EvoSets: Tracking the Sensitivity of Dimensionality Reduction Results Across Subspaces

Guodao Sun, Sujia Zhu, Qi Jiang, Wang Xia, and Ronghua Liang

Abstract—Dimensionality reduction is commonly used for identifying and analyzing patterns in the visual analysis of multi-dimensional datasets. The selection of subspaces is a core building block in projecting high-dimensional data to low-dimensional space, which is usually illustrated as a scatterplot for analysts to easily understand and explore. This process involves human prior knowledge and domain-specific requirements. Thus, quantifying and tracking the changes of dimensionality reduction results across subspaces remain challenging. Existing methods can neither quantify the subsets-based changes of dimensionality reduction results when switching subspaces, nor automatically and comprehensively display the overall and subtle differences among dimensionality reduction results. To address this, we developed *EvoSets*, a novel visual analytics system designed to help users understand how subspaces affect dimensionality reduction results. The effects are quantified based on the distribution of subsets within projections to tracking the sensitivity of dimensionality reduction results across subspaces. In addition, the system supports the exploration of the overall evolution of the dimensionality reduction results for helping users track the convergence and divergence behavior changes of subsets based on an extended *Bubble Sets* visualization. Similarities are intuitively illustrated, and dissimilarities are highlighted among the generated dimensionality reduction results across subspaces based on different layout constraints. The usefulness and effectiveness of the system are further evaluated with a user study and two case studies on multi-dimensional datasets.

Index Terms—Subspace, dimensionality reduction result, convergence and divergence behavior changes, comparison

1 Introduction

Subspace analysis is often performed before visual exploration for a complex high-dimensional dataset because the dataset is commonly mapped into an intuitive and easy-to-understand low-dimensional space according to a specific subspace for further patterns recognition. Hence, selecting a reasonable subspace is one of the most common and basic steps [40] in the process of dimensionality reduction. However, many aspects must be considered in determining subspaces, such as domain-specific requirements and experts' knowledge. A plethora of approaches has been proposed to obtain a particular dimensionality reduction result, such as automatic subspace analysis [3], projection pursuit [21], and the clustering method [18]. However, certain significant patterns may be uncovered in different subspaces, certain discernible patterns may be redundant in all subspaces, and certain relevant patterns may facilitate pattern identification in some subspaces. Thus, distinguished subspaces result in various dimensionality reduction results, where the relative correlation of data instances changed. For example, tigers and wolves are similar in terms of carnivory and habitat, but tigers and cats are more similar than tigers and wolves with respect to genetic composition and social grouping. Tracking the grouping changes across subspaces has the potential of helping domain experts to understand the evolutionary relationship among species. Thus, analyzing how subspaces affect dimensionality reduction results has theoretical significance and is necessary.

The analysis of how subspaces affect dimensionality reduction results is hindered by three major challenges. First, the combination of dimensions could increase exponentially with the increasing dimensions, hindering the efficient exploration of high-dimensional data. Thus, visually comparing all possible subspaces and their dimensionality reduction results is excessive, and visually recognizing all possible patterns and filtering redundant patterns are challenging for users. Automatic approaches were proposed in previous works to reduce the effort of exploring potential and reasonable subspaces [3], [35]. However, these methods or models must often be adapted to domain-specific requirements. Second, quantifying the changes of dimensionality reduction results across distinguished subspaces is challenging. Typical point-based approaches were proposed to merely compare the similarities or dissimilarities on the points' location in different dimensionality reduction results. However, these methods do not reveal the correlation changes of subsets within dimensionality reduction results across subspaces. To efficiently discover and intuitively track patterns, such as behavior changes of divergence or convergence among subsets, users must locate and correlate detailed variations with respect to the structure of subsets when switching subspaces. Third, the intuitive

• *The authors are with the Zhejiang University of Technology, HangZhou 310023, China. E-mail: {godoor.sun, zhusujiashiny}@gmail.com, {jiangqi, xiawang, rhliang}@zjut.edu.cn.*

Manuscript received 24 Oct. 2020; revised 18 Mar. 2021; accepted 6 May 2021. Date of publication 11 May 2021; date of current version 11 Nov. 2022. (Corresponding author: Ronghua Liang.) Recommended for acceptance by R. Hong. Digital Object Identifier no. 10.1109/TBDATA.2021.3079200

visual design remains a major challenge for tracking the co-evolution of multiple dimensionality reduction results across subspaces. In addition, visually ascertaining the similarities and dissimilarities among dimensionality reduction results is difficult for users. Much research has been devoted to investigating the evolution visualization of subsets [7], [12], [38]. However, existing methods merely focus on the presentation of a series of dimensionality reduction results, and neglect the detailed side-by-side comparison of subsets [22], [45]. Users need manually compare overall or subtle differences on dimensionality reduction results. Therefore, presenting the global evolution of subsets while preserving local information is also challenging.

