Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study

Jiazhi Xia, Yuchen Zhang, Jie Song, Yang Chen, Yunhai Wang and Shixia Liu

Fig. 1: Different dimensionality reduction results of the ExtYaleB dataset, which contains 320 images of five people. The twelve dimensionality reduction techniques fall into four categories: Non-linear & Local (NL&Lc), Non-linear & Global (NL&G), Linear & Local (L&Lc), and Linear & Global (L&G). Instances of task stimuli: (a) Cluster identification (T1), (b) Membership identification (T2), (c) Distance comparison (T3), and (d) Density comparison (T4). (e) lasso tool used in T1–T4.

Abstract—Dimensionality Reduction (DR) techniques can generate 2D projections and enable visual exploration of cluster structures of high-dimensional datasets. However, different DR techniques would yield various patterns, which significantly affect the performance of visual cluster analysis tasks. We present the results of a user study that investigates the influence of different DR techniques on visual cluster analysis. Our study focuses on the most concerned property types, namely the linearity and locality, and evaluates twelve representative DR techniques that cover the concerned properties. Four controlled experiments were conducted to evaluate how the DR techniques facilitate the tasks of 1) cluster identification, 2) membership identification, 3) distance comparison, and 4) density comparison, respectively. We also evaluated users’ subjective preference of the DR techniques regarding the quality of projected clusters. The results show that: 1) Non-linear and Local techniques are preferred in cluster identification and membership identification; 2) Linear techniques perform better than non-linear techniques in density comparison; 3) UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-Distributed Stochastic Neighbor Embedding) perform the best in cluster identification and membership identification; 4) NMF (Nonnegative Matrix Factorization) has competitive performance in distance comparison; 5) t-SNLE (t-Distributed Stochastic Neighbor Linear Embedding) has competitive performance in density comparison.

Index Terms—Dimensionality reduction, visual cluster analysis, perception-based evaluation

1 INTRODUCTION

Visual cluster analysis usually employs dimensionality reduction (DR) techniques to project high-dimensional data into 2D scatterplots, in which analysts can visually identify cluster patterns [53,59]. The cluster analysis is an inherent human-in-the-loop task that lacks a universal ground truth [6]. Therefore, visual cluster analysis, which assumes clusters in the projection are faithful representation matching with actual clusters, is widely used in evaluating automatic clustering algorithms [6] and clustering applications [7,21,33,58,60].

However, different DR techniques perform differently in a set of aspects in visual cluster analysis, e.g., cluster separation, membership preservation, distance preservation, and density preservation. As shown in Fig. 1, the twelve DR techniques generated various visual patterns out of the same dataset. Although each DR technique provides different insights into clusters, most of them are rarely used in real applications. For example, we have surveyed papers in the mainstream visualization journals and conferences from 2010 to 2020 with a focus on applying DR techniques in visual cluster analysis. Among the 51 resulted papers, we found that only PCA (Principal Component Analysis) [55], MDS (Multidimensional Scaling) [25], and t-SNE (t-Distributed Stochastic Neighbor Embedding) [47] are used for more than 2 times. A revisit to DR techniques is needed to guide the use of them for visual cluster analysis.

* Jiazhi Xia, Yuchen Zhang and Jie Song are with School of Computer Science and Engineering, Central South University, China. E-mail: {xiajiazhi,yuchenzhang,jiesong}@csu.edu.cn;
* Yang Chen is with I4 data. E-mail: chen1984yang@gmail.com.
* Yunhai Wang is with School of Computer Science and Technology, Shandong University, China. E-mail: cloudseawang@gmail.com.
* Shixia Liu is with School of Software, Tsinghua University, China. E-mail: shixia@tsinghua.edu.cn.
* Jiazhi Xia and Yang Chen are corresponding authors.

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

Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx


While a high-level view of the performances of DR techniques on visual cluster analysis is a common concern of researchers. When choosing DR techniques from a large set, analysts usually consider the property that the DR technique preserves. Among various properties, linearity and locality have gained the most attention [16-35]. For example, linear DR techniques are believed to outperform in preserving the density of clusters [11, 37]. On the other hand, non-linear DR techniques, such as t-SNE and UMAP (Uniform Manifold Approximation and Projection) [36], are widely used because of their performance in separating clusters. The locality property, i.e., if local or global pairwise relationships are preserved, also yields significant differences in the generated scatterplots [13, 16].

Nevertheless, many qualitative comparisons have been conducted towards understanding how different DR techniques perform in visual cluster analysis tasks [17, 49, 57, 61]. However, we argue that two significant gaps exist in current research. First, while the aforementioned high-level properties are of concern to both DR researchers and analysts, there lacks experimental studies that test and verify them against user’s visual perception in visual cluster analysis tasks. Second, none of the existing experimental studies has evaluated a set of DR techniques that can be representative of all the properties or cover the most popular DR techniques, calling for a comprehensive and systematic comparison of these methods.

