J Vis (2025) 28:205–221
https://doi.org/10.1007/s12650-024-01025-6

REGULAR PAPER

Kaya Okada · **Takayuki Itoh**

Scatterplot selection for dimensionality reduction in multidimensional data visualization

Received: 2 March 2024 / Revised: 26 May 2024 / Accepted: 30 July 2024 / Published online: 23 August 2024

Abstract Dimensionality reduction (DR) techniques for multidimensional data serve as powerful tools for visualization and understanding of the structure of the data. Various DR methods have been developed to extract specific features of the data over the years. However, selection of the optimal DR method and fine-tuning parameters are still challenging, as these choices vary based on the characteristics of the dataset. Consequently, data scientists often rely on their experience or undertake extensive experimentation to identify the most suitable approach. This paper proposes a semi-automatic method for selecting appropriate DR techniques through scatterplot evaluation. Initially, our approach applies a range of DR methods to the given multidimensional data to compute two-dimensional values. Next, we generate scatterplots from the two-dimensional data and calculate scores reflecting the distribution and spatial relationships among the points. Scatterplots that provide insights achieve higher scores, enabling an efficient selection of DR methods based on their visualization. We demonstrate the effectiveness of the presented method through two case studies: The first one is an e-commerce review dataset, and the second focuses on a dataset derived from music feature extraction.

Keywords Dimensionality reduction · Multidimensional data visualization · Scatterplot · Scatterplot selection · Evaluation of dimensionality reduction · Text data visualization

1 Introduction

The collection and utilization of multidimensional datasets have increasingly become commonplace, thanks to advancements in machine learning and the evolution of database systems utilization. These datasets contain not only transaction and machine processing data but also a wide variety of unstructured data types, including texts, music, images, and videos, which are vectorized into multidimensional forms. However, difficulties remain in the ability of humans to navigate and extract valuable insights from these large and complex datasets as stored in extensive tables. To address this problem, many studies have employed dimensionality reduction (DR) techniques to extract only the most significant features and visualize them. Such visualization facilitates an intuitive understanding of the data, ranging from an overview to detailed characteristics, and enables the evaluation of data processing techniques and models that involve vectorization.

K. Okada (✉) · T. Itoh
Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
E-mail: kaya@itolab.is.ocha.ac.jp

T. Itoh
E-mail: itot@is.ocha.ac.jp

There have been various types of DR techniques over recent decades. These methodologies fall into two primary categories: linear techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), and nonlinear techniques, including t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). As shown in Sect. 2.1, numerous studies have compared and evaluated these DR techniques in terms of their characteristics, parameters, and accuracy. Nevertheless, the appropriate selection and application of DR techniques remain challenging for several reasons. First, the optimal DR method varies for each dataset, dependent on its features and objectives. Consequently, data science experts often rely on their experience for selection, leading to a subjective determination of the "best" method. This subjectivity is compounded by the lack of a clear definition for "good" DR results; therefore, assessments of DR methods are ambiguous. Second, the computational cost associated with DR is significantly high. Here, DR is basically applied to vast multidimensional datasets, where manual implementation, computation, and assessment can be time-consuming. This challenge is accelerated especially in the case of unstructured data, such as text, where vectors are often sparse and dimensionality can be extremely high; thus, computation time increases. These issues have emerged in the processing of unstructured data, highlighting the need for more efficiency. Bioinformatics is one of the fields that extensively utilizes DR techniques. Numerous studies have demonstrated the prevalence of employing DR to analyze and comprehend the characteristics of genetic data, particularly RNA sequences, due to the large size of the data columns. However, the optimal DR method for a specific dataset remains an active area of research. Heiser and Lau (2020) applied single-cell RNA sequence (scRNA-seq) data to visualize clusters and identify dispersion trends in local and global distance distributions. Huang et al. (2022) evaluated the performance of various DR techniques on single-cell transcriptomic data. Remeseiro and Bolon-Canedo (2019) employed DR methods for efficient feature selection in medical imaging, biomedical signal processing, and DNA microarray data. DR techniques have also been applied to other domains, such as face recognition and text analysis (Ayesha et al. 2020). Ayesha et al. (2020) discussed the mapping of appropriate DR methods based on the type of data being analyzed. While DR methods have been widely adopted across various data types, the selection of the most suitable DR technique for a particular case remains a topic of ongoing debate.

In this study, we present a method designed for the semi-automatic selection of appropriate DR techniques for multidimensional data. Initially, we systematically apply a variety of DR methods, each with a range of potential parameters, to a multidimensional dataset. After calculating two-dimensional vectors by the above process, we generate scatterplots in two-dimensional spaces for each of DR methods. Then, we compute scores that reflect the distribution and spatial relationships of the points for every scatterplot corresponding to each of the DR methods. The selection process uses these scores to pick out only scatterplots worth a closer look. We calculate multiple scores for each scatterplot since what makes a scatterplot interesting can depend on many factors, like its shape and features. These scores are then integrated into a single score through weighting which is tailored to the specific objectives of the visualization. Finally, scatterplots that offer significant insights receive higher scores and they enable automatic scatterplot selection. This methodology not only serves as a proxy for selecting DR methods but also significantly reduces the time users spend in the cycle from implementation to assessment.

In this paper, we introduce two application cases to illustrate the efficacy of our proposed method. The first dataset comprises Amazon e-commerce review data, from which we have generated embeddings with 767 dimensions, derived from the review texts. Each reviews are associated with product categories. The second dataset is Japanese pop music features data. It consists of over 1,000 dimensions computed using machine learning techniques, and each rows are associated with the release year of the tracks. Figure 1 shows an example of scatterplot selection by this technique. We implemented four metrics, scoring with 3 types of weighting and a dashboard that shows the automatic scatterplot selection result based on these scores. It enables us to easily find the most preferable DR methods and understand the characteristics of data more easily.

2 Related work

This section first introduces comprehensive review papers on dimensionality reduction methods and experimental research applied to various types of data, with a particular focus on text embeddings. In the latter parts, we discuss recent advancements in multidimensional data visualization techniques, specifically the techniques of scatterplot evaluation.