To address the issues above, we propose *EvoSets*, an interactive visual analysis system, for revealing how the selection of subspace influences dimensionality reduction results. For the first issue, subspaces are generated automatically and/or manually to decrease the number of subspace explorations. We also illustrate the characteristic of dimensions by general statistical approaches to guide users in interactively selecting dimensions of interest. For the second issue, we propose an approach to quantifying how dimensionality reduction is affected by subspaces based on the calculation of the distribution of subspaces. For the third issue, the *EvoSets* allows users to participate in the visual exploration of pattern recognition, globally and locally. For parsing and illustrating the global correlation of dimensionality reduction results, we propose an approach to quantifying the changes of dimensionality reduction results from the perspective of subsets. We also utilize a matrix chart to visually quantify the correlations among subspaces and the overall impact of subspaces on dimensionality reduction results. For parsing and presenting a local comparison, we extended *Bubble Sets* visualization [10] to express the co-evolution of subsets' structural changes among the scatterplots across different subspaces. We also define three kinds of layout constraints to filter out special points, considering that the local detail changes are redundant.

In addition, we demonstrate the usefulness and effectiveness of our methods by conducting two case studies on commonly used datasets and a user study on visual design and task analysis. In summary, the key contributions of this work are as follows.

- We propose a measurement for quantifying how subspaces affect dimensionality reduction results based on the distribution of subsets within dimensionality reduction results.
- We designed *EvoSets*, a visual analytics system that aims to identify the correlation of subspaces and compare their dimensionality reduction results, globally and locally.
- We provide empirical findings from the dynamic exploration process of the dimensionality reduction results based on a user study and two case studies on real-world datasets.

2 Background and Related Work

This section reviews related works on searching for subspaces, comparison of scatterplots, and visualization of subsets transition.

2.1 Subspaces Searching

Subspaces play a vital role [49] in the process of dimensionality reduction because many visual analysis works on high-dimensional data are based on projection. Subspace selection in a large search space is a core building block in projecting high-dimensional data to low-dimensional space, which is susceptible to human prior knowledge and task requirements. Among of the issues, scalability is a key challenge in dealing with high-dimensional data in subspace projections. For a dataset with d dimensions, the number of possible subspaces could be 2^{d-1}, which impose a heavy cognitive load on analysts.

Exploring all subspaces is unnecessary for identifying patterns because some subspaces may contain redundant patterns. Therefore, many researchers have made some contributions to filtering valid and significant subspaces by efficiently comparing or ranking dimensionality reduction results, such as automatic subspace analysis [11], [44], projection pursuit [21], [26], or clustering method [27], [34], [45]. Automatic subspace analysis [2], [43], [51] decreases the number of subspace explorations by measuring the indicators of subspaces, such as interestingness, importance, and entropy. Projection pursuit approaches [2], [26] aim to find significant projections of high-dimensional datasets. Clustering approaches [3], [23] seek to identify clusters among various subspaces within a dataset-based clustering algorithm.

Considering that the above approaches may miss some significant subspaces of user interest, some studies focused on constructing and refining multi-dimensional subspaces. They leveraged interactive technology [33], [35] to filter subspaces [39] to decrease the cost of searching subspaces. Users could control the output of a subspace by adding or removing attributes. Guo *et al.* [17] and May *et al.* [30] helped analysts select interesting feature dimensions by interacting with the algorithm outputs. Krause *et al.* [24] proposed SeekAView, a transparent and flexible process-oriented approach that helps analysts build subspaces out of a high-dimensional dataset. Pérez *et al.* [35] provided users with more control over the process, including the weight strength and the feature extension. Paiva *et al.* [33] proposed an interactive projection technique that allows users to perfect the dimensionality reduction results by integrating their prior knowledge. Analysts can gradually improve the visual quality of the projection by manually making the objects in the same cluster nearer or separating non-components. However, the obtain satisfying results through the process may be tedious. In addition, these methods neglect the detailed side-by-side comparison of subspaces, which might make results difficult to interpret.

2.2 Projection Comparison

Researchers have proposed several techniques [6], [29], [48] to project a high-dimensional dataset into low-dimensional space, which is usually visualized as a two-dimensional scatterplot for analysts to easily understand and explore. However, the dimensionality reduction results generated by various subspaces have dissimilarities, which reflect how subspaces affect dimensionality reduction results. This subsection mainly introduces approaches to identifying these characteristics of multiple scatterplots.