To fill these gaps, we conduct controlled experiments to evaluate DR techniques in visual cluster analysis. Given the same high-dimensional datasets and certain cluster analysis tasks, our study compares the effectiveness of graphical outcomes generated by different DR techniques and investigates how their essential properties affect user’s visual perception and the task performance. The study starts with a comprehensive literature review of existing DR techniques, upon which twelve widely-used techniques are selected for evaluation (see Table 1). They are grouped based on two properties, namely locality and linearity, both of which are among the most concerned topics in the existing DR research. This grouping allows a high-level view of the DR techniques regarding their core features and thus the study results can be easily understood by data analysts. To ensure the study results are intuitive and generalized to various domains, we carefully define four analytical tasks that are essential for general visual cluster analysis. Each task is evaluated through a controlled experiment. The four corresponding controlled experiments are assessed based on the ground truth established in the data space.

Regarding the effectiveness of the selected DR techniques, a total of twelve hypotheses are formulated before the study. They are derived based on the summarized advantages of the DR techniques from our literature review. In the experiment design, three key design choices are made to ensure the soundness of the experiment and obtain compelling results. First, we select a total of eight datasets whose dimension count, data points count, and cluster patterns are balanced in the design. Then, a pre-study is conducted to tune and optimize the parameters for the selected DR techniques, ensuring their robustness for each of the datasets. In addition to the four controlled experiments, a subjective experiment is conducted where the participants rank the clustering outcomes based on their personal preference. It allows us to evaluate the DR from a more subjective perspective.

Based on the experiment results, we perform a comprehensive statistical analysis. The analysis results of the objective metrics validate seven of the hypotheses but leave five rejected. We found that although non-linear and local techniques show significant advantages in tasks such as cluster identification and cluster membership identification, there is not a technique that can outperform the others in all the four visual cluster tasks. Moreover, some linear techniques that receive less citation, such as NMF (Nonnegative Matrix Factorization) [29] and t-SNLE (t-Distributed Stochastic Neighbor Linear Embedding) [9], catch our eyes because they yield better performance in tasks such as cluster distance comparison and density comparison, respectively. We also discuss the impact of the two DR properties on the effectiveness of the visual perception and the potential improvement of the experiment design. The DR techniques, datasets, and records of the study are available at https://github.com/DR-approach/DR-approaches.
Wang et al. [51] used GONG [4] and SC (silhouette coefficient) [41] to compare the class separation performance of 10 DR techniques such as PCA and t-SNE on 93 datasets. Spahis et al. [45] used SC to evaluate the performance of interactive DR techniques such as LAMP, PLMP (Part-Linear Multidimensional Projection), and KELP (Kernel-based Linear Projection). Vernier et al. [50] evaluated performance of 11 DR techniques on 10 datasets using the neighborhood preservation.

**Evaluation based on user perception.** Lewis et al. [30] and Sedlmair et al. [44] reported that quantitative metrics might perform differently from human perception in cluster recognition and visual class separation tasks, respectively. A set of researches focus on perception based evaluations. Lewis et al. [31] designed a study to investigate human agreement on the embedding quality. Their results show that expert users are reasonably consistent judges of embedding quality, whereas novice users are very inconsistent with one another. Etemadpour et al. [17] conducted a controlled experiment to evaluate the performance of DR techniques on five visual cluster analysis tasks. Ventocilla et al. [49] conducted a similar perception evaluation but with a focus on PCA, t-SNE, Radviz, and SC. Xu et al. [57] conducted an experiment to compare advanced feature transformation technology (FT-high) and traditional DR techniques, such as PCA and MCML, regarding their performance in visual cluster analysis tasks. The results show that FT-high would yield better results than the DR techniques, especially in the accuracy of identifying clusters. Zhao et al. [61] conducted a controlled experiment to evaluate the performance of four DR techniques on fuzzy clusters analysis. Nevertheless, none of the above experimental studies has considered the two important properties, namely linearity and locality, in terms of their impact on the performance of the DR techniques. Another important difference lies in the task design and measure. Most existing studies are based on color-encoded scatterplots rather than monochrome scatterplots. In the task of identifying clusters [17], although monochrome scatterplots are used, their metric is based on the number of identified clusters only and lacks an accurate measure. Furthermore, most existing studies do not explore other cluster properties, such as density and membership.

3 Evaluation Landscape

In order to establish a solid basis for our study, we have conducted a comprehensive literature review, upon which the DR techniques, the datasets, and the analytical tasks of the study are derived.

### 3.1 Selection of DR techniques

To select the most representative techniques from the wealth of existing DR research, we consider the following three metrics: (1) influence - the techniques should be influential in general research fields, and more importantly, in the visualization field as well. In practice, we consider the citation as a major indicator of the influence; (2) task-suitability - the techniques should have been widely applied in visual cluster analysis, and considered essential for accomplishing the related analytical tasks. Specifically, we focus on those that are the most cited techniques for a certain visual cluster analysis task; (3) comparison-balance - since we intend to investigate the techniques based on their high-level categories, each category should have a consistent number of techniques for an effective comparison.

