case of label-based measures, both CA-T&C and the general CVM-based process (DSC and Silhouette) show mainly decreasing scores.

Experiment B In B-1, as the overlap between the discs grows, both Label-T [DSC] and Label-T [CH_{btwn}] decrease in a similar manner, while Label-Cs stay still. Such results validate our hypothesis, confirming Label-T&C's capability in properly detecting False Groups distortions. Meanwhile, S&C, T&C, and MRREs all decrease, while Steadiness, Trustworthiness, and MRRE [False] decrease more than Cohesiveness, CA-Continuity, and MRRE [Missing], respectively. Global measures stay still. CA-T&C partially succeed in properly detecting False Groups distortions; both CA-Continuity and CA-Trustworthiness decrease, but CA-Continuity's decrement was subtle compared to the one of CA-Trustworthiness. CVMs show a decreasing trend. In B-2, the amount of decrement becomes bigger than in B-1 for Label-T [DSC] and Label-T [CH_{btwn}] while Label-Cs again stay still, confirming our second hypothesis. The amount of decrement also becomes bigger than in B-1 for T&C, MRREs, and Cohesiveness, while Steadiness showed a similar drop as in B-1. In the case of KL divergence, DTM, and Silhouette, the patterns are almost identical to B-1 except that the scores rebound when the discs are nearly overlapped. The decrement becomes bigger also for CA-T&C and DSC.

Experiment C As the number of PCs decreases, Label-Ts decrease while Label-Cs stay still, validating our hypothesis. Global measures (KL divergence, DTM) stay still while all other measures decrease.

Experiment D As we increase the randomization probability, both Label-C [DSC] and Label-C [CH_{btwn}] decrease, while Label-Ts stay still, verifying our hypothesis. However, while Label-C [DSC] decreases right before the data are perfectly mixed, Label-C [CH_{btwn}] decreases from the start. For local measures, both T&C and MRREs decrease. Steadiness decreases, while Cohesiveness suddenly goes up after decreasing for a while. Global (KL divergence, DTM) measures increase in general. CA-Trustworthiness goes down while CA-Continuity stays still, and CVMs (DSC and Silhouette) stay still.

Experiment E. When the overlap between hyperballs increases, both Label-C [DSC] and Label-C [CH_{btwn}] decrease, while Label-Ts stay still, verifying our hypothesis. However, as in experiment D, Label-C [DSC] and Label-C [CH_{btwn}] decrease differently; while Label-C [DSC] decreases right before the hyperballs perfectly overlap, Label-C [CH_{btwn}] decreases before Label-C [DSC] does. Meanwhile, local measures (T&C, MRRE) decrease, while global measures (KL divergence, DTM) stay still. Steadiness decreases while Cohesiveness temporarily pops up when Steadiness starts to decrease. CA-Trustworthiness maintains a maximum score while the CA-Continuity score increases before the perfect overlap of the hyperballs. CVMs stay still.

Experiment F. The results confirm our hypothesis; as the starting index of the PCs that we slice increases, both Label-C [DSC] and Label-C [CH_{btwn}] decrease while Label-Ts stay still. Local measures (T&C, MRRE) decrease, and global measures (KL divergence, DTM) stay still. S&C decrease, while Steadiness decreases more than Cohesiveness. CA-T&C show a similar trend; CA-Trustworthiness decreases, while CA-Continuity decreases to a smaller extent. CVMs stay still.

5.1.3 Discussions

Label-T&C and competitors' capability in detecting cluster-level distortions. The results from experiments A-C confirm that Label-T is sensitive to False Groups distortions, while Label-C is not, as we intended. Moreover, the difference between the B-1 and B-2 results validates Label-T's accuracy at measuring the amount of False Groups distortions. The results from experiments D-F, on the other hand, confirm that Label-C accurately captures Missing Groups distortions, while Label-T ignores them.

The results also validate that previous measures fail to accurately detect the distortions or to distinguish specific distortion types. Global measures (KL Divergence, DTM) hardly discover distortions for all six experiments. Local measures (T&C, MRRE) fail to pinpoint specific distortion types; all measures decrease regardless of the type of distortion they aim to measure. Cluster-level measures (S&C) fail to distinguish False Groups distortions in experiments A-C. For experiments D-F, the situation is even worse; Steadiness reacts more sensitively to Missing Groups distortions although it was originally designed to aim at False Groups distortions. CA-T&C succeed in pinpointing False Groups distortions for B-1, but fails to do so for the remaining experiments.

