(Appendix) *Classes are not Clusters*: Improving Label-based Evaluation of Dimensionality Reduction

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

A Theorems and Proofs

In Sect. 4.2.2, we need to examine whether representative Internal Validation Measures (IVMs; e.g., Silhouette Coefficient [9], Davies-Bouldin index [2], and Distance Consistency (DSC) [10]) satisfy invariance requirements to be proper clustering validation measures to compute Label-T&C (**R1** to **R3** in Sect. 4.2.1).

A.1 Proofs on Silhouette & Davies-Bouldin Index

We first prove that the Silhouette Coefficient and Davies-Bouldin index do not satisfy requirement R2; hence, they cannot be used for evaluating CLM distortions.

Theorem 1. *Silhouette Coefficient does not satisfy shift invariance (R2).*

Proof. For any input clustering partition $\mathbf{P}$ over data $\mathbf{S}$ and Euclidean distance function d, Silhouette Coefficient SC is defined as:

$$SC(\mathbf{P},\mathbf{S},d) = \frac{1}{|\mathbf{P}|}\cdot\sum_{i=1}^{|\mathbf{P}|}\frac{1}{|P_i|}\sum_{x\in P_i}\frac{b(x,d)-a(x,d)}{\max(b(x,d),a(x,d))}.$$

For a data point $x \in P_i$,

$$a(x,d) = \frac{1}{|P_i|-1}\sum_{y\in P_i, x\neq y} d(x,y),$$

and

$$b(x,d) = \min_j \frac{1}{|P_j|}\sum_{y\in P_j} d(x,y).$$

As $a(x,d+\beta) = a(x,d)+\beta$ and $b(x,d+\beta) = b(x,d)+\beta$,

$$\begin{aligned} SC(\mathbf{P},\mathbf{S},d+\beta) &= \frac{1}{|\mathbf{P}|}\cdot\sum_{i=1}^{|\mathbf{P}|}\frac{1}{|P_i|}\sum_{x\in P_i}\frac{b(x)-a(x)}{\beta+\max(b(x),a(x))} \\ &\neq SC(\mathbf{P},\mathbf{S},d). \end{aligned}$$

Thus, Silhouette Coefficient SC is not shift invariant. □

Theorem 2. *Davies-Bouldin index does not satisfy shift invariance (R2).*

Proof. For any input clustering partition $\mathbf{P}$ over data $\mathbf{S}$ and Euclidean distance function d, Davies-Bouldin index DB is defined as:

$$DB(\mathbf{P},\mathbf{S},d) = \frac{1}{|\mathbf{P}|}\sum_{i=1}^{|\mathbf{P}|}\max_{j\neq i}\frac{\frac{1}{|P_i|}\sum_{x\in P_i} d(x,c_i)+\frac{1}{|P_j|}\sum_{x\in P_j} d(x,c_j)}{d(c_i,c_j)},$$

where c_n denotes the centroid of P_n in $\mathbf{S}$.

For any $\mathbf{P}$ and $\mathbf{S}$,

$$\begin{aligned} &DB(\mathbf{P},\mathbf{S},d+\beta) \\ &= \frac{1}{|\mathbf{P}|}\sum_{i=1}^{|\mathbf{P}|}\max_{j\neq i}\frac{\frac{1}{|P_i|}\sum_{x\in P_i} d(x,c_i)+\frac{1}{|P_j|}\sum_{x\in P_j} d(x,c_j)+2\beta}{d(c_i,c_j)+\beta} \\ &\neq DB(C,X,d). \end{aligned}$$

Thus, Davies Bouldin index DB is not shift invariant. □

A.2 Proofs on Distance Consistency (DSC)

We prove that DSC satisfies all four requirements, and thus can be used as a proper CVM for Label-T&C. For input clustering partition $\mathbf{P}$ over data $\mathbf{S}$ and Euclidean distance function d, DSC is defined as:

$$DSC(\mathbf{P},\mathbf{S},d) = \frac{|\{x\in S | x\in \arg\min_{P_i\in\mathbf{P}} d(x,c_i)\}|}{|\mathbf{S}|},$$

where c_i denotes the centroid of P_i in $\mathbf{S}$. DSC counts the proportion of data that belong to the same class as their nearest class centroid. As DSC does not require any hyperparameter other than d, it satisfies R4. Now, we provide proofs for R1, R2, and R3.