Based on the first two metrics, we conducted a comprehensive literature survey on recent research papers in the visualization community. Specifically, the source includes papers published in the journal of IEEE TVCG and three mainstream visualization conferences (IEEE VIS, EuroVis, and PacificVis) between 2010 and 2020. We distilled 937 papers that not only consist of the keywords “cluster” or “clustering” but also are relevant to DR techniques. Among them, we narrowed the keyword search related to the visual cluster analysis, which derived 51 papers as the finally refined source.

We conducted an in-depth survey on each of the papers and selected nine DR techniques based on our metrics. They were grouped based on the properties of linearity and locality. Table 1 summarizes the selected techniques with their categories, their google scholar citation counts, and the number of times it has been cited within the 51 papers. The classical techniques, such as PCA, have multiple related papers that have high citation counts. For clarity, we selected the paper with the highest citation counts for a technique. While there was no DR technique in the group of Linear&Local, we selected three more techniques, namely LPP, NPE, and t-SNLE, which are linear variants of LE, LLE, and t-SNE, respectively. 

| Techniques | Linearity | Locality | Citation | NO. of Applications |
|------------|-----------|----------|----------|---------------------|
| t-SNE [47] | non-linear| local     | 18655    | 30                  |
| Isomap [46] | non-linear| global   | 14394    | 2                   |
| UMAP [56]  | non-linear| local     | 2244     | 2                   |
| MDS [28]   | non-linear| global   | 7942     | 19                  |
| LE [4]     | non-linear| local     | 4803     | 2                   |
| LLE [42]   | non-linear| local     | 15811    | 2                   |
| PCA [19]   | linear    | global    | 8807     | 14                  |
| FA [18]    | linear    | global    | 14660    | 1                   |
| NMF [29]   | linear    | global    | 12365    | 1                   |
| LPP [23]   | linear    | local     | 4702     | 0                   |
| NPE [27]   | linear    | local     | 1829     | 0                   |
| t-SNLE [19] | linear    | local     | 87       | 0                   |

**Table 1:** DR techniques used in our evaluation.

### 3.2 Selection of Datasets

To reflect the common interests of visualization community, we select eight datasets out of 41 datasets that were used in the visualization applications in our survey. First, we considered the data size within a range from 100 to 1500 data instances and the cluster size from 3 to 10 to avoid scalability issue. Second, in order to make sure that the study can distinguish the performance of DR techniques, we select datasets for which cluster patterns can be visually separated in at least one projection result of a DR technique. Due to the huge decision space (41 datasets and 12 DR techniques), we assumed classes given by data labels formed clearly separable clusters. We used visual separation measures [5] GONG and KNNNG to quantify class separation and select the best projections for our study. Noting that we have not fixed the DR technique that can visually separate clusters, selecting datasets with visually distinguishable clusters introduces few bias into the study, if there is. As a result, the selected eight datasets are shown in Table 2.

We use PCA to reduce the dimensionality of ExtYaleB from 32,256 to 30 for a meaningful projection. All datasets are normalized by dimension.

**Table 2:** Datasets used in our evaluation.

| Datasets          | Data Items | Dimensions | Clusters |
|-------------------|------------|------------|----------|
| EcoliProteins [15]| 336        | 7          | 8        |
| Dermatology [15]  | 259        | 34         | 6        |
| ExtYaleB [20]     | 320        | 30         | 5        |
| World12 [16]      | 151        | 12         | 5        |
| Boston [15]       | 155        | 13         | 3        |
| MNist64 [15]      | 1083       | 64         | 6        |
| Weather [39]      | 366        | 194        | 7        |
| Olive [18]        | 572        | 8          | 3        |

### 3.3 Selection of Typical Tasks

The goal of the task formulation is to select the most common analytical tasks of the visual cluster analysis within general application domains. To this end, we extracted all the analytical tasks from the 51 surveyed cluster analysis papers and selected the top four most-cited tasks.

- **Cluster identification (TI):** given a scatterplot generated by a DR technique, identify dense and well-separated clusters. Input: projected data. Output: color-coded points from each identified cluster. Action: lasso selection.
3.1 Parameters

In this study, we focus on hyperparameters that significantly affect the projection results and require trial-and-error tuning. Specifically, we considered two parameters, namely neighborhood value and perplexity, as they significantly affect the projection quality of LPP, NPE, Isomap, UMAP, LE, LLE, t-SNE, and t-SNLE in different degrees [16,19,48]. For the other parameters, such as iteration times, we adopted recommended values based on previous studies [44,47]. Following the previous studies [16,48], we formulated 5-level ranges for the parameter neighborhood value and perplexity, from which the optimal values can be selected for each DR method. The parameter range formulation ensures a moderate degree of discrimination for the projection results while keeping a reasonable range to fit the optimal values. The settings are presented in Table 3.