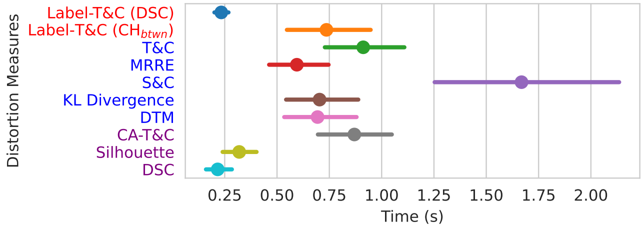

Fig. 5. Results of the scalability analysis. Name and line colors match with Fig. 4. Label-T&C [DSC] (dark blue) is on par with CVMs (Silhouette, DSC), while Label-T&C [CH_{btwn}] is similar to most of the other measures. S&C is the slowest.

The general process of label-based DR evaluation based on CVMs (DSC and Silhouette) succeeds in detecting the False Groups distortions in experiments A-C. However, in experiments D-F, the process fails to detect Missing Groups distortions. Moreover, the process does not have a specific focus on distortion type and thus cannot explain whether the False or Missing Groups distortions occurred. Such results confirm the threat of using the general label-based evaluation of DR in practice, providing clear evidence for adopting Label-T&C instead.

Effect of CVM choice on Label-T&C. Label-T&Cs with two different CVMs (DSC or CH_{btwn}) show a consistent pattern in experiments A-C. However, they behave differently in experiments D and E; Label-C [CH_{btwn}] starts decreasing for the lower level of generated CLM distortions than Label-C [DSC]. This observation may be CVM-specific as DSC and CH_{btwn} use different schemes in examining how the classes are clustered. In Label-C [DSC], the score only drops when classes overlap. Therefore, Label-C [DSC] is sensitive to Missing Groups distortions only if the *overlapped* classes in the original space are more separated in the embedding. In contrast, CH_{btwn} decreases as the proximity between classes increases, whether the classes overlap or not. Thus, when proximity increases, Label-C [CH_{btwn}] is more sensitive to Missing Groups distortions than Label-C [DSC]. The results indicate that CH_{btwn} has a larger range of variation, being more sensitive to CLM than DSC, but it is less sensitive to class overlap. Creating a CVM both sensitive to CLM and class overlap while fulfilling our requirements (Sect. 4.2.1) constitutes an interesting future work.

Discussions on the competitors. We discuss the patterns shown by competitors with more detail in Appendix B.

5.1.4 Sensitivity Analysis with the Class Labels Generated by Clustering Techniques

We want to validate whether the results of our study are replicable with the labels that come from other sources. We thus conduct experiments A-F while generating class labels with clustering techniques (Appendix F). We find that Label-T&C show consistent results regardless of the sources of labels, while the general label-based DR evaluation process (i.e., CVMs) fails to do so. Such results confirm the robustness of the Label-T&C in evaluating the quality of DR embeddings.

5.2 Scalability Analysis

5.2.1 Objectives and Design

We evaluate the scalability of Label-T&C against the competitors. We gather 96 labeled datasets [28] that vary in dimensionality, the number of data points, and the number of classes. We exclude two datasets as the implementation of S&C provided by the authors[1] fails to process them, resulting in 94 datasets (Appendix C). We generate embeddings using t-SNE, UMAP, PCA, and random projection for all 94 datasets.