Theorem 3. *DSC satisfy scale invariance (R1).*

Proof. $\forall\, x\in\mathbf{S}$, $\arg\min_{P_i\in\mathbf{P}}\alpha d(x,c_i) = \arg\min_{P_i\in\mathbf{P}} d(x,c_i)$. Thus, $\forall\, x\in\mathbf{S}$, $DSC(\mathbf{P},\mathbf{S},\alpha d) = DSC(\mathbf{P},\mathbf{S},d)$. Therefore, DSC satisfies scale invariance. □

Theorem 4. *DSC satisfies shift invariance (R2).*

Proof. $\forall\, x\in\mathbf{S}$, $\arg\min_{P_i\in\mathbf{P}}(d+\beta)(x,c_i) = \arg\min_{P_i\in\mathbf{P}}(x,c_i)$. Thus, $\forall\, x\in\mathbf{S}$, $DSC(\mathbf{P},\mathbf{S},d+\beta) = DSC(\mathbf{P},\mathbf{S},d)$. Therefore, DSC satisfies shift invariance. □

Theorem 5. *DSC satisfies range invariance (R3).*

Proof. DSC ranges from 0.5 to 1, where the higher value denotes better clustering. Thus, DSC is range invariant. □

Note that the satisfaction of the requirements of the between-dataset Calinski-Harabasz index (CH_{btwn}), which we use as another proper option for Label-T&C, has been previously proven by Jeon et al. [5].

B Additional Discussions on Sensitivity Analysis

In Sect. 5.1.3, we especially focus on discussing the difference between the patterns shown by Label-T&C and the general process of label-based evaluation (Silhouette, DSC). Here, we provide detailed discussions on the remaining measures: local measures (T&C, MRRE), global measures (KL-divergence, DTM), cluster-level measures (S&C), and CA-T&C. Note that this discussion is linked with Sect 5.1.2, 5.1.3, and Fig. 5 in the main document.

Local measures The analysis of the results indicate that local measures are able to *detect* the cluster-level distortions, but have no capability to distinguish False and Missing Groups distortions. For all experiments, all local measures decreased together regardless of the type of distortions generated for the experiment. Such results indicate that the occurrence of cluster-level distortions (Missing and False Groups distortions) also make Missing and False Neighbors distortions occur

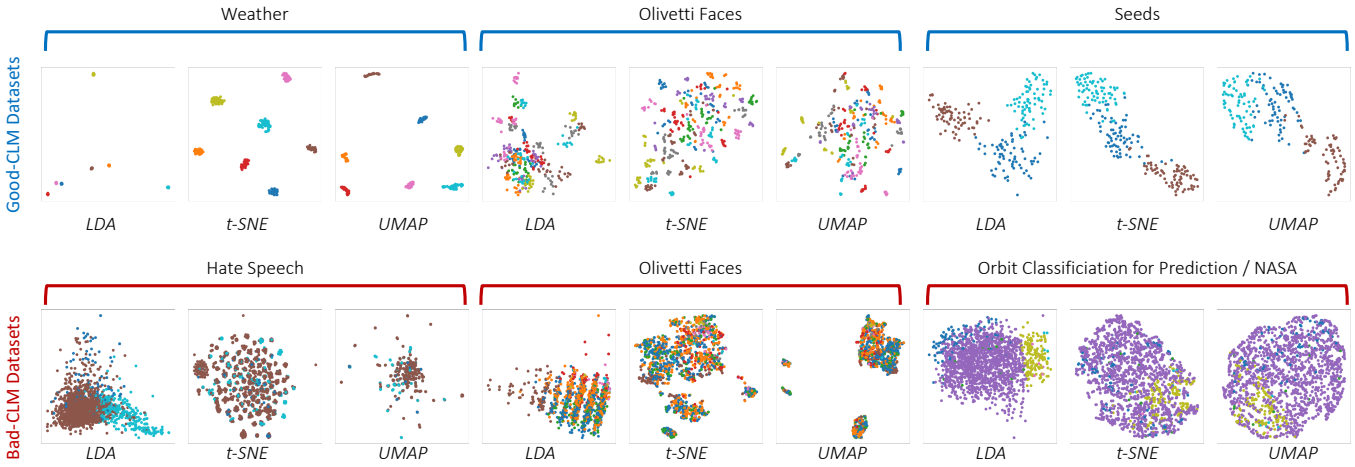

Fig. 1: The subset of the LDA, t-SNE, and UMAP embeddings used in the preliminary experiment (Sect. 3.3). The titles placed over the brackets depict the dataset name. If the CLM of the original dataset is good, CLM of the embeddings made by three DR techniques is all good. However, when the CLM of the original dataset is bad, LDA tends to separate class labels more than the other two unsupervised learning techniques.