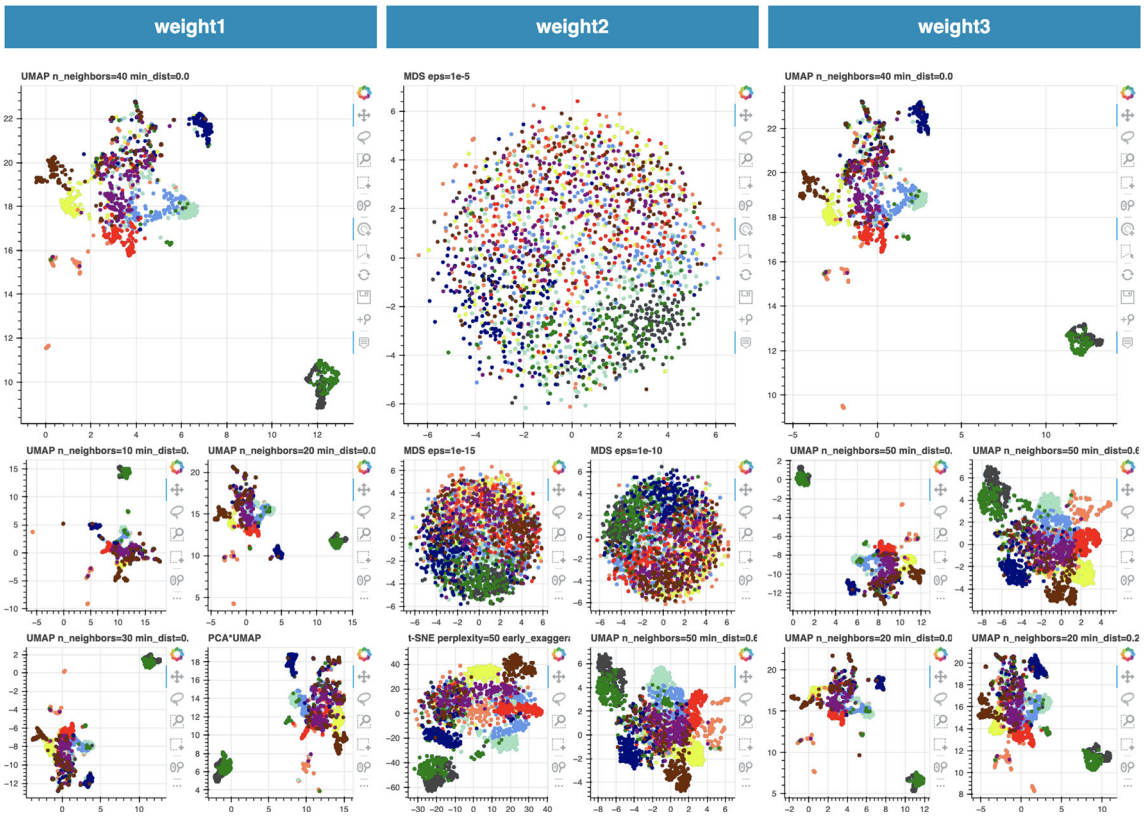

Fig. 1 Visualization dashboard application example. Initially, our proposed technique applies various DR techniques to high-dimensional data, reducing the dimensions to two. The resulting scatterplots are evaluated using four metrics with three weights each. The application automatically selects and visualizes the highest-scoring scatterplots, enabling efficient exploration and understanding of high-dimensional data

2.1 Dimensionality reduction methods for multidimensional data visualization

For decades, various types of dimensionality reduction (DR) techniques for analyzing multidimensional data have been developed and tailored to meet specific analytical requirements. Extensive survey and review studies (Fodor 2002; Van Der Maaten et al. 2009; Anowar et al. 2021; Ayesha et al. 2020; Saini and Sharma 2018; Malik et al. 2023) have been conducted to evaluate these methods and select the optimal approach for each data type. These studies provide comprehensive overviews of principal DR methodologies, detailing their characteristics and assessing them based on criteria such as classification accuracy, correlation, parameters, and computational efficiency. Particularly, Nanga et al. (2021) highlighted their applicability across diverse data types including text and images.

Similarly, Engel et al. (2012) reviewed various DR methods from a visualization perspective. The review concluded that no single method can be preferred over another, and the effectiveness of state-of-the-art methods mainly depends on the data and application. Consequently, many studies focused on applying specific DR methods to particular fields of datasets and evaluating them through comparative analysis. For instance, Wang et al. (2023) investigated the accuracy of DR results using time-of-flight data, while Heiser and Lau (2020) visualized the clusters and identified dispersion trends in local and global distance distributions by applying single-cell RNA sequencing (scRNA-seq) data. Nadia Syed and Jamil (2023) assessed the accuracy of PCA, LDA, and SVD while applying to cancer datasets, further suggesting that incorporating feature extraction and selection methods into DR processes can significantly enhance classification performance.

We apply the DR methods mainly to the text data in this study. Huang et al. (2005) proposed a mechanism for comparing and evaluating the effectiveness of DR in the visual exploration of text documents. Their approach assessed which DR best preserves the interrelationships within a set of text documents, as interpreted through numerous visualizations. Prior to implementing DR, text data are transformed

into an n x m document-term matrix (the number of documents and terms are n and m, respectively). Similar to this study, there have been various methods for processing text data as is. On the other hand, our proposed methods involve the creation of embeddings prior to DR to enhance the versatility of these methods. Singh et al. (2022) introduced a novel DR technique that employs the "GloVe" word embedding method to eliminate redundant features by assessing the similarity score between word vectors and comparing the results with those of existing DR methods. Vashisth and Meehan (2020) compared multiple Natural Language Processing (NLP) techniques such as Bag of Words, Word Embedding (W2Vec, GloVe), and traditional Machine Learning techniques (Logistic Regression, Support Vector Machine, and Naïve Bayes) for gender classification in tweets. Yamada et al. (2018) also proposed a new embedding model named Wikipedia2Vec. Both studies utilized DR methods to demonstrate the effectiveness of these models and systems. In our experimental section, we apply doc2vec and SimCSE (Gao et al. 2021) as preprocessing steps. The SimCSE results revealed more distinct clusters and characteristics across multiple datasets and DR methods; thus, this paper introduces only the results applying SinCSE.

Based on the related work mentioned above, we employed eight DR techniques in our case studies as shown in Table 1. We chose these methods because they are well-known and represent different combinations of features. Linear methods exhibit a propensity for effectively analyzing global structure, while nonlinear methods excel at capturing local structure, though they require more computational power. Anowar et al. (2021), Van Der Maaten et al. (2009), and Saini and Sharma (2018) have clearly and effectively summarized the features of each DR technique, so we refer to their work in our analysis.

2.2 Numerical evaluation of scatterplots

A certain number of metrics for evaluating two-dimensional scatterplots have been devised. Wilkinson et al. (2005) introduced Scagnostics, providing nine comprehensive metrics based on the distribution of points in a scatterplot to identify shapes, outliers, and correlations among others. Wang et al. (2019) have presented an improved Scagnostics method based on human visual perception for determining outliers and clusters in scatterplots. Based on this approach, Itoh et al. (2023) calculated scores based on several independent metrics for each scatterplot. Moreover, they implement a graph coloring algorithm to extract a set of various scatterplots while avoiding selecting sets of close vectorized scores that have similar features. In our proposed technique, we employed metrics grounded in the scoring approach in the above study.

There have been studies on improved calculation for the metrics presented by Wilkinson. Aupetit and Sedlmair (2016) and Sedlmair et al. (2012) have discussed the separability of classes within scatterplots. Harrison et al. (2014) have addressed the correlation among clusters of points. Sips et al. (2009) and Lee et al. (2013) have focused on the consistency and integrity of classes when mapping high-dimensional data to lower-dimensional scatterplots. Notably, Sedlmair et al. (2012) have defined the characteristics of data and distributions regarding class separability in scatterplots and applied these definitions to data after dimensionality reduction. Dang and Wilkinson (2014) and Matute et al. (2017) have proposed methods for organizing, summarizing, and exploring large sets of scatterplots.

3 Evaluation of dimension reduction methods using scatterplot selection

This section presents a processing flow of the dimensionality reduction (DR) and scatterplot selection technique. The technique calculates the scores of each scatterplot based on multiple metrics. Furthermore, the presented technique displays the high score scatterplots on the top of the screen space.

3.1 Dimensionality reduction methods

This technique applies eight different methods of dimensionality reduction, which helps us to visualize complex multi-dimensional data more easily. This section does not describe the details of these methods here because they have already been well-known and widely used. The DR process in this paper is specifically aimed at visualization purposes, thus converting multi-dimensional datasets into two-dimensional ones. The DR methods applied in this study include the following:

1. PCA (Principal Component Analysis)
2. t-SNE (T-distributed Stochastic Neighbor Embedding)
3. LDA (Linear Discriminant Analysis)

Table 1 The characteristics of DR techniques we selected

DR Method	Goal	Supervision	Linearity	Topology	Computational Complexity
PCA	Maximize variance	Unsuper	Linear	Random Projection	$O(d^2n + n^3)$
LDA	Maximize class separation	Super	Linear	Random Projection	$O(d^2n), n > dandO(d^3), d > n$
MDS	Preserve Euclidean pairwise distances	Unsuper	Nonlinear	Manifold	$O(n^3)$
t-SNE	Preserve local structure	Unsuper	Nonlinear	Manifold	$O(n^2)$
UMAP	Preserve local and global structure	Unsuper	Nonlinear	Manifold	$O(nlogn)$
ICA	Maximize statistical independence	Super	Linear	Random Projection	$O[2di(d+1)n]$
SVD	Minimize reconstruction error	Unsuper	Linear	Random Projection	$O(d^2n + n^3)$
KPCA	Linearly separate data	Unsuper	Nonlinear	Manifold	$O(n^3)$

4. MDS (Multi-Dimensional Scaling)
5. UMAP (Uniform Manifold Approximation and Projection)
6. SVD (Singular Value Decomposition)
7. ICA (Independent Component Analysis)
8. KernelPCA (Kernel Principle Component Analysis)

The application of a two-step dimensionality reduction process, which is the combination of different DR methods, has been shown to yield superior results compared to the use of a single DR method. Agis and Pozo (2019) employed a combination of PCA and t-SNE, while Padron-Manrique et al. (2022) and Stolarek et al. (2022) utilized a combination of PCA and UMAP. In all cases, the authors reported improved performance and quality compared to the results obtained by applying t-SNE or UMAP alone. Agis et al. discovered that the combination of PCA and t-SNE enhances the quality of clusters related to structural states. Stolarek et al. found that the performance of UMAP can be further improved by preprocessing the image input using PCA. Padron et al. introduced a novel method called sc-PHENIX, which utilizes the initialization of PCA-UMAP space and provides a closer approximation to the true underlying manifold of scRNA-seq data compared to UMAP, PHATE, and MAGIC.