[1] github.com/hj-n/steadiness-cohesiveness

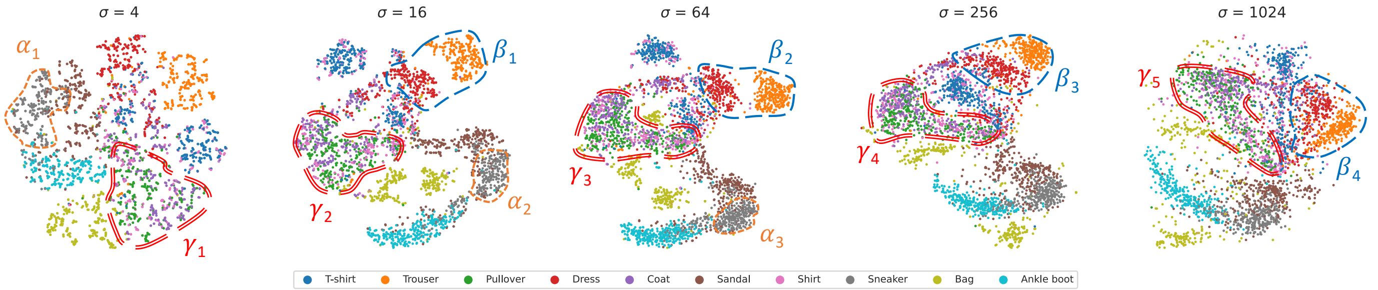

Fig. 6. t-SNE embeddings of Fashion-MNIST [72] data with diverse perplexity (σ) values. Combined with the class-pairwise CLM of the original dataset (Fig. 8), the patterns in the embeddings qualitatively support the findings about the effect of σ revealed by Label-T&C (Fig. 7; Sect. 6.1).

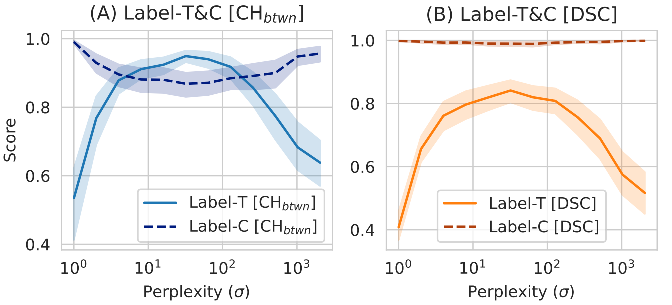

Fig. 7. Overall reliability of t-SNE embeddings according to the σ value quantified by Label-T&C [DSC] and Label-T&C [CH_{btwn}]. For each σ value, we average the score of the embeddings generated from 94 labeled datasets (95% confidence interval shaded).

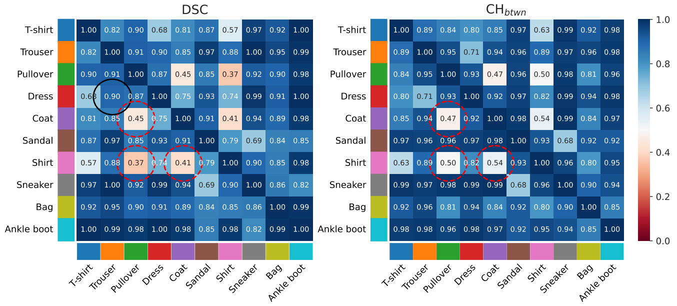

Fig. 8. Heatmaps detailing the CLM matrix of the Fashion-MNIST dataset ($M(X)$ in Sect. 4.1). The color of each cell depicts the CVM (DSC, CH_{btwn}) score measured for each pair of classes corresponding to rows and columns.

We check the overall execution time applying all measures to the embeddings, adding up the running times of the measures run in pairs (Label-T&C, T&C, MRRE, S&C, and CA-T&C). We use the provided implementation for S&C and scikit-learn [51] for the Silhouette. We implement the remaining measures in Python with Numba parallel computing [36] to maximize the scalability. We run the experiments on a Linux server with 40-core Intel Xeon Silver 4210 CPUs.

5.2.2 Results and Discussion

Fig. 5 show that the running time of Label-T&C highly depends on the CVM. Among all measures, DSC is the fastest, followed by Label-T&C [DSC]. If CH_{btwn} is used as the CVM, Label-T&C becomes less scalable. Still, Label-T&C [CH_{btwn}] has scalability similar to local (T&C, MRRE) and global (KL Divergence, DTM) measures and to CA-T&C, all being more than twice faster than S&C.

6 Case Studies

We report two case studies demonstrating the usefulness of Label-T&C to characterize DR techniques and their hyperparameters.