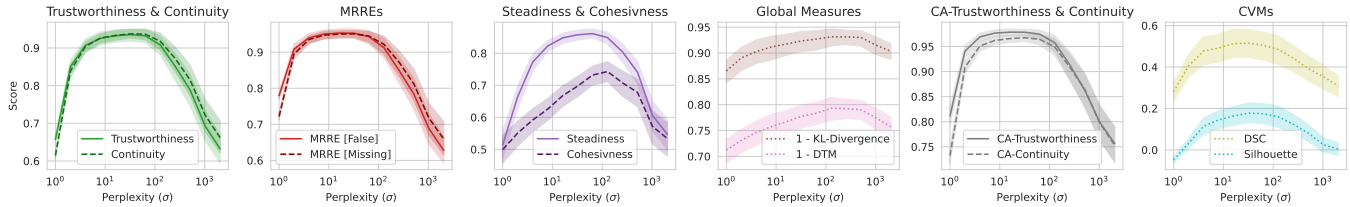

Fig. 2: Overall reliability of t-SNE embeddings according to the perplexity (σ) value, assessed by competitor distortions measures we used in our evaluations (Sect. 5). Unlike Label-T&C, which showed a clear tradeoff between Label-T and Label-C, all competitor measures promoted an intermediate range of σ.

simultaneously. Seeking how the distortions at different scales (cluster and neighbor) interplay with each other will be interesting future work.

Global measures Global measures often fail to detect False and Missing Groups distortions; for example, in experiments A and D, global measure scores go up while distortions increase. We can thus conclude that using global measures is not suitable to detect cluster-level distortions.

Steadiness & Cohesiveness (S&C) While Label-T&C precisely pinpoint Missing and False Group distortions, S&C—the only pair of measures that directly focus on False and Missing Group—failed to do so. This is mainly because S&C adopts Shared-Nearest-Neighbor (SNN) [6] as a default distance function. SNN is constructed based on k-nearest neighbors (kNN) graph. It assigns higher similarity to point pairs that share more kNNs. Therefore, adding more points between two points (i.e., fewer shared neighbors) increases the SNN distance even though the Euclidean distance between the two points does not change. This makes not only Steadiness but also Cohesiveness decrease in Experiment A-C; the overlap between low-dimensional data points made the SNN distances between the points within the same classes grow, which lead S&C to interpret that not only compression but also stretching occurred. In experiments D-F, such unintentional distance growth occurred in the original space, leading to inaccurate performances of S&C.

Class-Aware Trustworthiness & Continuity (CA-T&C) Patterns shown by CA-T&C are generally similar to T&C in experiments A to C. Such results indicate that CA-T&C can detect False Groups distortions. However, CA-T&C fails to accurately detect Missing Groups distortions in experiments D to F; in experiment E, the CA-Continuity score even increases while distortions increase. This indicates that CA-T&C can only detect False Groups distortions while hardly capturing Missing Groups, which is well aligned with the design of CA-T&C. Such results are well aligned with the design of CA-T&C, which genuinely detect the degradation of CLM but not the increment during the reduction, which is described in 2.1.3.