Our study also explores the two-step dimensionality reduction consists of the following steps:

1. Initially reduce the dimensions to a specified number based on the Cumulative Contribution Rate calculated by PCA.
2. Subsequently reduce to two-dimensions using the selected methods.

3.2 Data structure

This paper formalizes the data structure after applying the DR methods. m-dimensional dataset A has n individuals as $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n\}$. The i-th individual $\boldsymbol{a}_i$ has the m-dimensional values as $\boldsymbol{a}_i = (a_{i1}, a_{i2}, ..., a_{im})$. Additionally, each sample has been assigned to one or more classes. A set of scatterplots formed from every pair of dimensions is described as $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N\}$, where N is the total number of scatterplots. Each scatterplot has a set of scores calculated based on predefined metrics. This section describes the score of the j-th scatterplot as $\boldsymbol{s}_j = (s_{j1}, s_{j2}, ..., s_{jM})$, where M is the number of metrics.

3.3 Selection of metrics

Based on the objective of visualization and DR, we defined "good results of DR" as follows:

- Represent the features of the original data
- Preserve the structure of the original data

We implemented the following four metrics to assist in finding the methods that provide the above results.

1. Separability of clusters
2. Separability of classes
3. Continuity of classes

4. Conservation of distance between a pair of points

3.3.1 Separability of clusters

We can expect to discover through the observation of scatterplots if the points are clearly separated into a finite number of clusters after applying DR. The score can be high with the results of applying techniques that maximize separation between classes, such as LDA. The following metric evaluates this property.

Firstly, the technique applies hierarchical clustering to the set of points, forming into a certain number (e.g., 5 to 20) of clusters. The clustering result is denoted as c_i of the scatterplot s_i. For each number of clusters $u = 5, 6, \cdots, 20$, our technique evaluates the clustering results using the Calinski-Harabasz index (Caliński and Harabasz 1974) and assigns the maximum score calculated among the clusters as the score for scatterplot s_i. A high score indicates both high cohesion within each cluster and high separation between multiple clusters.

$$s_{k1} = max(CalinskiHarabasz(c_k)_u)$$

3.3.2 Separability of class

Scatterplots in which points belonging to specific classes are clearly separated from other points of other classes often provide valuable insights. The following metric evaluates this property.

The proposed technique quantifies the separability of classes based on information entropy. Specifically, for a scatterplot constructed from the two dimensions, the following value is calculated:

$$H(i,j) = -\sum_{k=1}^{n}\sum_{c=1}^{C} p(y_k = c|(a_{k1}, a_{k2}))\log p(y_k = c|(a_{k1}, a_{k2}))$$

y_k represents the class of the k-th sample, (a_{k1}, a_{k2}) denotes the coordinate values of the k-th sample, and C is the number of classes. In our implementation, the scatterplot is divided into L subregions and the entropy $H(i,j)_l$ of the l-th subregion is calculated using the aforementioned equation. Finally, the score of the k-th scatterplot is determined using the following equation:

$$s_{k2} = \left(H_{max} - \frac{1}{L}\sum H(i,j)_l\right)/H_{max}$$

Here, H_{max} is the maximum number of $H(i,j)_l$.

3.3.3 Continuity of class

When the classes associated with a set of points have an order and the classes are arranged in sequence (such that a gradient is observed when coloring is applied with consideration for order), we can explain that the data preserve the characteristics of the original points classes. The following metric evaluates this property.

Firstly we calculate the average position of the points included in each class. Secondly, we generate a Delaunay triangular mesh connecting the average positions of each class and calculate the score as the sum of the absolute differences in the sequential IDs of classes between the endpoints of each edge. We describe the score as the following equation that calculates the difference between the sequential IDs of classes of the v-th and $(v + 1)$-th vertices connected by an edge, where c_i denotes the sequential ID of a class assigned to each average position.

$$s_{k3} = \sum_{v=1}^{m} |(c_{v+1} - c_v)|$$

A lower score indicates higher continuity; therefore, we apply $s_{k3} = 1 - s_{k3}$ after normalization described in Sect. 3.4.

3.3.4 Conservation of distance between points

Effective DR methods maintain the original distances between points well. Specifically, certain DR techniques aim to preserve the global structure and the relationships among neighboring points. The following metric evaluates this property.

To verify the preservation of distances, we applied Spearman's rank correlation coefficient. Initially, for both the original multi-dimensional values and the resulting two-dimensional values, the method computes the distances across all possible pairs of points. Subsequently, the method establishes a ranking of these point pairs based on their distances. The degree of similarity between the ranking of original multi-dimensional values and that of the two-dimensional values is computed using Spearman's rank correlation coefficient, indicating that scatterplots closely reflecting the original dataset structure are deemed to have a higher score. The first score of the k-th scatterplot is calculated using the following equation:

$$s_{k4} = |Spear(rank(dist(i,j)_{multi-dimensional}), \quad rank(dist(i,j)_{two-dimensional}))^2|$$

Here, *Spear* is Spearman's rank correlation coefficient, *rank* is the ranking of the pairs, and $dist(i, j)$ is the distance of i-th and j-th point.

3.4 Score weighting

The technique generates the four-dimensional vector with the scores calculated by the metrics for each scatterplot and normalizes them with their minimum/maximum values. Here, we define the weight for each metric $w = \{w_1, w_2, w_3, w_4\}(w_1 + w_2 + w_3 + w_4 = 1.0)$. Finally, we calculate the score for each scatterplot as the sum of the weighted scores.

$$s'_j = \sum_{k=1}^{4} w_k(s_{jk} - min(s_k))/(max(s_k) - min(s_k))$$

3.5 Scatterplot selection and dashboard generation

After calculating the scores for each dimensionality reduction (DR) method, our technique ranks the scatterplots according to their respective scores within each predefined weight. Given the extensive number of scatterplots generated from various DR methods and parameters, it becomes impractical to display all on a single screen. Consequently, we prioritize the selection of scatterplots that achieve the highest scores for presentation. Our technique creates the dashboard which is separated into the number of weights. Scatterplots are organized according to the highest score within each weight category, with those achieving higher scores positioned at the top. Therefore, users can quickly identify the most effective scatterplots and DR methods for a given dataset at a glance.

4 Case study and result

This section introduces the example results applying the technique presented in Sect. 3. The initial dataset analyzed in this study is the Amazon Review Data, for which we provide a detailed discussion of the scoring and visualization outcomes. To demonstrate the versatility of our proposed technique beyond text data and include other forms of multidimensional data, we briefly present the results applied to music feature data as a second example.

4.1 Commonly applied methods to the two cases

In our examples, we apply the same DR methods, parameters, and scoring weights. The explanation of these common conditions precedes the presentation of data and results.