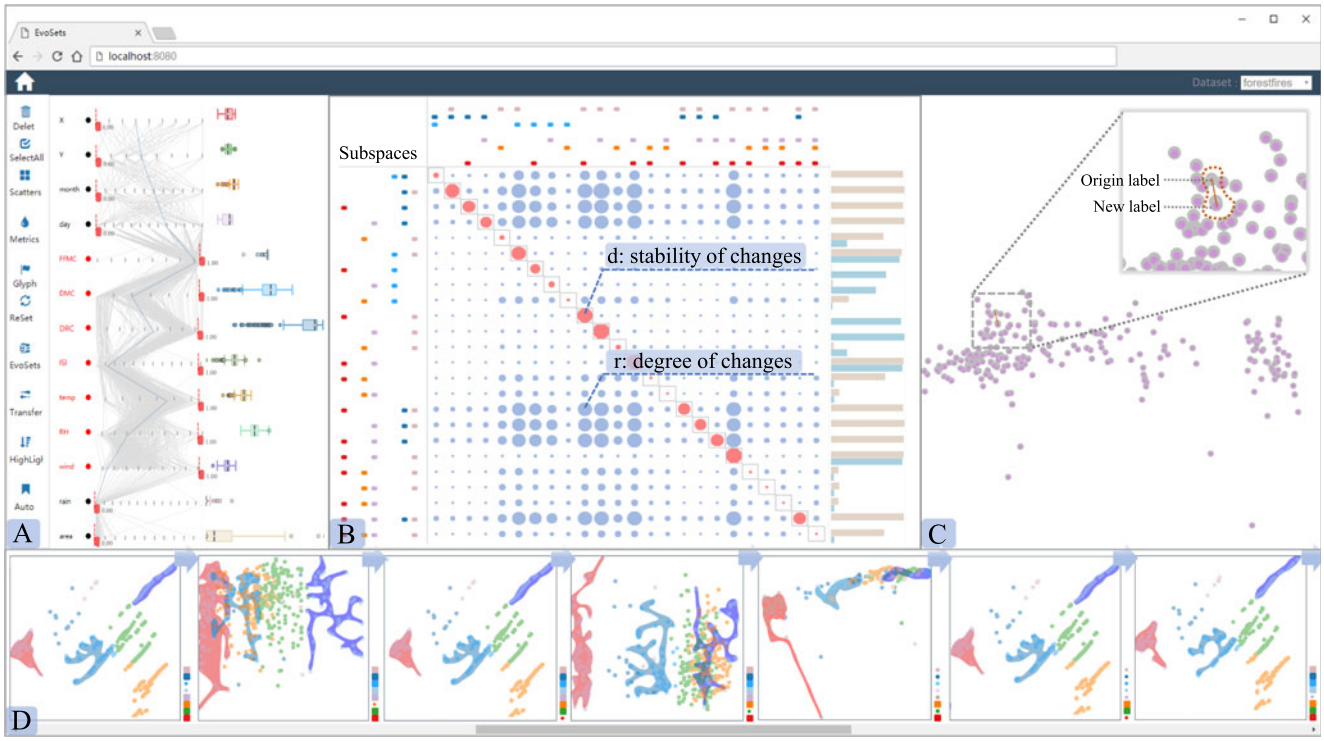


Fig. 1. *EvoSets*: The user interface starts with dimensions combination view (A), which illustrates characteristics of dimensions. The matrix-based correlation view (B), which allows users to identify the similarity among projections across various projections. The projections subtraction view (C), which assists analysts to efficiently identify the subtle information. The evo-subsets view with extended *Bubble Sets* visualization (D), which intuitively tracks the co-evolution of significant subsets.

The goal of most research on projection is finding high-quality dimensionality reduction results for visual exploration by comparing dimensionality reduction results [5], [25], [46]. Researchers have compared, quantified, and ranked scatterplots with some considerations such as the tightness within the cluster, the separation among clusters, and visual clutter [13], [15]. Sips *et al.* [41] qualified and compared different scatterplots in terms of distance and distribution consistency for the selection of high-quality projection. Schaefer *et al.* [40] integrated the quality measure of a given embedding based on structural preservation and visual clutter avoidance. Andrada *et al.* [43] presented the assessment of density and separateness for clusters in the distribution of unclassed and classed data. Some researchers [32], [42] leveraged the statistical information of scatterplots to compare groups within projections.

Some researchers also investigated the correlation of dimensionality reduction results by calculating dissimilarities [14] or similarities [22], [28] among scatterplots with defining layout algorithms or constraints [8]. For differnces-based computation of dimensionality reduction results, Dirk *et al.* [26] presented dissimilarity-based measurement by discarding affine transforms of projections and preventing redundancies of the same data patterns. Fujiwara *et al.* [14] proposed an automatic tracking approach, which can help analysts follow the common changes, such as the animated transitions of outliers and the newly forming cluster. For similarities-based computation of dimensionality reduction results, Wang *et al.* [45] utilized matrix visualization to illustrate dimensionality reduction results, and pixel-wise mean square error to quantify the similarity between two dimensionality reduction results. They also cluster and construct a hierarchical tree for efficiently selecting and comparing topological similarity in the matrix views based on Tanimoto distance. Jackle *et al.* [22] described a method for transforming point-based distance matrices to 1D feature vectors for computing a similarity between two subspace projections. Ma *et al.* [28] introduced a deep-learning-based approach for characterizing perceptual similarities among ScatterNets.

However, these methods do not automatically quantify how subspaces affect dimensionality reduction results, how subspaces affect the structure of subsets, and how to visualize the detailed correlation changes among data instances. In this work, we quantify how dimensionality reduction results change across subspaces, visualize the divergence and convergence behavior changes of subsets, and reveal the detailed changes of data records across subspaces with three kinds of layout constraints. Thus, users can comprehensively obtain the co-evolution of dimensionality reduction results across subspaces globally and locally.

2.3 Subsets Transition Visualization

This work includes revealing the co-evolution of subsets in different projections. We define this co-evolution of subsets in subspaces as subsets transition. This section reviews related works on the transition visualization of subsets.