### Table 3: The parameter settings of DR techniques in the pre-study.

| Techniques    | Key Parameters | Settings          | Other Parameters |
|---------------|----------------|-------------------|------------------|
| t-SNE         | Perplexity     | 5,15,30,40,50     | Early exaggeration=6 |
|               |                |                   | Iteration=3000   |
| t-SNLE        | Perplexity     | 5,15,30,40,50     | Early exaggeration=6 |
|               |                |                   | Iteration=3000   |
| UMAP          | Neighbors      | 4,7,10,13,16      | Minimum distance=0,1 |
|               |                |                   | Iteration=500    |
| LLE           | Neighbors      | 4,7,10,13,16      | Reg=0.001        |
|               |                |                   | Iteration=200    |
| LPP           | Neighbors      | 4,7,10,13,16      | None             |
| NPE           | Neighbors      | 4,7,10,13,16      | None             |
| Isomap        | Neighbors      | 4,7,10,13,16      | None             |
| LE            | Neighbors      | 4,7,10,13,16      | None             |
| MDS           | None           |                   | Iteration=500    |
| FA            | None           |                   | Iteration=2000   |
| NMF           | None           |                   | Iteration=400    |
| PCA           | None           |                   | No parameters    |

4.1 Pre-Study: Parameters of DR Techniques

The projection of the DR technique is usually significantly affected by their hyperparameters [19]. Improper parameters will lead to unreliable comparison results. In order to optimize the parameters of selected DR techniques for each dataset, we have conducted a preliminary experiment.

### 4.2 Task and Procedure

The same set of datasets was used in both formal study and pre-study. For a given dataset and a DR technique, participants were provided with 5 projections resulting from 5 parameter settings (see Table 5). The number and size of classes, which are taken as ground truth for clusters, are also explicitly mentioned to the participants of the pre-study.
H2.1–H2.3 Regarding T2, although it is a different perception task from T1, it also requires the ability to separate clusters and preserve near neighbors. Therefore, we have hypotheses H2.1, H2.2, and H2.3 in T2 that are identical to H1.1, H1.2, and H1.3 in T1, respectively.

H3.1–H3.3 Regarding T3, global DR techniques are designed to preserve global pairwise distances and thus could enable more faithful projections of global structures, e.g., cluster relationships, of high-dimensional spaces [26]. Compared to the non-linear embeddings, the linear embeddings are more reliable in terms of metric-preserving properties [27]. This might facilitate the representation of cluster relationships. Moreover, PCA is the most widely used global linear technique. Therefore, we have the following hypotheses:

- H3.1 Global DR techniques perform better than local ones in T3 when they have the same linearity;
- H3.2 Linear DR techniques perform better than non-linear ones in T3 when they have the same locality;
- H3.3 PCA performs better than other techniques in T3.

H4.1–H4.3 When considering T4, there are many practical experiences [37][32] pointing out that non-linear or local techniques neglect the density information. It is also considered that PCA performs better than non-linear techniques [11]. Therefore, we have hypotheses H4.1, H4.2, and H4.3 in T2 that are identical to H3.1, H3.2, and H3.3 in T3, respectively.

5.2 Experiments
Under the guidance of the four tasks (T1–T4) and the associated twelve hypotheses (H1.1–H4.3), we designed four controlled experiments: E1–E4 for T1–T4 to verify H1–H4, respectively. In addition to the four controlled experiments, a subjective evaluation E5 was designed where we studied participants’ preferences for different DR techniques based on their free exploration of datasets. All DR techniques and datasets for E1–E5 are the same. There were 12(techniques) × 8(datasets) = 96 trials for each participant in E1–E4 and 1(3 × 4 layout for 12 techniques) × 8(datasets) = 8 trials for each participant in E5.

E1: T1 was performed to test if projected clusters were separated cleanly and easily recognized by participants. In each trial, the participants were asked to use a lasso tool to interactively select well-separated clusters, if any, from the visualization generated by a DR technique. Compared with other evaluation methods such as counting the cluster number [1][7], interactive selections enable more accurate detection of participants’ visual perceptions, especially for handling aggregated clusters that might lead to bias in counting the cluster number [3]. Since color would interfere with the judgments of the clusters, we rendered all data points in grey color and full opacity.

E2: T2 was performed to test whether a meaningful belonging relationship between data points and clusters can be easily identified. In each trial, a pre-defined data point was highlighted in red color. Participants were asked to use a lasso tool to interactively select a well-separated cluster, if any, that contains the highlighted data point. When selecting the pre-defined data point, we prefer challenging ones that are close to the cluster boundary in the high-dimensional space. To this end, we extend the clustering metric of silhouette coefficient [41] to high-dimensional space, upon which the silhouette coefficient values of each data point are calculated. Then, the point with the lowest value was selected.

E3: T3 was performed to evaluate the ability of different DR techniques to preserve the relative distance among clusters in terms of visual perception. In each trial, a pre-defined cluster was highlighted in red color. Participants were asked to use a lasso tool to interactively select a well-separated cluster, if any, that is closest to the given cluster.