4.1.1 Employed DR methods

In our approach, we applied a range of DR techniques with various parameters, as shown in Table 2. We adjusted parameters for PCA, MDS, LDA, t-SNE, and UMAP, and notably for t-SNE and UMAP, we

Table 2 The DR and parameters applied in our cases

DR Method	Parameter	Range
PCA	svd_solver	"auto," "full," "arpack," "randomized" (default = "auto")
MDS	eps	1e-15 ∼1.0 (default = 1e-3)
LDA	solver	"svd," "eigen" (default = "svd")
t-SNE	perplexity	10 ∼50 (default = 30)
t-SNE	early_exaggeration	10 ∼50 (default = 12)
UMAP	n_neighbors	10 ∼50 (default = 15)
UMAP	min_dist	0.0 ∼0.99 (default = 0.1)
ICA	As default	None
SVD	As default	None
KPCA	As default	None
PCA and t-SNE	As default	None
PCA and UMAP	As default	None
MDS and t-SNE	As default	None
MDS and UMAP	As default	None

experimented with different combinations of parameters in a round-robin manner. This resulted in the creation of 72 scatterplots, showcasing the diverse outcomes of these DR techniques.

Furthermore, integrating multiple DR techniques often enhances the results. During our preliminary experiment, we employed a round-robin to test various combinations of DR methods and identified that linear method and nonlinear method combinations tend to provide superior results. The first step in the first DR is to decide how many dimensions to reduce to. We calculated the Cumulative Contribution Rate using PCA, which revealed that reducing the data to 83 dimensions accounts for 80% of the explained variance, 35 dimensions for 60%, and 22 dimensions for 50%. Our pre-experiment indicated that reducing the data to 22 dimensions using the first DR method, followed by the application of a second DR method, produced the most effective visualization. Consequently, this paper will focus exclusively on those findings.

4.1.2 Weights of scores and dashboard arrangement

In our study, we define three types of weights for evaluating scatterplots as summarized in Table 3. Weight 1 aims to highlight scatterplots with clear clusters with distinguishable classes, indicating that the reduced dimensions preserve the original characteristics of the data. Several DRs tend to distort the actual distance to emphasize features or clusters instead of preserving the global relation of points. Weight 2 selects scatterplots that maintain the original distances between data points, using metrics for continuity and conservation of distance to ensure spatial relationships are preserved. Weight 3 offers a balanced approach, averaging scores across various metrics to select well-rounded scatterplots. These weights are adaptable and should be modified based on the specific goals of each visualization task. In this study, we layouted the each weighted scatterplot on the dashboard horizontally. The correlation between the dashboard's arrangement and the score ranking is detailed in Fig. 2.

4.2 Case study 1: Amazon review data

4.2.1 Dataset

In the initial case study presented in this paper, we applied the Amazon Review Data (2018) (Ni 2018), specifically selecting from "Small subsets for experimentation." Here, we limited our analysis to a total of 2000 records, owing to the substantial size of the dataset, which renders direct visualization on scatterplots impractical. The applied records consist of each 200 records from the following 10 categories:

1. Gift Cards
2. All Beauty
3. Grocery and Gourmet Food
4. Amazon Fashion
5. Arts, Crafts, and Sewing
6. Industrial and Scientific
7. Automotive

Table 3 The weight applied to our cases

Name	Separability of Clusters	Separability of Class	Continuity of Class	Conservation of Points Distance
Weight 1	0.7	0.3	0.0	0.0
Weight 2	0.0	0.0	0.5	0.5
Weight 3	0.25	0.25	0.25	0.25

Fig. 2 Dashboard layout to show only selected scatterplots

8. Cell Phones and Accessories
9. Digital Music
10. CDs and Vinyl

We extracted the review text data from the original JSON format and then subsequently transformed them into 767-dimensional vectors through SimCSE (Gao 2021) and saved in CSV format. Here, each record is associated with a product category name. The specific model utilized for this transformation was "princeton-nlp/sup-simcse-bert-base-uncased." Before settling on SimCSE, we also explored the use of the doc2vec model. However, the comparative analysis showed that SimCSE provided superior visualization outcomes. Thus, this paper focuses only on the results using SimCSE.

4.2.2 Evaluation score and visualization result

According to scoring by the first weight (Sect. 4.1.2), the high-scoring scatterplots were predominantly generated using UMAP with small min_dist, as detailed in Table 4. Conversely, scatterplots derived from MDS and several linear DR techniques were found to achieve lower scores, as evidenced by the data presented. Figure 3 clearly illustrates the differences between the high- and low-score scatterplots. The high-score scatterplots exhibit clusters, a hallmark of effective DR methods. This is particularly characteristic of UMAP, which is designed to preserve the relative proximity of data points from high-dimensional space in the reduced low-dimensional space, ensuring that points close in the original space remain close,

Table 4 DR methods which have the highest and lowest scores with Weight 1

Rank	DR name	Parameters	Score
1	UMAP	n_neighbors = 40 min_dist = 0.0	0.79969
2	UMAP	n_neighbors = 10 min_dist = 0.0	0.59000
3	UMAP	n_neighbors = 20 min_dist = 0.0	0.58845
4	UMAP	n_neighbors = 30 min_dist = 0.0	0.58715
5	PCA*UMAP	default	0.47821
68	MDS	eps = 1e-15	0.05036
69	MDS	eps = 1e-10	0.04915
70	MDS	eps = 1e-5	0.04656
71	SVD	default	0.03668
72	MDS	eps = 1.0	0.00394

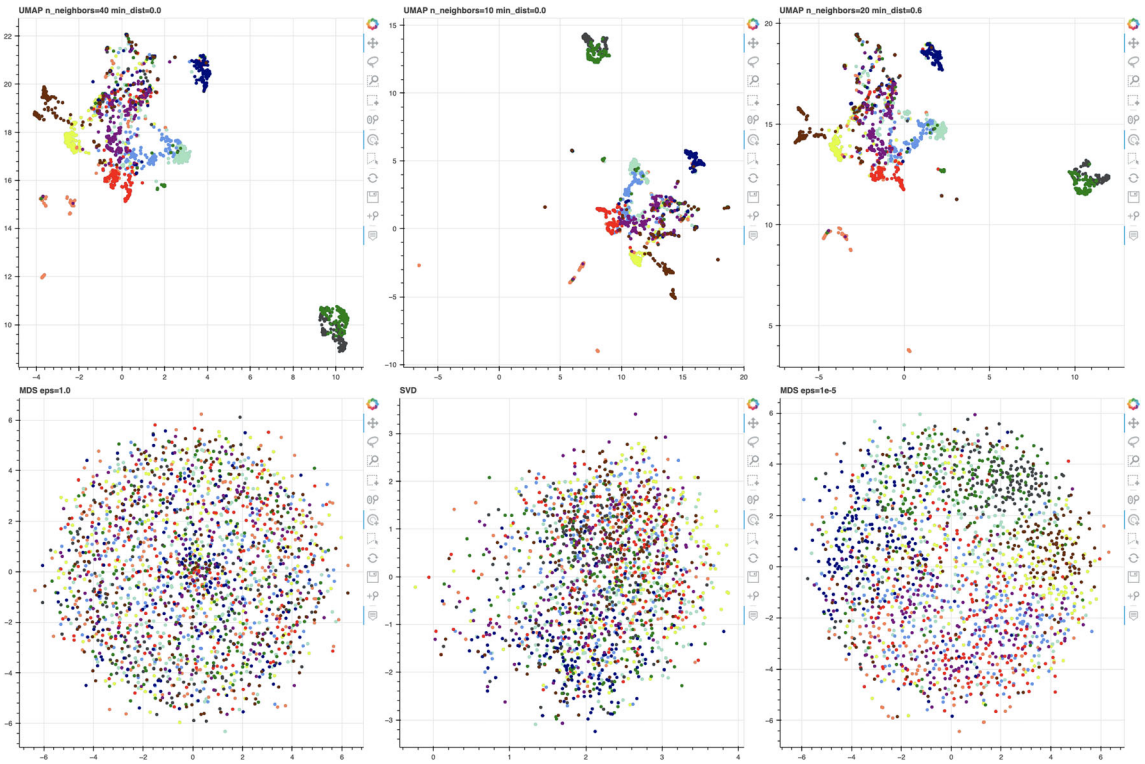

Fig. 3 Scatterplots evaluated as the highest and lowest scores with Weight 1

and those far apart continue to be distant in the transformed space. Conversely, the scatterplots that received lower scores failed to form clusters even though exhibiting some degree of class separability. This indicates that while these methods may differentiate between different classes to certain subregions, they do not effectively group similar data points into cohesive clusters in the reduced-dimensional space.