For the visualization of various subsets in one set, Dinkla *et al.* [1] presented Kelp Diagrams, a novel method for depicting set relations over points. The algorithm considers the aspects of aesthetic standards and efficiency and balances visual complexity from the data. Meulemans *et al.* [31] proposed KelpFusion, which uses the shortest path diagram

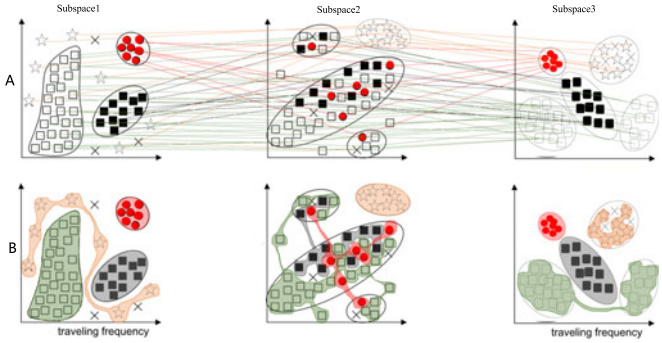


Fig. 2. High-dimensional dataset is projected into scatterplots and the group correlation have been changed across three subspaces. The transitions of structural changes are depicted using two visual approaches. Jäckle *et al.* [22] leveraged lines to connect and track data instances of selected subsets (A), Our proposed visual analysis method utilizes *Bubble Sets* visualization to wrap the subset, which has been evaluated an effective approaches [10] to maximally distinguish subsets, and it also express subsets' convergence or divergence behavior changes (B).

to define the linear characteristics of the visualization and integrates hull and line-based technologies to visualize collection members. All these methods can distinguish different subsets and perform well in scenarios of accuracy, time complexity, and adaptation. However, expressing the structural state of the set is challenging, especially the convergence and divergence in our research.

On the basis of the idea of tracking a single record over time and visualizing the changes of such single records in the projections, some approaches have been proposed in recent years [4], [20], [47]. Yuan *et al.* [50] presented two kinds of visual approaches, namely, "Dimension Projection Matrix" and "Dimension Projection Tree", which are designed for visually exploring high-dimensional data correlation and attributes correlation. Liu *et al.* [27] introduced the transition graph of dynamic projections to navigating exploration of dimensionality reduction results. Wang *et al.* [45] utilized multiple matrix visualization to track the relative distance changes of data items. However, matrix visualization can not intuitively illustrate the cluster information. Jäckle *et al.* [22] proposed *Pattern Trails*, shown in Fig. 2A, to visualize projections of subspaces side-by-side and indicate changes among adjacent patterns in subspaces through linked representation. Poco *et al.* [36] leveraged enclosed surfaces to visualize 3D projections, which illustrate the geometrical information derived from clusters. However, these works only focused on the visual presentation among subsets and did not quantify how "changeable" these dimensionality reduction results are. Thus, we propose a layout algorithm based on the distribution of subsets for automatically revealing the convergence or divergence behavior changes. We also present extended *Bubble Sets* visualization [10] for tracking these changes. Compared with *Pattern Trails* in Fig. 2A, the proposed visualization of *Bubble Sets* in Fig. 2B could not only distinguish different subsets in a specific projection result, but also help express the convergence or divergence behavior of subsets among different results.

3 Task Analysis and Workflow

This section describes user tasks after our literature review that has discussed above. We survey the relevant works on subspace searching and projection comparison and provide summary of these researches in Fig. 4. We also introduce the workflow of exploring the impact of subspaces on dimensionality reduction results.

3.1 Task Analysis

We aim to automatically ascertain the changes of multiple dimensionality reduction results across subspaces. To help users understand how subspaces affect dimensionality reduction results, we summarized the following core analysis tasks that appear most frequently in the domain literature.

- T1 *Supporting interactive subspaces generation.* The design must guide users in creating random subspaces. Automatic approaches to subspaces generation may lead to the loss of certain significant subspaces. Therefore, considering that subspaces are combined with dimensions, the characteristics of dimensions must be visualized to guide users in interactively generating subspaces.
- T2 *Summarizing overall changes of dimensionality reduction results globally.* The design must provide a compact visual summary of the projections' transition. An automatic approach is proposed to quantify the changes of dimensionality reduction results based on the distribution of subsets. Considering that the information is directed correlation and we must illustrate the transition of projections, we employ a matrix chart to visually summarize the overall changes of the scatterplots across subspaces.
- T3 *Providing a visual metaphor of subsets' convergence and divergence behavior changes.* The design The design is required to efficiently distinguish various subsets and intuitively interpret the subsets' divergence or convergence behavior changes. Dimensionality reduction results inevitably appear in groups, which may exhibit changes in group members across different subspaces. The visual encoding of convergence and divergence patterns is motivated by *Bubble Sets* visualization in which the former occurs when two or more subsets merge into a subset and the latter occurs when a subset spread into two or more subsets. This visual metaphor of bubble subsets allows us to highlight significant information where subsets converge to similarity or diverge from similarity.
- T4 *Relaxing visual clutter.* The design must provide an intuitive and pleasing visualization of the co-evolution of the structural changes of subsets. The crossing and overlapping of bubble subsets may create visual clutter, which can hinder users from exploring and investigating pattern recognition. The visualization of bubble subsets should be extended to relax visual clutter.
- T5 *Highlighting and unfolding differences between dimensionality reduction results locally.* The design must allow a user to compare dimensionality reduction results in detail. Considering that visualizing all differences may increase the burden on users and block the exploration and identification of significant patterns, we defined specific layout constraints (i.e.,