6.1 Examining the Effect of *t*-SNE Perplexity

6.1.1 Objectives and Design

We want to use Label-T&C to evaluate the reliability of the cluster structures from t-SNE embeddings (Sect. 5.1) depending on its perplexity hyperparameter σ. σ adjusts the balance between local and global cluster structures [13, 66]. We generate the t-SNE embeddings of the 94 labeled datasets used for the scalability analysis (Sect. 5.2) using different σ values ($\sigma \in \{2^i \mid i = 0, \cdots, 10\}$) and evaluate them using Label-T&C [CH_{btwn}] and Label-T&C [DSC]. We also inspect the t-SNE embeddings of the Fashion-MNIST [72] dataset with various perplexity values ($\sigma \in \{4, 16, 64, 256, 1024\}$; Fig. 6) to gain more qualitative insights. Moreover, we compute the "ground-truth" CLM matrix of the Fashion-MNIST dataset (Fig. 8), where the (i, j)-th cell represents the CVM score (CH_{btwn} or DSC) of the i-th and j-th classes. Note that this CLM matrix is identical to $M(\mathbf{X})$ in Sect. 4.1.

6.1.2 Results and Discussions

In the case of Label-T&C [CH_{btwn}], we found a clear tradeoff between Label-T and Label-C (Fig. 7A). When σ is low or high, Label-T [CH_{btwn}] gives low scores to t-SNE embeddings, indicating more False Groups distortions, while Label-C [CH_{btwn}] gives high scores, meaning fewer Missing Groups distortions. This means that t-SNE underrepresents the extent to which classes are clustered. In contrast, when σ has an intermediate value, Label-T&C [CH_{btwn}] indicate more Missing Groups and fewer False Groups distortions; hence, t-SNE exaggerates the degree to which classes are clustered.

These results align well with the intent of σ. With low σ, t-SNE focuses more on a small number of neighbors, likely fewer than the clusters' sizes, interpreting each cluster as made of loosely-connected components in the data space. Thus, the embedding is more likely to split classes into several clusters in the embedding. This phenomenon occurs in the Fashion-MNIST embedding (Fig. 6); the *Sneaker* class is less dense if σ is low (region α_1) and relatively condensed when σ has intermediate values (α_2 and α_3). For the latter, the number of neighbors that t-SNE focuses on will likely match the size of natural clusters within the original data. Therefore, t-SNE embeddings will tend to dismiss the inter-cluster connections, exaggerating the between-cluster distances. The number of neighbors that t-SNE focuses on with high σ values will likely be bigger than the clusters' sizes. Thus, t-SNE will detect all data clusters as one densely-packed component and generate embeddings with smaller inter-cluster distances.

The relation between the *Trouser* and *Dress* classes of the Fashion-MNIST embeddings (Fig. 6) qualitatively verifies these hypotheses. Their DSC scores are almost maximum (the black circle in Fig. 8), meaning they slightly overlap in the data space. However, their distance in the embedding is exaggerated with intermediate σ (β_1 and β_2) compared to high σ (β_3 and β_4). The same effect was observed qualitatively by Jeon et al. [29] while Label-T&C does so quantitatively.

Meanwhile, Label-C [DSC] decreases slightly for intermediate values of σ (Fig. 7B dotted line). As Label-T&C [DSC] focuses more on class overlaps and less on between-class distances compared to Label-T&C [CH_{btwn}] (see D and E in Sect. 5.1), it indicates that t-SNE preserves well the extent to which classes overlap regardless of σ. To

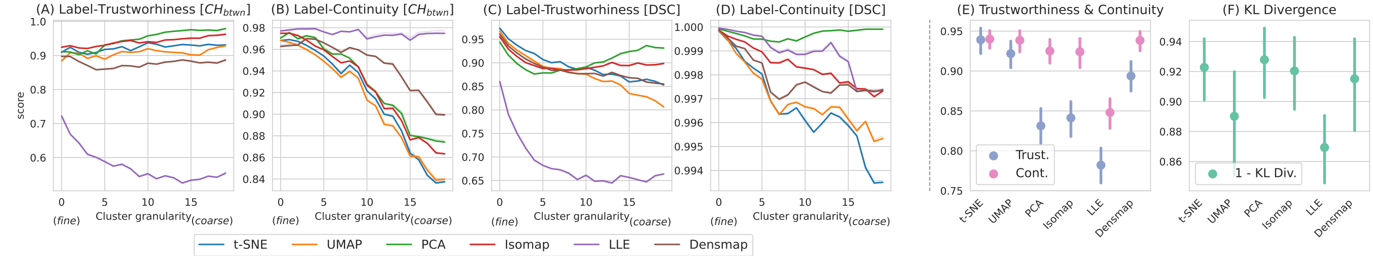

Fig. 9. CLM distortion evaluation of a linear (PCA) and five nonlinear (t-SNE, UMAP, Isomap, LLE, and Densmap) unsupervised DR techniques. (A-D) Evaluation results with Label-T&C [CH_{btwn}/DSC] where class labels are obtained from the hierarchical clustering of the original data at multiple granularity levels (x-axis). Label-T&C evaluates more coarse-grained (global) clusterings for higher levels. See details in Sect. 6.2. (E-F) Evaluation results of the techniques with T&C (E) and KL Divergence (F). Note that for all figures, higher scores indicate better embeddings.