Table 5 demonstrates the DR methods that achieved the highest and lowest scores, respectively, using Weight 2. Furthermore, Fig. 4 illustrates the correlation between colors and product categories. We applied DR technique to visualize the category names while setting the order number to the categories. Specifically, MDS (Fig. 5) arranged categories so that those with intuitive similarities were positioned in proximity to one another. The order of these categories is provided in Sect. 4.2.1.

Particularly, MDS and some of UMAP and t-SNE emerged as superior in preserving the distances among data points. This preservation helps us understand the global relation in the dataset, enabling an accurate interpretation of the spatial distribution of clusters and individual points. Fundamentally, MDS achieved high scores in metrics assessing the conservation of distance between pairs of points, while UMAP with

	Gift Cards		Industrial and Scientific
	All Beauty		Automotive
	Grocery and Gourmet Food		Cell Phones and Accessories
	Amazon Fashion		Digital Music
	Arts, Crafts and Sewing		CDs and Vinyl

Fig. 4 Colors correspond to product categories

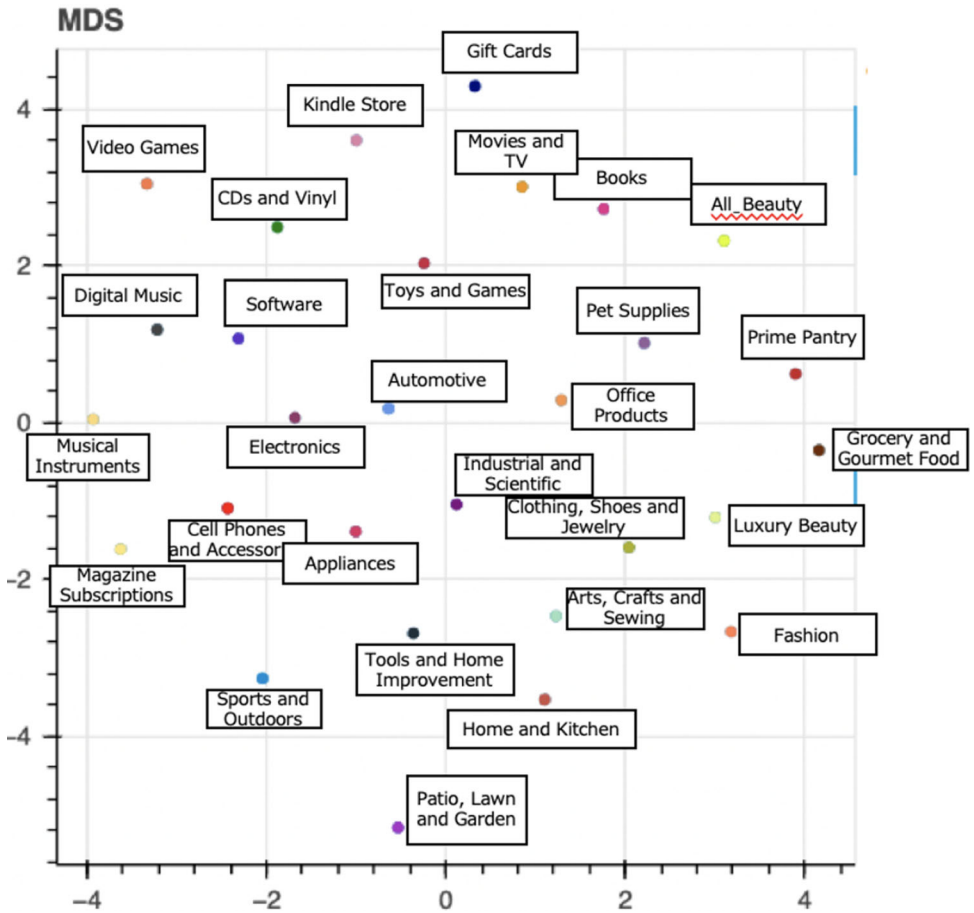

Fig. 5 Similar categories are arranged close

Table 5 DR methods which have the highest and lowest scores with Weight 2

rank	DR name	Parameters	Score
1	MDS	eps = 1e-5	0.84883
2	MDS	eps = 1e-15	0.82617
3	MDS	eps = 1e-10	0.81123
4	t-SNE	perplexity = 50 early_exaggeration = 20	0.63056
5	UMAP	n_neighbors = 50 min_dist = 0.6	0.61857
68	MDS*UMAP	default	0.21700
69	UMAP	n_neighbors = 10 min_dist = 0.99	0.20436
70	t-SNE	perplexity = 40 early_exaggeration = 30	0.20235
71	LDA	solver = “svd”	0.19767
72	LDA	solver = “eigen”	0.19767

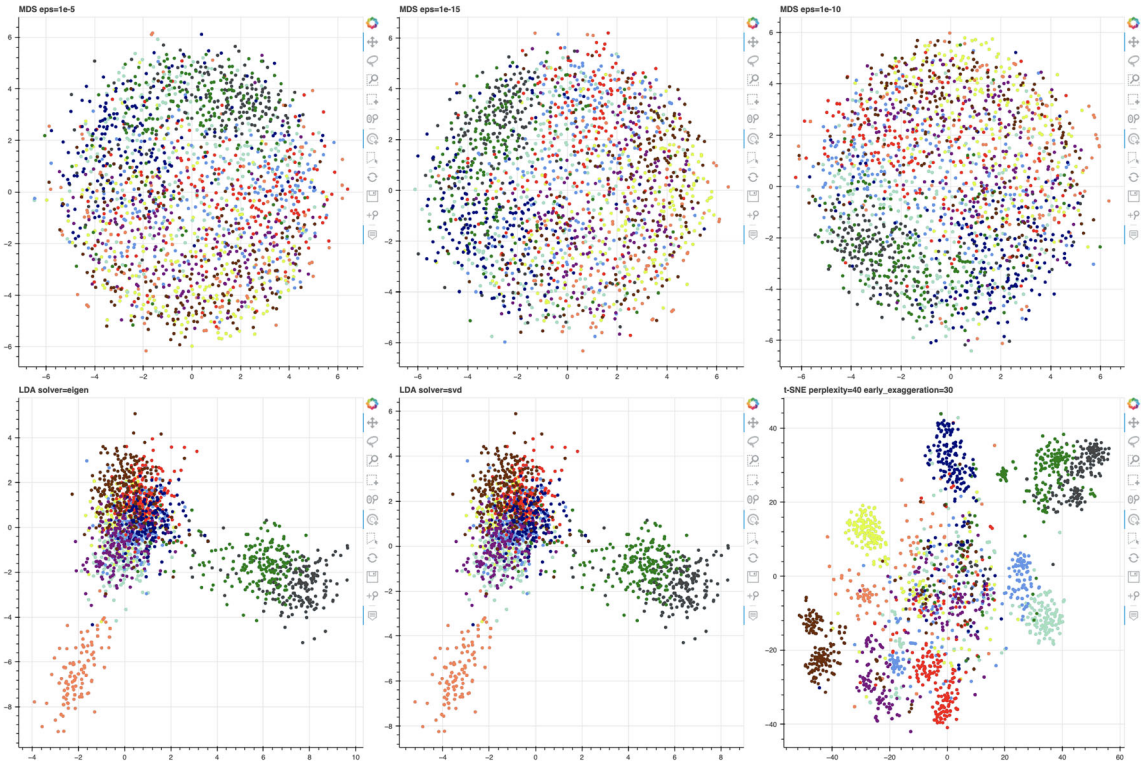

Fig. 6 Scatterplots evaluated as the highest and lowest scores with Weight 2

min_dist = 0.4 to 0.6 performed well in metrics of continuity of classes. This suggests that MDS maintains the spatial relationships accurately, and certain outcomes of UMAP and t-SNE are adept at capturing and interpreting the relationships among product categories derived from review texts. For instance, it is remarkable that "Amazon Fashion" and "All Beauty," "Digital Music," and "CDs and Vinyl" are located closely. Conversely, DR methods that scored lower on this weight, especially LDA, appeared to significantly alter the structure of the dataset to accentuate clustering, potentially at the expense of distorting the true distances between points. While considering with the outcomes of Weight 1, this observation indicates a fundamental trade-off in the application of DR techniques. This trade-off highlights the importance of selecting DR methods that align with the specific objectives of the analysis (Fig. 6).