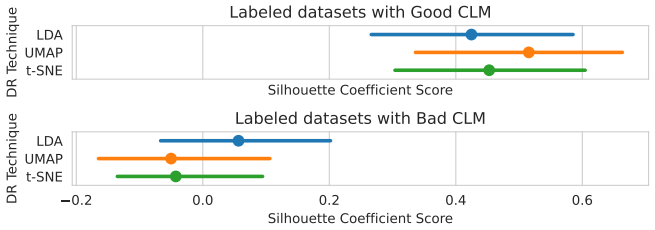

Fig. 3: The evaluation of three DR techniques (LDA, t-SNE, UMAP) using the general process of assessing DR based on class labels. We gathered good-CLM and bad-CLM datasets, applied DR techniques, and then measured their quality using the Silhouette Coefficient in the embedded space. While UMAP and t-SNE showed competitive or better performance compared to LDA with good-CLM datasets (top), they fell behind LDA with bad-CLM datasets (bottom). The error bars depict 95% confidence intervals. As LDA tries to maximize class separation rather than preserving the intrinsic structure of the original data, such results indicate that the general label-based evaluation process can be incorrect with bad-CLM datasets. Refer to Appendix D to see the embeddings.

C Labeled Datasets

We provide 94 labeled datasets we used in our preliminary experiment (Sect 3.3), scalability analysis (Sect. 5.2), and the application examining t-SNE perplexity hyperparameter (Sect 6.1). Please refer to the full list of the labeled dataset in Table 1. The datasets are sorted in descending order based on CLM measured by CH_{btwn} developed by Jeon et al. [5].

Dataset	Points	Dims	Classes
Weather	365	192	
Olivetti Faces	400	4,096	40
MNIST64	1,082	64	6
Optical Recognition of Handwritten Digits	3,823	64	10
Seeds	210	7	3
Wireless Indoor Localization	2,000	7	4
COIL20	1,440	400	20
Iris	150	4	3
Pen-Based Recognition of Handwritten Digits	7,494	16	10
Rice Seed (Gonen&Jasmine)	18,185	10	2
Breast Cancer Wisconsin (Original)	683	9	2
pH-recognition	653	3	15
Echocardiogram	61	10	20
Fashion-MNIST	3,000	784	10
Mobile Price Classification	2,000	20	4
Human Stress Detection	2,001	3	3
Dry Bean	13,611	16	7
HAR	735	561	6
Rice Dataset Cammeo and Osmancik	3,810	7	2
Wine Customer Segmentation	178	13	3
Wine	178	13	3
Zoo	101	16	7
Image Segmentation	210	19	7
Boston	154	13	3
Statlog (Image Segmentation)	2,310	19	7
User Knowledge Modeling	258	5	4
Ecoli	336	7	8
Website Phishing	1,353	9	3
Date Fruit	898	34	7
Music Genre Classification	1,000	26	10
Pistachio	2,148	28	2
Crowdsourced Mapping	10,545	28	6
Raisin	900	7	2
Breast Cancer Wisconsin (Prognostic)	569	30	2
Yeast	1,484	8	10
Dermatology	358	34	6
Glass Identification	214	9	6
Classification in Asteroseismology	1,001	3	2
Breast Tissue	106	9	6
Mammographic Mass	830	5	2
Banknote Authentication	1,372	4	2
Birds Bones and Living Habits	413	10	6
ExtyaleB	319	30	5
Flickr Material Database	997	1,536	10
CNAE-9	1,080	856	9
Fetal Health Classification	2,126	21	3
Durum Wheat Features	9,000	236	3

Dataset	Points	Dims	Classes
Smoker Condition	1,012	7	2
Student Grade	395	29	2
Turkish Music Emotion	400	50	4
CIFAR10	3,250	1,024	10
Ionosphere	351	34	2
SPECTF Heart	80	44	2
Hate Speech	3,221	100	3
Predicting Pulsar Star	9273	8	2
Parkinsons	195	22	2
HTRU2	17,898	8	2
Siberian Weather Stats	1,439	11	9
Patient Treatment Classification	4,412	10	2
SMS Spam Collection	835	500	2
MAGIC Gamma Telescope	19,020	10	2
Orbit Classification For Prediction / NASA	1,748	11	6
Harberman's Survival	306	3	2
IMDB	3,250	700	2
Pumpkin Seeds	2,500	12	2
World12d	150	12	5
Heart Attack Analysis & Prediction	303	13	2
Diabetic Retinopathy Debrecen	1,151	19	2
Seismic Bumps	646	24	2
Hepatitis	80	19	2
Statlog (German Credit)	1,000	24	2
Wine Quality	4,898	11	7
Sentiment Labeled Sentences	2,748	200	2
Pima Indians Diabetes Database	768	8	2
Blood Transfusion Service Center	748	4	2
Heart Disease	297	13	5
Cardiovascular Study	2,927	15	2
Insurance Company Benchmark	5,822	85	2
Street View House Numbers	732	1,024	10
SkillCraft1 Master Table Dataset	3,338	18	7
HIVA	3,076	1,617	2
Spambase	4,601	57	2
Wilt	4,339	5	2
Breast Cancer Coimbra	116	9	2
SECOM	1,567	590	2
Customer Classification	1,000	11	4
Credit Risk Classification	976	11	2
Planning Relax	182	12	2
Taiwanese Bankruptcy Prediction	6,819	95	2
Labeled Faces in the Wild	2,200	5,828	2
Water Quality	2,011	9	2
Epileptic Seizure Recognition	5,750	178	5
Paris Housing Classification	10,000	17	2
Fraud Detection Bank	20,468	112	2