quantitatively validate these findings, we searched for the overlapped classes within the CLM matrices, assuming that t-SNE accurately depicts class overlap for all σ values. We observed that the *Pullover*, *Coat*, and *Shirt* classes overlap in the high-dimensional space (red circles in Fig. 8; both their DSC and CH_{btwn} class-pairwise scores are low). We found that these classes overlap in all embeddings in Fig. 6 (γ_1 to γ_5), confirming our assumption.

In summary, we can conclude that for non-overlapping classes in t-SNE embeddings, the amount of proximity between them depends essentially on σ and is not indicative of the proximity of these classes in the data space: t-SNE is not trustworthy regarding the original distance between visually separated classes. However, classes with strong overlaps in the data are depicted as overlapping in the embedding too: t-SNE is more trustworthy for overlapping classes. Such results align with the qualitative findings of Wattenberg et al. [66].

Overall, these findings demonstrate the effectiveness of Label-T&C to enhance our understanding of the effect of σ on t-SNE results. We conduct the same analysis utilizing the competitor measures we used in our evaluation (Sect. 5.1); refer to Appendix E for the results.

6.2 Analyzing DR Techniques' Performance in Detail

6.2.1 Objectives and Design

We use Label-T&C to analyze the quality of unsupervised DR techniques across fine-grained to coarse-grained cluster structures. We embed each of the previous 94 datasets using six DR techniques: t-SNE, PCA, UMAP, Isomap, LLE, and Densmap [48]. We also apply hierarchical clustering, getting 20 clustering partitions with different granularity levels for each of these datasets. The levels of granularity are obtained by thresholding the pairwise distances computed by Ward linkage [33] into 20 equal ranges. We use Label-T&C [CH_{btwn}] and Label-T&C [DSC] to evaluate the embeddings using each of the 20 clusterings as class labels.

We also want to check whether the results obtained by Label-T&C align with the ones made by previous measures. We thus evaluate the embeddings using T&C and KL divergence as representative local and global measures, respectively. We use the same hyperparameter setting with the sensitivity analysis (Sect. 5.1).

6.2.2 Results and Discussions

Fig. 9 depicts the results. LLE generates few Missing Groups distortions (highest Label-C score; Fig. 9B, D) at any level, but more False Groups distortions as the granularity level increases (Label-T decreases; Fig. 9A, C). This finding aligns with the fact that LLE obtains the worst KL divergence score among all techniques (Fig. 9F). Such results are coherent with how LLE works, trying to reconstruct the "local patches" consisting of each point and its nearest neighbors while neglecting the overlap between the patches.

There is a Label-C downward trend across all other techniques as the level increases, while Label-C [DSC] shows higher scores than Label-C [CH_{btwn}] (Fig. 9B, D). This implies that Missing Groups distortions generally occur more for coarse-grained structures than for fine-grained ones; DR techniques exaggerate the separation between clusters at a global level. t-SNE and UMAP especially give the worst Label-C scores because they focus on the preservation of local neighborhoods, casting doubts on their reliability in identifying global clusters. T&C and KL divergence score provide strong evidence to the reliability of that claim. t-SNE and UMAP are in the top-2 highest ranks for T&C but fail to do so for KL divergence.

For Label-T&C except Label-C [CH_{btwn}], PCA gets the best score at higher granularity, suggesting that PCA is more reliable to conduct global tasks such as the density and similarity identification of clusters. These results align with the fact that PCA earns the best score for KL divergence. The phenomenon confirms the experimental observation made by Xia et al. [69]. This is also coherent with the fact that PCA embeds the data along the top two principal axes that preserve most of their variance, better representing coarse-grained structures than fine-grained ones.