In the evaluation applying Weight 3, as shown in Table 6 and Fig. 7, UMAP prominently features in achieving balanced results, demonstrating both effective clustering and preservation of point relationships. Results from t-SNE are scattered in the ranking. Basically, they have high scores for class separability and low scores for cluster separability and conservation of distance. The class continuity seems to vary greatly depending on the parameters. Moreover, SVD and MDS particularly with excessively large eps parameters resulted in the generation of scatterplots that did not offer significant insights or value for observation. This observation concludes the importance of selecting appropriate parameters and techniques based on the specific requirements and characteristics of the dataset to ensure the production of meaningful and informative visual representations.

The comprehensive analysis across the three distinct weighting schemes, from Weight 1 to Weight 3, illustrates that the effectiveness of visualization varies significantly based on the objectives, DR techniques, and parameters. In other words, we can obtain deeper insights and understanding of the data by examining a wide range of "effective" visualizations. This is the reason why we developed a dashboard (shown in Fig. 8) to showcase the best visualizations for a comprehensive and comparative analysis of the data.

Table 6 DRs which have the highest and lowest scores with Weight 3

Rank	DR name	Parameters	Score
1	UMAP	n_neighbors = 40 min_dist = 0.0	0.59365
2	UMAP	n_neighbors = 50 min_dist = 0.0	0.55819
3	UMAP	n_neighbors = 50 min_dist = 0.6	0.51560
4	UMAP	n_neighbors = 20 min_dist = 0.0	0.49730
5	UMAP	n_neighbors = 20 min_dist = 0.2	0.49521
68	UMAP	n_neighbors = 10 min_dist = 0.99	0.28875
69	t-SNE	perplexity = 50 early_exaggeration = 30	0.27910
70	t-SNE	perplexity = 40 early_exaggeration = 30	0.26674
71	SVD	default	0.21018
72	MDS	eps = 1.0	0.11266

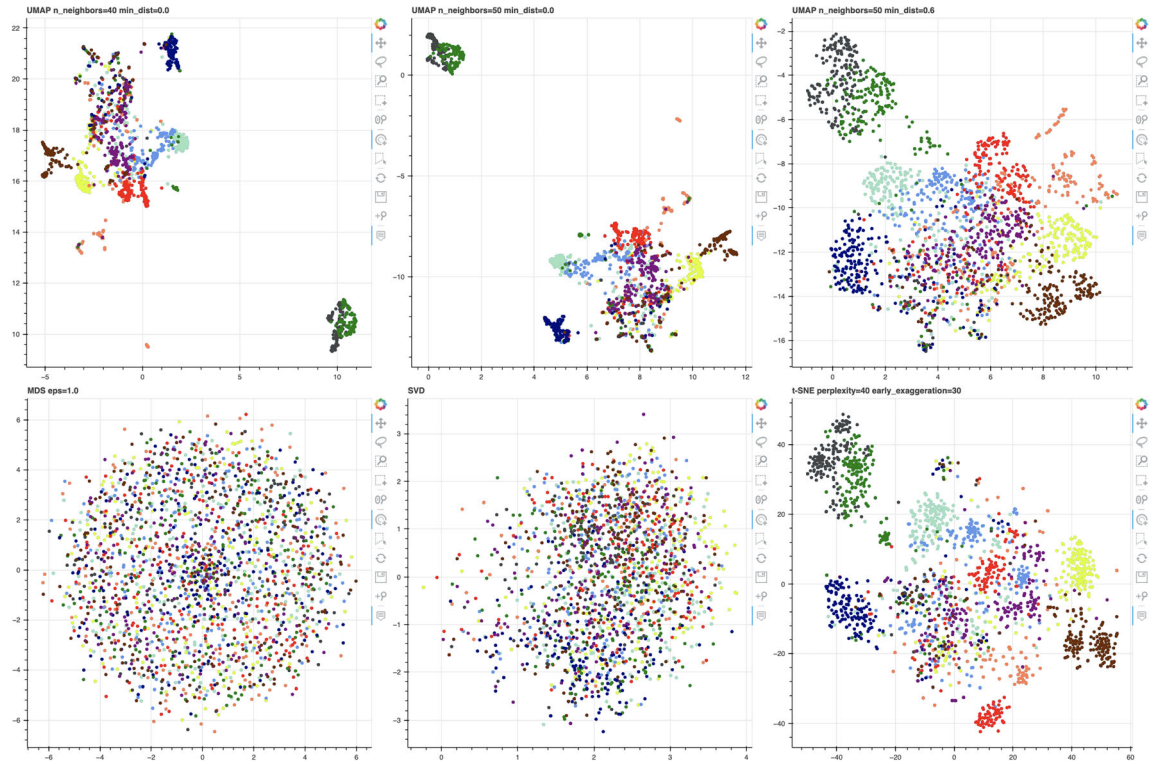

Fig. 7 Scatterplots evaluated as the highest and lowest scores with Weight 3

4.3 Case study 2: music features data

4.3.1 Dataset

In this example, we apply a dataset comprising 1,315 Japanese hit songs released between 1986 and 2018. The features of each song were extracted using RP extract (Wien 2015), consisting of dimensions 1,440 dimensions of RP, 60 dimensions of RH, and 168 dimensions of SSD. For the application example, we assigned the release decades of the songs as categories. Based on the criteria illustrated in Fig. 9, the release years of the songs were classified into six categories, with each category being assigned a unique color.

4.3.2 Evaluation score and visualization result

In a comprehensive view, it becomes apparent that the optimal DR method selected for case 2 (shown in Fig. 10) differs significantly from that chosen for case 1. As detailed in Table 7, LDA, which was not prominent in the rankings for case 1, achieves the highest score in case 2, with the resulting scatterplot exhibiting clear and distinguishable clusters. Especially, when compared scores to UMAP with smaller