Table 1: 94 Labeled datasets sorted in descending order by CLM, gathered by the previous research [5]. The left table shows the top half datasets, while the right one shows the bottom half datasets. The bad-CLM datasets and the good-CLM datasets are used in the preliminary experiment (Sect 3.3) are emphasized with bold and bold with underline, respectively.

D Preliminary Experiment

D.1 Objectives and Design

To reveal the threat of the general label-based evaluation process, we conduct a preliminary experiment (Fig. 3) comparing three DR techniques (LDA [3], t-SNE, UMAP) based on that process. LDA is supervised; it takes predefined class labels as input and produces embeddings that maximize the separation among classes. In contrast, t-SNE and UMAP are unsupervised; they ignore class labels and try to preserve the original structure of the data. We deliberately choose these two techniques as they are previously shown to be more capable of capturing the cluster structure, compared to other widely used unsupervised DR techniques [12]. To evaluate the techniques, we construct two distinct sets of labeled datasets: one featuring 10 datasets with good CLM and another one comprising 10 datasets with poor CLM. We select the datasets from a collection of 96 datasets previously compiled in a related study [5], guided by their CLM rankings as provided within that research. We then generate embeddings of the datasets in the two sets using the three DR techniques and measure the embeddings' CLM using the Silhouette Coefficient. We used the Silhouette Coefficient as it is the most widely used CVM for visualization research. The detailed settings we used are as follows:

Datasets Good- and bad-CLM datasets are picked as top-10 and bottom-10 CLM datasets in the list of 94 datasets (Appendix C). Note that we only picked the datasets with more than two classes, as it is necessary to run LDA.