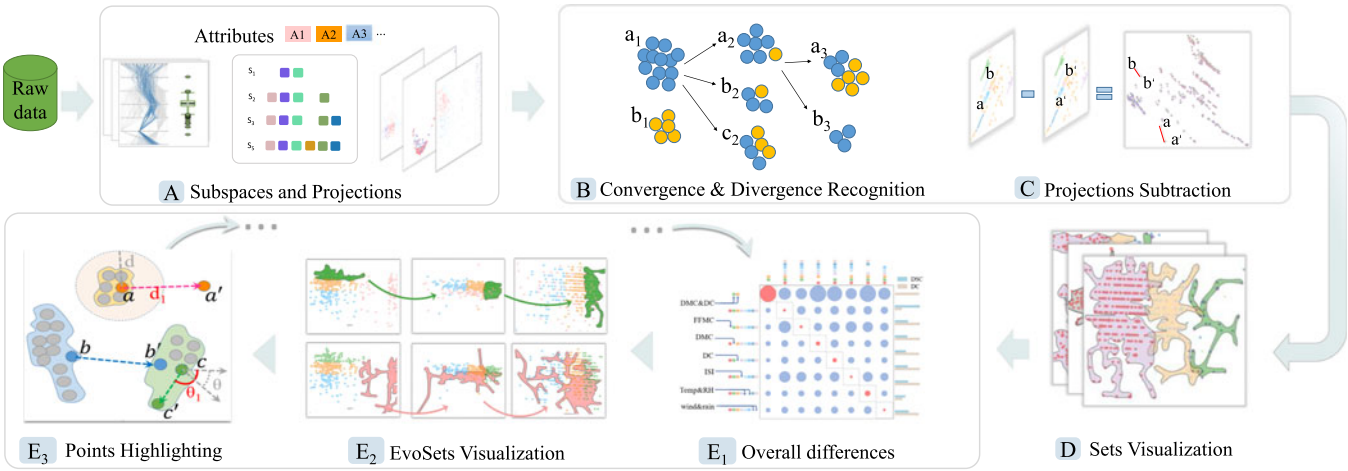


Fig. 3. The explanation pipeline can be divided into the following steps: Dimensions statistics and subspaces generation automatically or manually for creating dimensionality reduction results (A). Ascertaining subsets' behavior changes of convergence or divergence based on the distribution of subsets across various subspaces (B). Comparison of two projections with geometric transformation (C). Extended *Bubble Sets* visualization for quickly distinguishing subsets, recognizing subsets' behavior changes, and maintaining group correlations of data points (D). Visualization of comparison globally and locally (E_1, E_2, E_3). Overall correlation among different dimensionality reduction results (E_1). Tracking of the co-evolution of subsets within dimensionality reduction results (E_2). Highlighting of notable data points with tree kinds of layout constraints (E_3).

distance, angle, and cluster constraints) to quickly find notable data instances.

These requirements help us obtain appropriate design principles and make informed decisions about our visual design.

3.2 Analysis Workflow

The analysis workflow of our proposed system is shown in Fig. 3. First, the workflow starts from the general statistical analysis of dimensions for illustrating the characteristics of dimensions, which guide users in implementing dimensions combination to generate dimensionality reduction results (A). Second, the effects of subspaces on dimensionality reduction results globally (B) and locally (C) are revealed. For parsing global information, we ascertain the convergence or divergence behavior changes of subsets based on the distribution of subsets. For parsing local information, we identify the detailed changes between projections with three kinds of layout constraints. Third, for tracking the transition of dimensionality reduction results across distinguished subspaces, we extended *Bubble Sets* visualization (D) and defined difference-based comparison to visually display the changes of projections from different perspectives, the overall correlation of different projections (E_1), and the tracking of the subsets' co-evolution within projections (E_2), and highlight the data points of notable changes (E_3).

Research	FCS	IGC	SC	CC	PC	DC	VDM	RVC	ILC
Liu et al.[27]	✓	✓	-	-	-	-	-	-	-
Lehmann et al.[26]	-	-	-	✓	✓	-	-	-	-
Fujiwara et al.[14]	-	-	-	-	-	-	-	-	✓
Jackle et al.[22]	✓	✓	✓	-	✓	-	✓	✓	-
Wang et al.[45]	-	✓	✓	-	✓	-	-	-	-
Ma et al.[28]	-	-	✓	-	✓	-	-	✓	-
Tatu et al.[44]	-	✓	✓	-	✓	-	-	-	✓
EvoSets	✓	✓	-	✓	-	✓	✓	✓	✓

Fig. 4. Task analysis. FCS represents the free creation of subspaces. IGC signifies the identification of global changes on projections. SC and CC denote similarity-based computation and changes-based computation of projections respectively. PC and DC denote point-based computation and subsets-based computation of changes on projections respectively. VDM describes the visual design metaphor. RVC serves for visual clutter relaxing. ILC depicts the identification of local changes on projections. ✓ depicts the existence of a certain function, and - depicts the (possible) absence of a certain function.