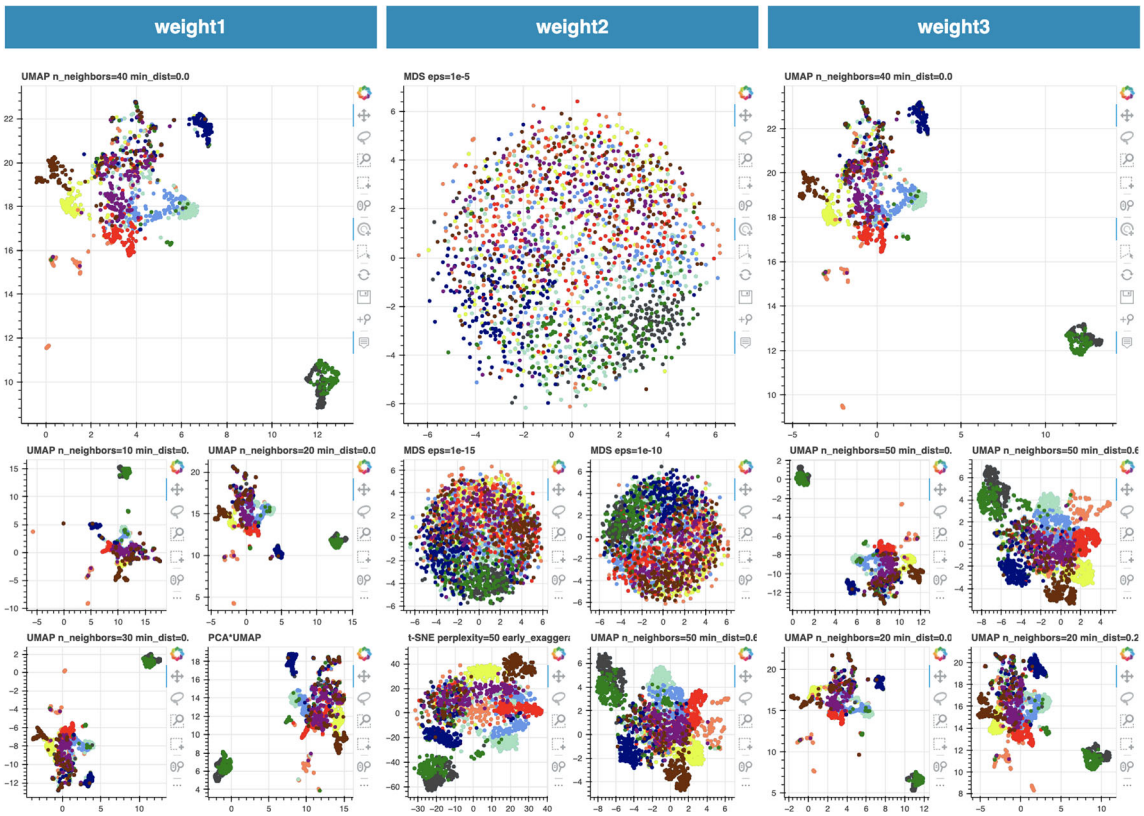

Fig. 8 Visualization dashboard shows the selected scatterplots

	1986 - 1991		2002 - 2007
	1992 - 1996		2008 - 2012
	1997 - 2001		2013 - 2018

Fig. 9 Colors assigned to the year in case study 2

Table 7 DR methods which have top five and worst five scores with Weight 1

rank	DR name	Parameters	Score
1	LDA	solver = svd	1.0000
2	UMAP	n_neighbors = 10 min_dist = 0.0	0.38420
3	UMAP	n_neighbors = 30 min_dist = 0.0	0.36984
4	UMAP	n_neighbors = 40 min_dist = 0.0	0.35892
5	UMAP	n_neighbors = 20 min_dist = 0.0	0.34876
68	UMAP	n_neighbors = 10 min_dist = 0.4	0.15754
69	UMAP	n_neighbors = 10 min_dist = 0.99	0.15692
70	t-SNE	n_neighbors = 40 min_dist = 0.99	0.15529
71	ICA	default	0.15289
72	MDS	eps = 1.0	0.13465

min_dist values ranking in second place and below, it indicates that LDA was uniquely effective in generating well-defined clusters for this dataset. On the other hand, several UMAP instances also scored low, a trend not observed in case 1. This indicates that UMAP was unable to produce distinct clusters in case 2.

In the evaluation applying Weight 2 (shown in Table 8), MDS and PCA got the highest scores. This indicates their superior capability to preserve the actual distances and continuity inherent in the original data. Intriguingly, LDA scored the lowest with Weight 2. This result provides a compelling insight: While LDA excels in forming distinct clusters, it does not accurately reflect the true distances between data points.

Table 8 DR methods which have top five and worst five scores with Weight 2

rank	DR name	Parameters	Score
1	MDS	eps = 1e-10	0.82142
2	MDS	eps = 1e-15	0.81765
3	MDS	eps = 1e-5	0.81615
4	PCA	svd_solver = auto	0.78959
5	PCA	svd_solver = randomized	0.78959
68	t-SNE	perplexity = 20 early_exaggeration = 40	0.32528
69	t-SNE	perplexity = 10 early_exaggeration = 50	0.31526
70	t-SNE	perplexity = 10 early_exaggeration = 10	0.31231
71	t-SNE	perplexity = 30 early_exaggeration = 30	0.29424
72	LDA	solver = svd	0.21428

Table 9 DR methods which have top five and worst five scores with Weight 3

Rank	DR name	Parameters	Score
1	LDA	solver = svd	0.60714
2	MDS	eps = 1e-10	0.55833
3	MDS	eps = 1e-5	0.55204
4	MDS	eps = 1e-15	0.53243
5	SVD	default	0.52898
68	t-SNE	perplexity = 10 early_exaggeration = 10	0.27613
69	t-SNE	perplexity = 40 early_exaggeration = 50	0.25986
70	t-SNE	perplexity = 30 early_exaggeration = 30	0.25191
71	t-SNE	perplexity = 20 early_exaggeration = 40	0.23382
72	t-SNE	perplexity = 10 early_exaggeration = 50	0.22099

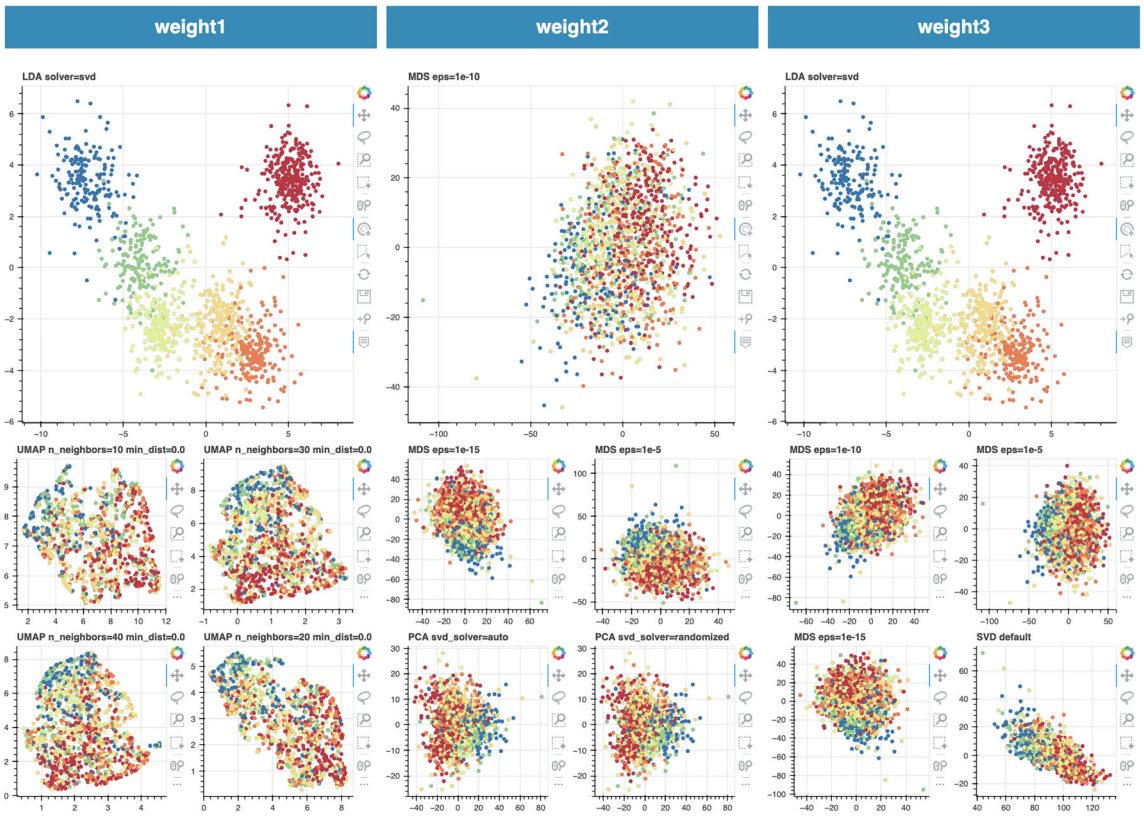

Fig. 10 Visualization dashboard shows the selected scatterplots in case 2

This suggests that LDA may compromise on distance accuracy to enhance the visibility of distinct characteristics, thereby distorting the spatial relationships to achieve clearer cluster differentiation.