Hyperparameters For t-SNE and UMAP, we used the default hyper-

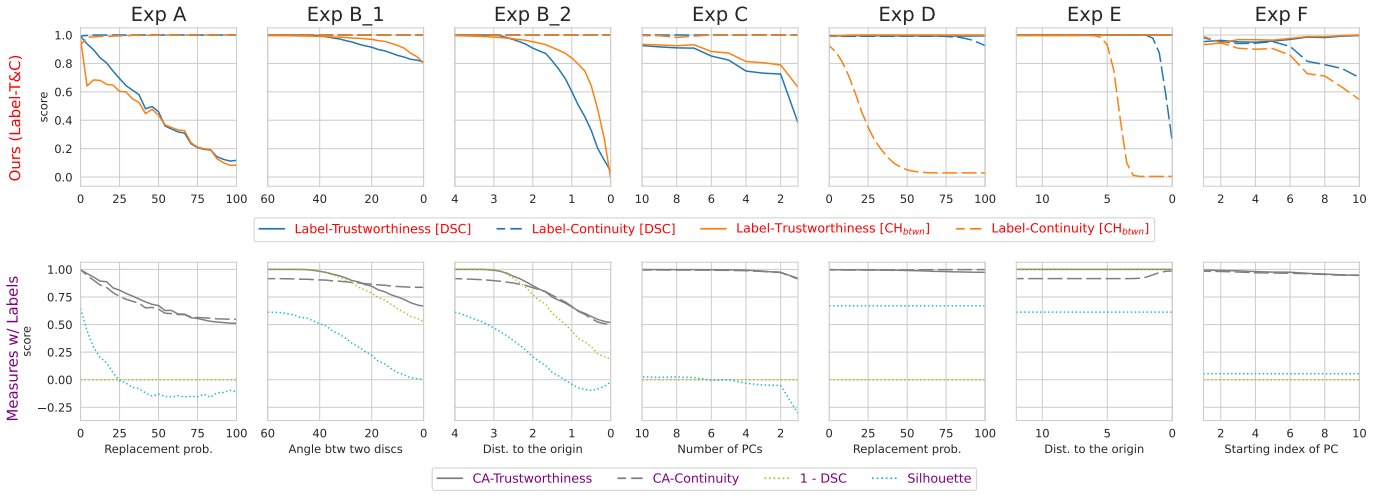

Fig. 4: The results of sensitivity analysis (Sect. 5.1) replicated with the class labels generated by HDBSCAN [7] clustering technique.

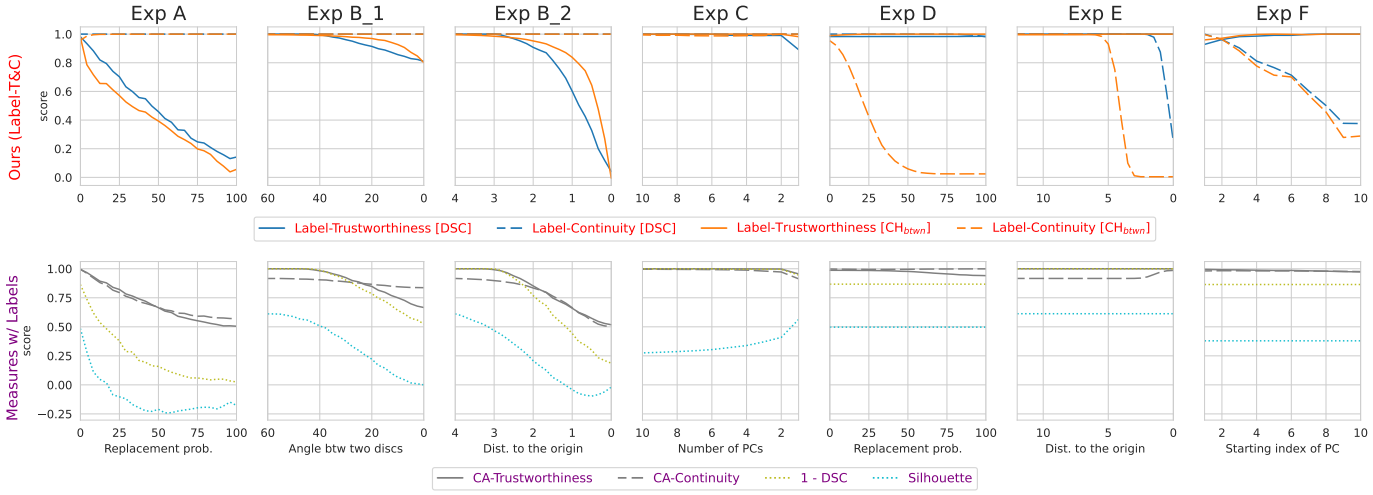

Fig. 5: The results of sensitivity analysis (Sect. 5.1) replicated with the class labels generated by K-Means [4] clustering technique.

parameter setting provided by `scikit-learn` [8] and `umap-learn`[1] library. We used the default setting for t-SNE and UMAP to compare them fairly against LDA, which has no hyperparameter and generates embeddings via a fixed algorithm.

DR Embeddings We depict the subset of the embeddings we used in the preliminary experiment in Fig. 1.

D.2 Results

The results (Fig. 3) demonstrate the issue of using the general label-based DR evaluation process. We expect a proper measure to always prefer UMAP and t-SNE to LDA, as LDA focuses on the linear separation of the classes by design and is thus less sensitive to the intrinsic structure of the original data. With good CLM datasets, the general process provides a better score for t-SNE and UMAP than for LDA as expected. However, with bad-CLM datasets, LDA outperforms both t-SNE and UMAP. This result shows that the general process of label-based DR evaluation erroneously assesses CLM distortions when the original data has bad CLM. Our goal in this work is to introduce a new way of using class labels for DR evaluation that mitigates such a bias.