We also find that Densmap, which is a variant of UMAP better preserving cluster density [48], gets worse Label-T [CH_{btwn}] scores than UMAP (Fig. 9A) but better Label-C [CH_{btwn}] scores (Fig. 9B), at all levels. This means that Densmap generates fewer Missing Groups but more False Groups distortions than UMAP. As Densmap approximately maintains the cluster locations of UMAP [48], such difference indicates that the clusters generally become bigger in Densmap compared to UMAP, hence the cluster density is relatively lower. Meanwhile, Densmap gets better Label-T&C [DSC] scores than UMAP for high granularity levels, confirming Densmap's advantage in investigating the overlap of clusters. The result is consistent with the KL divergence scores, indicating Densmap's advantage in preserving global structures when compared to UMAP (Fig. 9F).

These findings confirm the ability of Label-T&C to reveal the characteristics of DR methods over a wide range of clustering granularities. Although typical evaluation approaches of DR quality using both local and global measures (Fig. 9E, F) [19, 30, 46] show consistent results, they cannot reveal how the quality changes across granularity levels, as different measures are incomparable.

7 Conclusions

The general process of label-based DR evaluation relies on the assumption that the original data has good CLM, which can lead to erroneous conclusions when this assumption is violated. We introduce two new distortion measures—Label-Trustworthiness and Label-Continuity (Label-T&C)—that use class labels for DR evaluation while eliminating the need to check the validity of the CLM assumption. Our quantitative experiments show that Label-T&C outperforms previous DR measures in terms of precision and sensitivity in detecting Missing and False Groups distortions. Use cases show that Label-T&C can be used to characterize DR techniques and their hyperparameters.

As future work, we will study new CVM to make Label-T&C more sensitive to the CLM distortions than using DSC or CH_{btwn}. Enriching the embedding with CLM distortions [40] could also better inform analysts about the credibility of visual patterns. Yet another direction would be to evaluate supervised DR techniques with Label-T&C. We also believe that supervised DR techniques using class labels in their optimization process could benefit from incorporating Label-T&C in their loss function. Overall, our proposal aims toward getting more trustworthy DR-based visual analysis.

Acknowledgments

This work was supported by NAVER Corporation (Cloud Data Box) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C200520911). The authors thank Seokhyeon Park for his valuable feedback in improving the figures. The authors also appreciate Sungbok Shin and SNU HCIL members for their feedback in improving the language.

References

[1] M. Angelini, G. Blasilli, S. Lenti, A. Palleschi, and G. Santucci. Effectiveness error: Measuring and improving radviz visual effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4770–4786, 2022. doi: 10.1109/TVCG.2021.3104879

[2] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing*, 70(7):1304–1330, 2007. doi: 10.1016/j.neucom.2006.11.018

[3] M. Aupetit. Sanity check for class-coloring-based evaluation of dimension reduction techniques. In *Proc. of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, p. 134–141, 2014. doi: 10.1145/2669557.2669578

[4] M. Aupetit and A. Ali. Classsplom – a scatterplot matrix to visualize separation of multiclass multidimensional data, 2022. doi: 10.48550/ARXIV.2201.12822

[5] M. Aupetit, A. Ali, A. Baggag, and H. Bensmail. Classmat: a matrix of small multiples to analyze the topology of multiclass multidimensional data. In *2022 Topological Data Analysis and Visualization (TopoInVis)*, pp. 70–80, 2022. doi: 10.1109/TopoInVis57755.2022.00014

[6] M. Aupetit and T. Catz. High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing*, 63:139–169, 2005. New Aspects in Neurocomputing: 11th European Symposium on Artificial Neural Networks. doi: 10.1016/j.neucom.2004.04.009

[7] M. Aupetit, N. Heulot, and J.-D. Fekete. A multidimensional brush for scatterplot data analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 221–222, 2014. doi: 10.1109/VAST.2014.7042500

[8] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019. doi: 10.1038/nbt.4314

[9] R. Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966. doi: 10.1126/science.153.3731.34

[10] S. Ben-David and M. Ackerman. Measures of clustering quality: A working set of axioms for clustering. In *Advances in Neural Information Processing Systems*, vol. 21. Curran Associates, Inc., 2008. doi: 10.5555/2981780.2981796

[11] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pp. 217–235. Springer, 1999. doi: 10.1007/3-540-49257-7_15

[12] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proc. of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, BELIV '14, p. 1–8. ACM, New York, NY, USA, 2014. doi: 10.1145/2669557.2669559

[13] Y. Cao and L. Wang. Automatic selection of t-sne perplexity, 2017. doi: 10.48550/ARXIV.1708.03229

[14] L. D. Caro, V. Frias-Martinez, and E. Frias-Martinez. Analyzing the role of dimension arrangement for data visualization in radviz. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 125–132. Springer, 2010. doi: 10.1007/978-3-642-13672-6_13

[15] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011. doi: 10.1007/s10208-011-9098-0

[16] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017.