With Weight 3 (shown in Table 9), LDA and MDS achieved high scores, attributed to the reasons previously mentioned. Conversely, t-SNE shows the poorest performance, primarily due to failure in forming clusters and its inability to preserve the relations among points.

4.4 Discussion

Through the comparison of the presented two case studies, it is evident that the optimal DR technique varies significantly across different datasets. The comparison reveals both commonalities and differences in the results of various DR as follows:

- UMAP with small min_dist consistently performs well in generating clusters across both datasets.
- Linear DR methods, such as MDS and PCA, accurately preserve the original distances between data points in both cases.
- LDA excels in creating well-separated class clusters within the music data, yet produces class mixed clusters when applied to review data.
- t-SNE is effective under certain parameter settings for review data, but it does not work for music feature data
- While several DR methods generate clusters effectively in the review data, only LDA manages to do so for the music features data.

DR techniques are underpinned by their unique algorithms, which inherently influence its performance, as evidenced by the characteristics observed in the above commonalities. Here, these cases serve as a clear illustration of the principle that the suitability of a DR can vary greatly depending on the specific characteristics in the data. It highlights the importance of choosing the preferable techniques to effectively reveal the desired features in a dataset.

5 Conclusion and future work

This paper proposed the technique for selecting DR methods for multidimensional data, based on scoring generated scatterplots. This technique automatically computes scores using various weighted metrics for scatterplots produced by applying different DR methods and parameters. We explored 72 DR patterns in this study; here, the potential combinations of DR methods, parameters, DR combinations, the dimensionality in the initial DR phase when combining methods, and class assignments, suggest an expansive number of possibilities. Efficient scatterplot selection techniques are therefore required. We presented case studies involving e-commerce review data and music feature data, demonstrating that the optimal DR method varies according to the characteristics of the dataset.

As future work, we would like to evaluate the weighting of each metric used in scatterplot scoring. In our case studies, the selection of weights was based on the hypotheses and subjective of the authors. We would like to develop methods to determine optimal weights for a more effective selection of insightful scatterplots. Exploration of alternative metrics for scatterplot evaluation would be another issue. In our dashboard, while we comprehensively list scatterplots with the highest scores, the inclusion of similar scatterplots could be avoided. Since the dashboard should provide the selected DR methods, parameters, and characteristics of data, thus displaying every similar scatterplot may not be necessary. Further scatterplot selection techniques such as those proposed by Itoh et al. (2023) could be utilized to refine the redundant scatterplot selection results. Lastly, we would like to apply this application to other types of datasets and verify their effectiveness.

References

Agis D, Pozo F (2019) A frequency-based approach for the detection and classification of structural changes using t-sne. Sensors 19(23):5097

Anowar F, Sadaoui S, Selim B (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). Comput Sci Rev 40:100378

Aupetit M, Sedlmair M (2016) Sepme: 2002 new visual separation measures. In: 2016 IEEE pacific visualization symposium (PacificVis), pp. 1–8. IEEE
Ayesha S, Hanif MK, Talib R (2020) Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fus 59:44–58
Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27
Dang TN, Wilkinson L (2014) Scagexplorer: Exploring scatterplots by their scagnostics. In: 2014 IEEE Pacific visualization symposium, pp 73–80. IEEE
Engel D, Hüttenberger L, Hamann B (2012) A survey of dimension reduction methods for high-dimensional data analysis and visualization. In: Visualization of large and unstructured data sets: applications in geospatial planning, modeling and engineering-proceedings of IRTG 1131 Workshop 2011. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
Fodor IK (2002) A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US)
Gao T (2021) Simcse: simple contrastive learning of sentence embeddings
Gao T, Yao X, Chen D (2021) Simcse: simple contrastive learning of sentence embeddings. arXiv preprint. arXiv:2104.08821
Harrison L, Yang F, Franconeri S, Chang R (2014) Ranking visualizations of correlation using weber's law. IEEE Trans Visual Comput Graph 20(12):1943–1952
Nadia Syed HS, Jamil NW (2023) A comparative study of hybrid dimension reduction techniques to enhance the classification of high-dimensional microarray data. In: 2023 IEEE 11th conference on systems, process & control (ICSPC), pp 240–245
Heiser CN, Lau KS (2020) A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. Cell Rep 31(5)
Huang H, Wang Y, Rudin C, Browne EP (2022) Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. Commun Biol 5(1):719
Huang S, Ward MO, Rundensteiner EA (2005) Exploration of dimensionality reduction for text visualization. In: Coordinated and multiple views in exploratory visualization (CMV'05), pp 63–74. IEEE
Itoh T, Nakabayashi A, Hagita M (2023) Multidimensional data visualization applying a variety-oriented scatterplot selection technique. J Visual 26(1):199–210
Lee JH, McDonnell KT, Zelenyuk A, Imre D, Mueller K (2013) A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. IEEE Trans Visual Comput Graph 20(3):351–364
Malik HK, Al-Anber NJ, Al-Mekhlafi FAE (2023) Comparison of feature selection and feature extraction role in dimensionality reduction of big data. J Tech 5(1):184–192
Matute J, Telea AC, Linsen L (2017) Skeleton-based scagnostics. IEEE Trans Visual Comput Graph 24(1):542–552
Nanga S, Bawah AT, Acquaye BA, Billa MI, Baeta FD, Odai NA, Obeng SK, Nsiah AD (2021) Review of dimension reduction methods. J Data Anal Inf Process 9(3):189–231
Ni J (2018) Amazon review data
Padron-Manrique C, Vázquez-Jiménez A, Esquivel-Hernandez DA, Martinez Lopez YE, Neri-Rosario D, Sánchez-Castañeda JP, Giron-Villalobos D, Resendis-Antonio O (2022) Diffusion on PCA-UMAP manifold captures a well-balance of local, global, and continuum structure to denoise single-cell RNA sequencing data. bioRxiv, pp 2022–06
Remeseiro B, Bolon-Canedo V (2019) A review of feature selection methods in medical applications. Comput Biol Med 112:103375
Saini O, Sharma S (2018) A review on dimension reduction techniques in data mining. Comput Eng Intell Syst 9(1):7–14
Sedlmair M, Tatu A, Munzner T, Tory M (2012) A taxonomy of visual cluster separation factors. In: Computer graphics forum, vol 31, pp 1335–1344. Wiley Online Library
Singh KN, Devi SD, Devi HM, Mahanta AK (2022) A novel approach for dimension reduction using word embedding: an enhanced text classification approach. Int J Inf Manage Data Insights 2(1):100061
Sips M, Neubert B, Lewis JP, Hanrahan P (2009) Selecting good views of high-dimensional data using class consistency. In: Computer graphics forum, vol 28, pp 831–838. Wiley Online Library
Stolarek I, Samelak-Czajka A, Figlerowicz M, Jackowiak P (2022) Dimensionality reduction by umap for visualizing and aiding in classification of imaging flow cytometry data. Iscience 25(10)
Van Der Maaten L, Postma EO, van den Herik HJ et al (2009) Dimensionality reduction: a comparative review. J Mach Learn Res 10(66-71):13
Vashisth P, Meehan K (2020) Gender classification using twitter text data. In: 2020 31st Irish signals and systems conference (ISSC), pp 1–6. IEEE
Wang K, Yang Y, Fangjiang W, Song B, Wang X, Wang T (2023) Comparative analysis of dimension reduction methods for cytometry by time-of-flight data. Nat Commun 14(1):1836
Wang Y, Wang Z, Liu T, Correll M, Cheng Z, Deussen O, Sedlmair M (2019) Improving the robustness of scagnostics. IEEE Trans Visual Comput Graph 26(1):759–769
Wien T (2015) Music information retrieval
Wilkinson L, Anand A, Grossman R (2005) Graph-theoretic scagnostics. In: Information visualization, IEEE symposium on, pp 21–21. IEEE Computer Society
Yamada I, Asai A, Sakuma J, Shindo H, Takeda H, Takefuji Y, Matsumoto Y (2018) Wikipedia2vec: an efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. arXiv preprint. arXiv:1812.06280