E4: T4 was performed to evaluate the ability of different DR techniques to preserve the density in the aspect of visual perception. In each trial, participants were asked to use a lasso tool to interactively select a well-separated cluster, if any, that has the highest density among all clusters.

E5: Subjective experiment was used to compare the overall performance of different DR techniques based on participants’ personal preferences. In each trial, we assigned a participant twelve visualizations of a dataset generated by twelve DR techniques, respectively. They were asked to rank the rank the visualization from 1 to 12 (1 is the best) by their personal preference in terms of cluster separation performance, identical ranking for multiple visualizations is allowed in case their qualities are hard to distinguish. The visualizations were numbered and placed in a 3 × 4 layout with a random order (see Fig. 3).

5.3 Participants, Apparatus, and Testing Data
Participants. We recruited 60 participants (40 males and 20 females) for the formal study. Their age was from 19 to 32 with an average of 24. 56 of them are graduate students and four of them are bachelor students with research experience. They come from majors including computer science (46), engineering (9), art (4), and mathematics (1). 19 participants have a data visualization background. 6 participants study clustering algorithms. 11 participants have performed cluster analysis in bioinformatics. Each participant was rewarded $20 per hour for completing the experiments. None of them reported color blindness or color weakness. Participants in the formal study are different from participants in the pre-study.

Apparatus. The formal study was conducted online on a pre-built web application in order to involve more participants. The application allows participants to complete all the controlled experiments and subjective studies (E1–E5), providing functionality such as visualizations, interactions, and time counter to facilitate the experiments. All the participants remotely took the experiments on standard laptops with a 1,920x1,080 screen resolution and chrome browser. The participants were asked to share their screens with the researcher during the experiments to enable
Testing Data. Before the formal study, we generated scatterplots based on the eight selected datasets. For each dataset, we created twelve scatterplots by each of the selected DR techniques, respectively. The parameter of each DR technique was determined based on the results of the pre-study. The same layout was used for a specific DR technique and a dataset for different experiments. Except for the given points in E2 and the given clusters in E3, which were coded in red, the data points in the scatter plots were all coded in gray. The points were rendered with a radius of 3 pixels without transparency. The size of the scatterplots was 600 × 550 pixels in E1–E4, and 300 × 300 pixels in E1.

5.4 Procedure

A training session was conducted before the formal experiments. For the training session, we utilized the Iris dataset [15] to generate scatterplots for each DR technique. In the training session, the instructor explained the related concepts and the experimental procedure. Then, the participant was asked to complete a total of thirteen training trials, including three trials for each controlled experiment (E1–E4) and one trial for the subjective experiment (E5). The DR technique of the trial was randomly selected for each experiment. After the training session, the participant was asked to complete E1–E5 in order. A 5-minutes break was allowed before each experiment. The completion order of the trials in each experiment was randomly assigned to each participant. The participants used the application to complete each trial, upon which the trial completion time and results were automatically stored by the application. There was no time limit for each experiment. All the participants finished the five experiments within 120 minutes.

Toward the end of E5, the participants were asked to complete a questionnaire about their background and subjective feedback of the experiments. The questionnaire includes three sections. The first section is about the participant background, such as the education level, visualization experience, and knowledge about the DR techniques. In the second section, the participant was asked to list the factors, including variance of density, cluster separation, variance of count, variance of size, and variance of shape. If any, they relied on identifying the clusters in E1. In the last section, the participant was asked to rate the importance of the four analytical tasks (T1–T4) in their data analysis experience using a five-point Likert scale. After the questionnaire, we conducted a subjective interview to understand how the scatterplots could affect participants’ trial completion and their preference and/or experience on any DR techniques and visual clustering techniques in daily work.

6 Statistical Analysis

6.1 Result Measurement

To measure results of each trial for the five experiments, two conditions were considered: the trial completion time (the shorter the better), precision and recall (1 is best) were calculated for E1–E4; and the average ranking score of DR techniques was calculated for E5. The ground truth for computing the precision and recall are based on the class label of the datasets.

E1: While participants would identify different cluster structures, e.g., number of clusters, from the ground truth, matching the identification and the ground truth should be performed before the computation of precision and recall. We denote the identified clusters by participants as \( \{ C_i \}_{i=1}^K \), where \( K \) is the number of identified clusters. The ground-truth clusters are denoted as \( \{ G_j \}_{j=1}^N \), where \( N \) is the number of labels in the dataset. A matching function \( f \) matches an identified cluster \( C_i \) to \( f(C_i) \), where \( f(C_i) \in \{ G_j \}_{j=1}^N \). Note that \( f(C_i) \) could be an empty set when there is no matching for \( C_i \). The match is based on two rules. First, if an identified cluster \( C_i \) contains data points from multiple ground-truth clusters, it is matched with the ground-truth cluster contributing the most data points to \( C_i \). Second, if multiple identified clusters are matched to the same ground-truth cluster following the first rule, only the identified cluster containing the most data points from the ground-truth cluster is accepted. In each trial, we compute the precision as

\[
\frac{\sum_{i=1}^K | C_i \cap G_j | }{ \sum_{i=1}^K | f(C_i) \cap G_j | } \]

and the recall is the ratio of \( | C_i \cap G_j | \) to \( | C_i | \). The recall is computed as

\[
\frac{\sum_{i=1}^K | C_i \cap G_j | }{ \sum_{j=1}^N | G_j | } \]