[17] B. Colange, J. Peltonen, M. Aupetit, D. Dutykh, and S. Lespinats. Steering distortions to preserve classes and neighbors in supervised dimensionality reduction. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 13214–13225. Curran Associates, Inc., 2020. doi: 10.5555/3495724.3496832

[18] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909

[19] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182

[20] R. Etemadpour, L. Linsen, C. Crick, and A. Forbes. A user-centric taxonomy for multidimensional data projection tasks. In *Proc. of the 6th International Conference on Information Visualization Theory and Applications*, pp. 51–62, 2015. doi: 10.5220/0005313400510062

[21] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94, 2015. doi: 10.1109/TVCG.2014.2330617

[22] S. G. Fadel, F. M. Fatore, F. S. Duarte, and F. V. Paulovich. Loch: A neighborhood-based multidimensional projection technique for high-dimensional sparse spaces. *Neurocomputing*, 150:546–556, 2015. Special Issue on Information Processing and Machine Learning for Applications of Engineering Solving Complex Machine Learning Problems with Ensemble Methods Visual Analytics using Multidimensional Projections. doi: 10.1016/j.neucom.2014.07.071

[23] I. Färber, S. Günnemann, 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*, p. 1, 2010.

[24] D. Francois, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007. doi: 10.1109/TKDE.2007.1037

[25] T. Fujiwara, Y.-H. Kuo, A. Ynnerman, and K.-L. Ma. Feature learning for dimensionality reduction toward maximal extraction of hidden patterns, 2022. doi: 10.48550/ARXIV.2206.13891

[26] N. Heulot, M. Aupetit, and J. D. Fekete. Proxilens: Interactive exploration of high-dimensional data using projections. In *EuroVis Workshop on Visual Analytics using Multidimensional Projections*. The Eurographics Association, Jun 2013. doi: 10.2312/PE.VAMP.VAMP2013.011-015

[27] H. Jeon, M. Aupetit, S. Lee, H.-K. Ko, Y. Kim, and J. Seo. Distortion-aware brushing for interactive cluster analysis in multidimensional projections, 2022. doi: 10.48550/ARXIV.2201.06379

[28] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo. Sanity check for external clustering validation benchmarks using internal validation measures, 2022. doi: 10.48550/ARXIV.2209.10042

[29] H. Jeon, H.-K. Ko, J. Jo, Y. Kim, and J. Seo. Measuring and explaining the inter-cluster reliability of multidimensional projections. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):551–561, 2021. doi: 10.1109/TVCG.2021.3114833

[30] H. Jeon, H.-K. Ko, S. Lee, J. Jo, and J. Seo. Uniform manifold approximation with two-phase optimization, 2022. doi: 10.1109/VIS54862.2022.00025

[31] Z. Ji and H. Ji. Discussion of "exponential-family embedding with application to cell developmental trajectories for single-cell rna-seq data". *Journal of the American Statistical Association*, 116(534):471–474, 2021. doi: 10.1080/01621459.2021.1886106

[32] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571, 2011. doi: 10.1109/TVCG.2011.220

[33] J. H. W. Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845

[34] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):142–151, 2018. doi: 10.1109/TVCG.2017.2745085

[35] C.-H. Lai, M.-F. Kuo, Y.-H. Lien, K.-A. Su, and Y.-S. Wang. Parametric dimension reduction by preserving local structure. pp. 75–79, 2022. doi: 10.1109/VIS54862.2022.00024

[36] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015. doi: 10.1145/2833157.2833162

[37] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer-Verlag New York, 2007. doi: 10.1007/978-0-387-39351-3

[38] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4:538–547, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011. doi: 10.1016/j.procs.2011.04.056