4 Measuring Changes

Our goal is to track the co-evolution of multiple dimensionality reduction results globally and uncover the detailed changes between dimensionality reduction results locally.

4.1 Projections Quantification

To illustrate the impact of subspaces on the dimensionality reduction results, we evaluate a single projection and calculate the changes among multi-dimensional reduction results across various subspaces. For the former, we employ a projection assessment approach [41], which leverages the distance consistency (DC) and distribution consistency (DSC) of scatterplots to quantify the dimensionality reduction results. For the latter, we introduce an approach to measuring the changes of dimensionality reduction results across various subspaces based on the distribution of subsets.

Data Space and Subsets Definition. Let $X \subseteq R^d$ be a dataset with d dimensions. Datasets are divided into m subsets under certain subspace $C(X) = \{c_1, c_2, c_3, \ldots, c_m\}$, while c_i denotes a subset.

Measuring the Changes Among Multiple Projections. Qualitative or quantitative measurement is complicated and redundant when measuring the dissimilarities of dimensionality reduction results from the perspective of data points. Because the changes of projections include not only the position and the direction, but also the shifting of different classes, which complicates the measurement of the changes of dimensionality reduction results. Speak more specifically, a scatterplot is composed of a series of subsets whose behavior

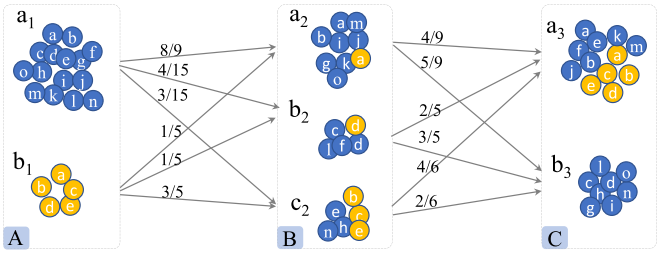


Fig. 5. Illustration of the identifying the subsets' behavior changes of convergence or divergence based on the distribution of subsets to reveal evolution of projections across subspaces.

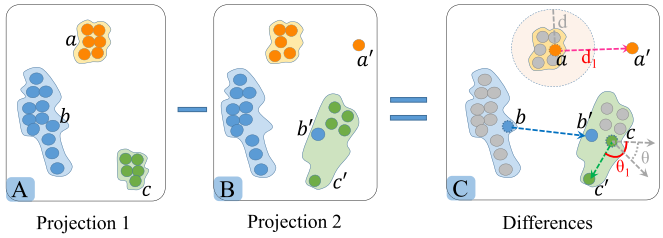


Fig. 6. The differences identification between Projection 1 and Projection 2. Abnormal points are highlighted with constraints, such as distance constraint(from point a to point $a^{'}$), clustering constraint(from point b to point $b^{'}$), and angle constraint(from point c to point $b^{'}$).

changes reflect the correlations of the inner members within subsets. These structure changes on scatterplots indicate how subspaces affect dimensionality reduction results. Therefore, we quantify the evolution of dimensionality reduction results across various subspaces from the perspective of the distribution of subsets, rather than data points.

Fig. 5 illustrates the process of qualifying the structural changes of dimensionality reduction results. First, three projections (i.e., A, B, and C) are generated with three subspaces, respectively. The dimensionality reduction results are automatically grouped after applying the cluster algorithm. The correlation of members within scatterplots are changed across different subspaces (such as $\{a_1, b_1\}$, $\{a_2, b_2, c_2\}$, and $\{a_3, b_3\}$). Second, the inner members' group labels of subsets are changed, forming the co-evolution of the subsets across subspaces, and $sum(c_i \cap c_j)$ denotes the number of items that subset c_i and subset c_j share. Let $c_{(i,j)}$ be the distribution of c_i across c_j, which also presents the similarity of two subsets because the higher the ratio of the same elements in two subsets is, the greater the value will be.

$$c_{(i,j)} = 1 - sum(c_i \cap c_j)/sum(c_i). \tag{1}$$

Third, to obtain all subsets' distribution, $M(P_i, P_j)$ is defined as the changes between projection P_i and P_j. This matrix summarizes the convergence and divergence behavior changes based on the distribution of subsets within one projection P_i (n + 1 clusters are identified) on another projection P_j (m + 1 clusters are identified). The sum of each row equals to 1. In addition, the first row of this matrix represents the transition of the first subset in P_i across each of the m subsets in P_j.

$$M(P_i, P_j) = \begin{bmatrix} c_{(0,0)} & c_{(0,1)} & \cdots & c_{(0,m)} \\ c_{(1,0)} & c_{(1,1)} & \cdots & c_{(1,m)} \\ \vdots & \vdots & \ddots & \vdots \\ c_{(n,0)} & c_{(n,1)} & \cdots & c_{(n,m)} \end{bmatrix}. \tag{2}$$