E2–E4: The computation of precision and recall in E2–E4 is simpler than E1, because there is only one identified cluster \( C_i \). We denote the ground-truth cluster as \( G_j \). In each trial, the precision is the ratio of \( | C_i \cap G_j | \) to \( | C_i | \), and the recall is the ratio of \( | C_i \cap G_j | \) to \( | G_j | \). E5: We calculated the distance among clusters in the high-dimensional space for the ground truth. As Etemadpour et al. [17], for each pair of clusters, we calculated the average pairwise Euclidean distances of all pairs of points belonging to the two clusters, respectively. Based on the distance among clusters, we randomly selected one cluster as the pre-defined cluster and its nearest cluster as the ground truth. In E4, the participants’ selection was tested against the ground truth that has the highest density based on the minimum spanning tree-based measure in the high-dimensional space [17].

E5: For the subjective experiment E5, we compare the average ranking score of DR techniques in all trials. In each trial, the first one gets 12 points, the second one gets 11, ..., and the last one gets 1 point.

In E1–E4, for evaluating the performance of a DR technique in an experiment, we calculate the average precision and recall values on all participants and on all corresponding trials of the experiment. It is worth noting that both precision and recall are important for accurate measure of the quality. On one hand, when participants mistakenly identify unrelated points to the target cluster, the precision decreases. On the other hand, if they correctly identify partial target cluster, the recall decreases. As shown in the confusion matrix, a DR technique with high precision is capable of well separating ground truth clusters. Otherwise, multiple ground truth clusters that mixed in the projection would result in low precision. A low recall indicates that the DR technique cannot gather points together in one ground truth cluster, leading to a peak splitting event [56]. If a DR technique projects multiple ground truth clusters as one cluster in the 2D layout, participants would select all the connection points as one cluster, resulting in low precision and high recall.

6.2 Statistical Analysis Approaches

We utilize the average value and confidence interval of the result measures as indicators to test our hypotheses. Based on a Shapiro-Wilk test, we found that the results of E1–E4 do not follow a normal distribution. Therefore, we employed a Friedman test to examine any significant differences among the various indicators for the DR techniques. For any resulted significant differences from the test, we further conducted detailed comparisons based on Tukey’s HSD test. Here, we consider the standard significance level \( \alpha = 0.05 \) as a statistical significance for the results. When we take the effect size as 0.25 and the power as 0.95, the power analysis show that we need at least 18 samples in each experiment [40]. In our experiment, we have 60 samples for each trial that is sufficient for the p-value test.

6.3 Result Analysis

6.3.1 Objective Experiment Results

H1.1–H1.3 The results of E1 are displayed in Fig. 4. Among all techniques, UMAP and t-SNE have the best precision (89.07% and 85.34%, respectively) and recall (88.98% and 84.99%, respectively). UMAP also has the shortest completion time (13.65s). The Friedman tests show statistical significance of precision (\( \chi^2(11)=524.90, p<0.05 \)) and recall (\( \chi^2(11)=524.89, p<0.05 \)) and average completion time (\( \chi^2(11)=261.82, p<0.05 \)). Fig. 5 shows the pairwise significance relationships for all the DR techniques.

- H1.1 Local DR techniques perform better than global ones in T1 if they have the same linearity type. H1.1 is partially confirmed.
In the non-linear group (NL&Lc and NL&G), local techniques t-SNE and UMAP perform significantly better than all global techniques in terms of precision and recall. However, the local technique LLE performs significantly worse than global techniques MDS and Isomap. In the linear group (L&Lc and L&G), the local techniques t-SNLE and LPP performs significantly better than all the global techniques in terms of precision and recall. In terms of completion time, local techniques UMAP and LE perform significantly better than all global techniques.

- **H1.2** Non-linear DR techniques perform better than linear ones in T1 if they have the same locality type. H1.2 is partially confirmed. In terms of precision and recall, in the local group (NL&Lc and L&Lc), non-linear techniques t-SNE and UMAP perform significantly better than all linear techniques. However, the non-linear technique LLE performs significantly poorer than linear techniques t-SNLE and LPP. In the global group (NL&G and L&G), non-linear techniques MDS and Isomap perform significantly better than all the linear techniques except for PCA.

- **H1.3** t-SNE and UMAP perform better than other techniques in T1. H1.3 is confirmed.

In E1, UMAP and t-SNE perform significantly better than all other techniques in precision and recall. They also perform well in completion time. In terms of completion time, UMAP performs significantly better than other techniques except LE.