[1] https://umap-learn.readthedocs.io/

E Examining the Effect of t-SNE Perplexity with Other Distortion Measures

Objectives and design As an extension of Sect. 6.1., we want to analyze the effect of t-SNE perplexity hyperparameter using the competitor distortion measures we used in our evaluations (Sect. 5). We conduct the same analysis using the following measures: local measures (T&C, MRRE), cluster-level measures (S&C), global measures (KL Divergence, DTM), CA-T&C, and the general process of label-based DR evaluation (i.e., Silhouette, DSC).

Results Fig. 2 depicts the results. We found that all distortion measures we considered assigned higher scores to the embeddings with intermediate σ than the ones with low and high σ, regardless of the target distortion (stretching or compression) evaluated in the experiment. The only insight we can obtain is that t-SNE with intermediate σ generates better embeddings in terms of both stretching and compression. Such results contradict the conclusion with Label-T&C: Missing and False Groups distortions have a clear tradeoff. Regarding our qualitative findings with Fashion-MNIST dataset [13] (Fig. 7 and 9) support the conclusion of Label-T&C, and also considering that Label-T&C precisely captures cluster-level distortions compared to competitors (Sect. 5.1), we believe that the conclusion with Label-T&C is more reliable than the one with competitors.

Acronym	Definition
DR	Dimensionality Reduction
CLM	Cluster-Label Matching
CVM	Clustering Validation Measures
Label-T&C	Label-Trustworthiness & Label-Continuity
KL Divergence	Kullback-Liebler Divergence
DTM	Distance-to-Measure
T&C	Trustworthiness & Continuity
MRREs	Mean Relative Rank Errors
S&C	Steadiness & Cohesiveness
CA-T&C	Class-Aware Trustworthiness & Continuity
IVM	Internal CVM
EVM	External CVM
DSC	Distance Consistency
CH_{btwn}	Between-Dataset Calinski-Harabasz Index

Table 2: The list of acronyms used in the paper and their definitions.

F Conducting Sensitivity Analysis with the Labels Generated by Clustering Techniques

Objectives and Design In Sect. 5.1, we conduct six experiments validating distortion measures' sensitivity in quantifying Missing and False Groups distortions. Here, we replicate the experiments while feeding the labels made by clustering techniques. To generate class labels, we use *K*-Means [4] and HDBSCAN [1], with the implementations provided by `scikit-learn` and McInnes et al. [7]. We exploit Bayesian optimization [11] to obtain optimal clustering results, using the hyperparameter range suggested by Jeon et al. [5].

Results Fig. 4 and Fig. 5 present the results with labels created by HDBSCAN [7] and *K*-Means [4], respectively. Overall, our findings show that Label-T&C and CA-T&C results consistently align with the original experiment (see Sect. 5.1; Fig. 4 in the main document). The results indicate that Label-T&C produces stable scores regardless of the change in class labels. However, the difference with the original experiment is that the amount of decrement made in experiments C and D decreases when the class labels are made with *K*-Means. Such results indicate that Label-T&C result is not completely independent of the CLM of the high-dimensional space. Examining the relationship between label characteristics and Label-T&C result will be interesting future work.

Note that for experiments B-1, B-2, and E, the result is exactly the same as the original experiment. This indicates that optimal clustering results made by *K*-Means and HDBSCAN exactly match the original class labels. Such a situation happens as we set individual hyperspheres, which are clearly separated, as classes.

However, the general process of label-based DR evaluation, such as DSC and Silhouette, fails to replicate these results. Particularly in experiments D-F, the DSC and Silhouette scores diverge substantially from those of the original experiment. These findings suggest that the general process of label-based DR evaluation is susceptible to changes in class labels. Therefore, our results underscore the need to consider high-dimensional data when evaluating DR embeddings. The results also highlight the potential pitfalls of employing a general label-based DR evaluation process.

G Acronyms

Here, we organize the acronyms and their long-form definitions used in the paper. Please refer to Table 2 for the list.

References

[1] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds., *Advances in Knowledge Discovery and Data Mining*, pp. 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-37456-2_14

[2] 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

[3] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x

[4] 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)*, 28(1):100–108, 1979.

[5] 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

[6] R. Liu, H. Wang, and X. Yu. Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Information Sciences*, 450:200–226, 2018. doi: 10.1016/j.ins.2018.03.031

[7] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017.

[8] 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.

[9] 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

[10] 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

[11] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012.

[12] 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

[13] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.