[39] J. A. Lee and M. Verleysen. Two key properties of dimensionality reduction methods. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 163–170, 2014. doi: 10.1109/CIDM.2014.7008663

[40] S. Lespinats and M. Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. *Computer Graphics Forum*, 30(1):113–125, 2011. doi: 10.1111/j.1467-8659.2010.01835.x

[41] S. Lespinats, M. Verleysen, A. Giron, and B. Fertil. Dd-hds: A method for visualization and exploration of high-dimensional data. *IEEE Transactions on Neural Networks*, 18(5):1265–1279, 2007. doi: 10.1109/TNN.2007.891682

[42] 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*, pp. 911–916, 2010. doi: 10.1109/ICDM.2010.35

[43] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu. Understanding and enhancement of internal clustering validation measures. *IEEE Transactions on Cybernetics*, 43(3):982–994, 2013. doi: 10.1109/TSMCB.2012.2220543

[44] R. M. Martins, D. B. Coimbra, R. Minghim, and A. Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, 41:26–42, 2014. doi: 10.1016/j.cag.2014.01.006

[45] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. doi: 10.48550/arXiv.1802.03426

[46] M. Moor, M. Horn, B. Rieck, and K. Borgwardt. Topological autoencoders. In H. D. III and A. Singh, eds., *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7045–7054. PMLR, 13–18 Jul 2020.

[47] R. Motta, R. Minghim, A. de Andrade Lopes, and M. Cristina F. Oliveira. Graph-based measures to assist user assessment of multidimensional projections. *Neurocomputing*, 150:583–598, 2015. doi: 10.1016/j.neucom.2014.09.063

[48] A. Narayan, B. Berger, and H. Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature biotechnology*, 39(6):765–774, 2021. doi: 10.1038/s41587-020-00801-7

[49] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). 1996.

[50] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[52] N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eisemann, and A. Vilanova. Hierarchical stochastic neighbor embedding. *Computer Graphics Forum*, 35(3):21–30, 2016. doi: 10.1111/cgf.12878

[53] G. J. Quadri, J. A. Nieves, B. Wiernik, and P. Rosen. Automatic scatterplot design optimization for clustering identification. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16, 2022. doi: 10.1109/TVCG.2022.3189883

[54] G. J. Quadri and P. Rosen. Modeling the influence of visual density on cluster perception in scatterplots using topology. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1829–1839, 2021. doi: 10.1109/TVCG.2020.3030365

[55] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7

[56] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632

[57] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013. doi: 10.1109/TVCG.2013.153

[58] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012. doi: 10.1111/j.1467-8659.2012.03125.x

[59] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009. doi: 10.1111/j.1467-8659.2009.01467.x

[60] T.Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101

[61] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[62] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006. doi: 10.1016/j.neunet.2006.05.014

[63] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(13):451–490, 2010. doi: 10.5555/1756006.1756019

[64] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, p. 1073–1080. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1553374.1553511

[65] Y. Wang, K. Feng, X. Chu, J. Zhang, C.-W. Fu, M. Sedlmair, X. Yu, and B. Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(5):1828–1840, 2018. doi: 10.1109/TVCG.2017.2701829

[66] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002

[67] J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, S. Leman, and C. North. Towards a systematic combination of dimension reduction and clustering in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):131–141, 2018. doi: 10.1109/TVCG.2017.2745258

[68] 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*, KDD '09, p. 877–886. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1557019.1557115

[69] J. Xia, L. Huang, W. Lin, X. Zhao, J. Wu, Y. Chen, Y. Zhao, and W. Chen. Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2022. doi: 10.1109/TVCG.2022.3209423

[70] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3114694

[71] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, and X. Chen. A comparison for dimensionality reduction methods of single-cell rna-seq data. *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.646936

[72] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. doi: 10.48550/arXiv.1708.07747

[73] Y. Yang, H. Sun, Y. Zhang, T. Zhang, J. Gong, Y. Wei, Y.-G. Duan, M. Shu, Y. Yang, D. Wu, et al. Dimensionality reduction by umap reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell reports*, 36(4):109442, 2021. doi: 10.1016/j.celrep.2021.109442

[74] A. Zubaroğlu and V. Atalay. Online embedding and clustering of data streams. In *Proceedings of the 2019 3rd International Conference on Big Data Research*, ICBDR 2019, p. 142–146. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3372454.3372481