**H2.1–H2.3** As shown in Fig. 4, UMAP has the highest precision (85.81%) and shortest completion time (3.87s) in E2. t-SNE is the second in precision (85.14%) and completion time (4.29s). The Friedman tests show significant differences among different DR techniques in terms of precision ($\chi^2(11)=534.79, p<0.05$) and recall ($\chi^2(11)=506.43, p<0.05$) and average completion time($\chi^2(11)=297.76, p<0.05$). The pairwise significance relationships between each pair of DR techniques can be seen in Fig. 5.

- **H2.1** Local DR techniques perform better than global ones in T2 if they have the same linearity type. H2.1 is partially confirmed. In the non-linear group (NL&Lc and NL&G), local techniques t-SNE and UMAP perform significantly better than all global techniques in terms of precision. However, the local technique LLE performs significantly poorer than the global technique LLE. In the linear group (L&Lc and L&G), in terms of precision and recall, local techniques t-SNLE and LPP perform significantly better than all global techniques except PCA. In terms of completion time, local techniques t-SNE and UMAP also perform significantly better than all global techniques.

- **H2.2** Non-linear DR techniques perform better than linear ones in T2 if they have the same locality. H2.2 is partially confirmed. In the local group (NL&Lc and L&Lc), non-linear techniques t-SNE and UMAP perform significantly better than all linear techniques in terms of precision and recall. However, the non-linear technique LLE performs significantly poorer than the linear technique t-SNLE in precision. In the global group (NL&G and L&G), the non-linear technique Isomap performs significantly better than all linear techniques. In terms of completion time, non-linear techniques t-SNE and UMAP perform significantly better than all techniques except LPP.

- **H2.3** t-SNE and UMAP perform better than other techniques in T2. H2.3 is partially confirmed.

In E2, UMAP and t-SNE perform significantly better than all linear techniques in precision. Though they have lower recall than LE and LLE, there is no significant difference between the four techniques. In terms of completion time, UMAP and t-SNE also perform significantly better than the other techniques except LLE, LE and LPP.

**H3.1–H3.3** The results of E3 are displayed in Fig. 6. UMAP has the highest precision (43.94%) but ranked 8th in recall (54.42%). NMF has the second-highest precision (43.69%) and highest recall (75.42%). The Friedman tests show statistical significance among different DR techniques in terms of precision ($\chi^2(11)=506.43, p<0.05$) and recall ($\chi^2(11)=534.79, p<0.05$). However, Tukey’s HSD test shows there are no pairwise significance relationships exist between each pair of DR techniques in terms of completion time. The pairwise significance relationships between each pair of DR techniques can be seen in Fig. 6.

- **H3.1** Global DR techniques perform better than local ones in T3 when they have the same linearity. H3.1 is rejected.

In the non-linear group (NL&Lc and NL&G), the local technique UMAP performs significantly better than all global techniques in terms of precision. In terms of recall, global techniques MDS and Isomap only significantly better than local techniques t-SNE. In the linear group (L&Lc and L&G), all global techniques perform significantly better than all local techniques except NPE in terms of precision and recall.

- **H3.2** Linear DR techniques perform better than non-linear ones in T3 when they have the same locality. H3.2 is rejected.

In the local group (NL&Lc and L&Lc), the non-linear technique UMAP performs significantly better than all linear techniques in terms of precision. The linear technique NPE performs significantly better than non-linear techniques t-SNE and LLE in precision. In the global group (NL&G and L&G), linear techniques PCA, FA, and NMF perform significantly better than the non-linear technique MDS in terms of precision and recall.
The results of $E_5$ are shown in Fig. 7. The rankings of the most preferred techniques are: UMAP (9.26), t-SNE (8.60), LE (7.07), LPP (6.74), LLE (6.38), t-SNLE (6.26), Isomap (6.15), NPE (6.00), FA (5.95), PCA (5.86), MDS (5.16), NMF (4.53). Local techniques obtain the top four preferred techniques with UMAP and t-SNE ranked the highest positions. It suggests a strong preference for the local techniques from the participants. Regarding linearity, LPP is ranked as the 4th preferred technique. It indicates that there is no strong tendency toward a particular linearity type by the participants. It is worth noting that there could be misunderstandings in interpreting the projection results. For example, if a single cluster in the data space was scattered into two clusters by the projection, the participants could view it as a correct visual cluster. Similar to Lewis et al. [31], we made a correlation analysis between the results of the subjective experiment and the results of objective experiments. The correlation coefficients between the rankings of $E_5$ and $E_1$ are 0.73 (precision) and 0.76 (recall). The correlation confirms the consistency and differences between the subjective experiment and objective experiments.