Finally, to quantify this transition of dimensionality reduction results, we further obtain the statistics of matrix $M(P_i, P_j)$ to express the changes between two projections better as follows.

$$D(P_i, P_j) = \frac{1}{n}\sum_{k=0}^{n} \sqrt{\frac{\sum_{p=0}^{m} |c_{(k,p)} - \overline{c_{(k)}}|}{m-1}}, m > 1, \tag{3}$$

where $c_{(k,p)}$ is in the *k-th* row and *p-th* colum of $M(P_i, P_j)$ when $\overline{c_{(k)}}$ is the average of all values of the *k-th* row of $M(P_i, P_j)$. A low $D(P_i, P_j)$ value indicates that the behavior tends to be divergent from convergent when the subsets within projection P_i are dispersed into other subsets within projection P_j, and vice versa.

Furthermore, evolution matrix Evo is generated after all projection changes are computed. Evo indicates the co-evolution of dimensionality reduction results across various subspaces.

$$Evo = \begin{bmatrix} D_{(P_0,P_0)} & D_{(P_0,P_1)} & \cdots & D_{(P_0,P_n)} \\ D_{(P_1,P_0)} & D_{(P_1,P_1)} & \cdots & D_{(P_1,P_n)} \\ \vdots & \vdots & \ddots & \vdots \\ D_{(P_m,P_0)} & D_{(P_m,P_1)} & \cdots & D_{(P_m,P_n)} \end{bmatrix}. \tag{4}$$

4.2 Projections Subtraction

Selecting interesting subspaces and carefully observing the subtle differences among each cluster are usually time-consuming and inefficient when we must focus on the detailed changes. In this work, we defined the following three kinds of layout constraints to assist users in finding key information from subtle transitions locally and efficiently (T5).

- *Distance Constraint.* When switching subspaces, the dimensionality reduction results will cause the data point location changes. We define points that beyond the distance threshold as distinctive points. As shown in Fig. 6, the change of distance is out of the threshold d) from point a within projection 1 to $a^{'}$ within projection 2. Thus, we consider this data record as a potential special point. In addition, deciding on threshold d is challenging. Thus, we provide users dynamic d by interactively deciding the parameter of the quantile statistical approach. We also leverage the common parameters as default (quartile) to automatically detect the threshold and visually highlight the outliers, which help users in efficiently identifying the abnormalities.
- *Cluster Constraint.* The dimensionality reduction results may result in changes in the points' group labels with the subspace changes. We define these points as distinguished data points. As shown in Fig. 6, the group label of point b changed from point b of the blue subset within projection 1 to point $b^{'}$ of the green subset within projection 2. In addition, we use the clustering algorithm to detect groups and then automatically identify the point whose group label changes by matching the group labels. Finally, our system visually highlights these special points.

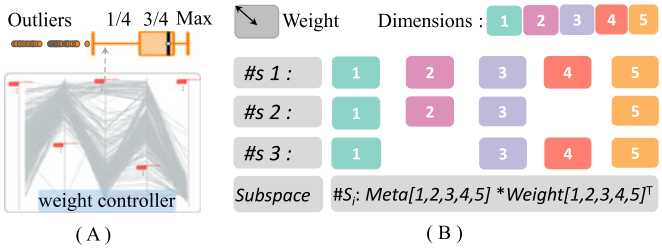


Fig. 7. Dimensions visualization with small multiple boxplots for illustrating the characteristics of dimensions (A). Subspaces visualization with combined weighted-color blocks, in which each color block represents an individual dimension and the radius of color blocks encode the weight of dimensions (B).

- *Angle Constraint*. The dimensionality reduction results will cause the offset of data points with the transforming subspace. We define the points beyond the angle threshold as notable points. As shown in Fig. 6, the change in angle is out of the c threshold θ) from a point within projection 1 to $c^{'}$ within projection 2. Thus, we consider this data entry as a potential special point. In addition, deciding the value of θ is challenging. We utilize the quantile statistical approach to providing a dynamic θ to automatically identify and visually highlight point outliers.

Based on above considerations, we provide a superposition design [15] to present a detailed comparison of two projections encoded by distinguished colors shown in Fig. 1C. All special points are highlighted and connected with lines to present the changes (distance outliers with pink lines, angle outliers with blue lines, the group changings with green lines, and hybrid with red lines), while other points are visualized with low saturation color and act as the context.

5 Visualization Design

This section describes the details of each visual component, visual mapping, and supported interactions. Fig. 1 shows the user interface that has four major components: the dimensions statistic panel (Fig. 1A) for showing the characteristics of each dimension, the correlation panel(Fig. 1B) for displaying an overview of the co-evolution among dimensionality reduction results, the projection subtraction panel (Fig. 1C) for illustrating the subtle changes in projections, and the evo-subsets view (Fig. 1D) for presenting the subsets' divergence or convergence behavior changes across subspaces.