6.4 Summary

Our experiment results indicate that there is no universally better DR technique and high-level property. Therefore, it is essential to choose appropriate techniques for enhancing particular tasks involving visual cluster analysis. (1) In cluster identification, local DR techniques perform better than global DR techniques in cluster identification if they have the same linearity type, except LLE; non-linear DR techniques perform better than linear DR techniques if they have the same locality type, except LLE; t-SNE and UMAP have an outstanding performance. (2) In membership identification, local DR techniques perform better than global DR techniques if they have the same linearity type, except...
Cluster identification and class separation. In this study, our evaluation is based on the perception of cluster identification. Participants were provided with monochrome scatterplots without color-coded class information. It is worth noting that we have also used the class label information with the datasets as the ground truth of clusters. Therefore, in selecting the datasets, we check the visual separation of clusters with the visual class separation measure. However, when a class is scattered into several clusters, the class separation measure also presents a good but misleading result. This issue can be addressed by additional user controlled selection with colored scatterplot. More discussion on this issue is referred to Aupetit [3].

Parameter robustness of DR techniques. Our pre-study provides an opportunity to analyze the DR techniques in the aspect of parameter robustness. In our pre-study, participants were asked to select the result that best fits the structure of datasets among five settings. If the selections are uniformly distributed in all five settings, we can infer that the tuning of parameters has few effects on the quality of projection. Otherwise, a big variation of the setting distribution indicates that the DR technique is sensitive to the parameter. We calculated the average standard deviation across datasets of each tested eight DR techniques in the pre-study. The rankings are: t-SNLE (3.25), t-SNE (3.15), LLE (3.11), UMAP (3.04), LE (2.57), NPE (2.39), Isomap (2.35), and LPP (1.85). The results show that although t-SNE is one of the most popular techniques, its sensitivity to the perplexity parameter should be considered. This observation is the same as Espadoto et al. [16]. It is also worth noting that UMAP is less sensitive than t-SNE while having similar performance in most tasks.

Conducting remote studies. Remote study lacks a face-to-face communication environment and yields several challenges for a control study. First, the test should be performed independently. In our study, participants are required to set an independent environment to avoid possible disturbance. Second, the test could be interrupted due to network issues. We ask participants to test their network to ensure a stable connection. During the test, if a participant is disconnected for more than ten minutes, we discard this sample to ensure the validity of tests. Third, remote communication would be hindered in laboratory study. To address this issue, we monitor the experimental screen of participants through screen sharing throughout the test. We also communicate with participant during the test in voice.

Order of experiments and trials. A limitation of our study is the order of experiments and trials. Because E3 is conducted before E4, it may reveal a part of ground truth clusters and make a potential threat to validity of E4. If the given cluster in E3 happens to be the cluster with the highest density in E4: 1) when several classes overlap in a single cluster, then an increase of precision would occur; 2) when a single class is scattered in several clusters, then an increase of recall would occur. In addition, learning effects should be considered. A Latin square design could address this issue.

Limitations and future work. Due to the different characteristics of datasets, DR techniques will inevitably produce different performance in different types of datasets. Therefore, the experimental results may have a certain correlation with the datasets. In this study, we focus on the characteristics of locality and linearity and four objective visual clustering tasks. In the future, it is valuable to perform further studies on other characteristics (for example, false and missed neighbors could have an impact on clustering tasks that is worth evaluating [38]) and tasks. The linear-nonlinear pairs of techniques, such as t-SNE and t-SNLE, also shed light on the development of DR techniques and design of comparative analysis.

8 Conclusion

In this paper, we present an empirical evaluation of DR techniques from the perspective of perception in visual cluster analysis tasks. Twelve representative DR techniques are identified from a literature review. They are grouped into two groups based on linearity and locality. We first conduct a pre-study to determine the proper input parameters for each DR technique. Then, the techniques are formally evaluated in four controlled experiments each of which focuses on hypotheses on a particular analytical task. A subjective task is also conducted to collect users’ preferences on each technique. Several guidelines and interesting insights are provided by the result analysis.

Acknowledgments

We would like to thank the helpful comments from the anonymous reviewers. This work is supported by the National Natural Science Foundation of China (No. 61872389).
[51] 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, 2017.

[52] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

[53] J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, 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, 2017.

[54] L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17(2):473–491, 2008.

[55] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[56] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu. Interactive correction of mislabeled training data. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, pp. 57–68, 2019.

[57] K. Xu, L. Zhang, D. Pérez, P. H. Nguyen, and A. Ogilvie-Smith. Evaluating interactive visualization of multidimensional data projection with feature transformation. *Multimodal Technologies and Interaction*, 1(3):13, 2017.

[58] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):3–36, 2021.

[59] J. Yuan, S. Xiang, J. Xia, L. Yu, and S. Liu. Evaluation of sampling methods for scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1720–1730, 2020.

[60] X. Yue, J. Bai, Q. Liu, Y. Tang, A. Puri, K. Li, and H. Qu. sPortfolio: Stratified visual analysis of stock portfolios. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):601–610, 2019.

[61] Y. Zhao, F. Luo, M. Chen, Y. Wang, J. Xia, F. Zhou, Y. Wang, Y. Chen, and W. Chen. Evaluating multi-dimensional visualizations for understanding fuzzy clusters. *IEEE transactions on visualization and computer graphics*, 25(1):12–21, 2018.