5.1 Subspace Visualization

Dimensions Characteristics. The dimensions of a raw dataset usually suffer from a lack of interpretability, especially for users without prior knowledge. Thus, the characteristics of dimensions must be presented to provide users a preliminary understanding of dimensions and guidance in combining dimensions. In Fig. 7A, the dimensions are encoded by the color of small multiple boxplots to visualize the outliers after general interquartile statistical analysis. In addition, a parallel coordinate plot is employed to illustrate the overall trend of the raw data.

Dimensions Combination. A combination of various dimensions forms a subspace. Except for the scenario where one subspace contains one dimension, the number of possible subspaces is 2^n-1-n, where n is the number of attributes. If each dimension is further equipped with various weights, the number of subspaces tends to infinity, presenting a major challenge in data exploration. Existing works present automatic approaches to reducing the effort of exploring potential and reasonable subspaces. However, problems remain, such as missing significant attributes that are of interest to users. Thus, our system supports users in creating subspaces automatically and/or interactively(T1).

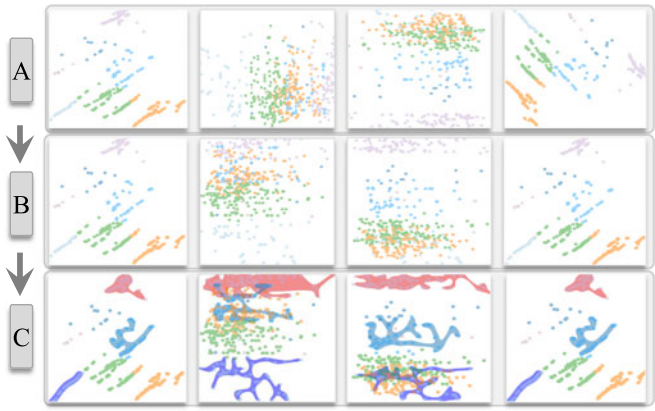


Fig. 8. Original dimensionality reduction results across various subspaces (A). Projections with geometric transformation (B). Subsets visualization after applying the extended *Bubble Sets* visualization (C).

First, we employ SURFING algorithm [3] and also provide the function of automatically generating subspaces to minimize the burden caused by subspace exploration considering many subspaces, while being able to explore all the attributes. Second, users can interactively adjust the weight (w) of each dimension, which implies that users could put emphasis on dimensions during analysis. If the weight (w) of a dimension is set to zero, then this dimension does not play any role in dimensionality reduction results because the dissimilarities of all data records are zero under this dimension. As shown in Fig. 7B, each color block acts as a visual metaphor of a dimension, the radius of these color pixels encodes the weight of dimensions, and the subspaces are encoded with the combination of the weighted-color blocks. Therefore, the amount of change in the dimensionality reduction results could directly indicate the influence of the added, removed, or replaced dimensions on the dimensionality reduction result. On the basis of certain selected subspaces, the high-dimensional dataset is projected into low-dimensional space, which is commonly visualized as scatterplots.

Subsets Visualization. Our work is designed for comparing the changes between dimensionality reduction results, which involves subsets visualization. First, to preserve the relative correlation among data instances, we apply the common dimensionality reduction technique [6] to project a multi-dimensional dataset as a scatterplot. We utilize the Birch algorithm to automatically identify groups. As shown in Fig. 1D, each scatterplot represents a selection of specific dimensions, and the color encodes the clustering category of each point. Second, the visualization must maintaining the viewer's mental map and distinguishing various subsets. Many studies on maintaining the users' mental map have been conducted [9], [14], [16], [37]. In Fig. 8B, we

leverage the Procrustes transformation [14], [16] to solve flipping and arbitrary rotation issues, Third, We employ extended *Bubble Sets* visualization. Along with the switching of subspaces, the distribution of subsets in the scatterplots indicates the subsets' convergence and divergence behavior changes. Furthermore, as the subspaces increase, identifying the convergence and divergence behavior among the small multiples of scatterplots quickly is another challenge. To address these issues, we extend *Bubble Sets*-based algorithm to relax the visual clutter, strengthen the correlations of members in subsets, and highlight subsets' behavior changes. As shown in Fig. 8C, each cluster is encircled by a bubble to highlight the corresponding points' locations, and same clusters in various projections are circled by the same color bubble.

Subspace Correlation. We must visually summarize how subspaces affect dimensionality reduction results (T2), as interpreted by Equation (4). Our goal is to track the co-evolution of dimensionality reduction results across subspaces, and Equation (4) expresses the directed correlation information. Thus, we leverage a circle-based matrix diagram to illustrate the differences on dimensionality reduction results, as shown in Fig. 1B. The radius of one blue circle encodes the changes of two dimensionality reduction results. The radius of one red circle represents the stability of dissimilarity in one row of dimensionality reduction results. In this paper, the degree of dissimilarity is obtained by measuring the variance of the repeatability among subsets. A large radius indicates that the two dimensionality reduction results are more different across subspaces and vice versa. In addition, the sequence of the left and top color blocks in Fig. 1B represents the selected dimensions. The histogram at the right part of Fig. 1B shows DSC and DC to help users comprehensively understand the quantity of dimensionality reduction results.

5.2 Extended Bubble Sets Visualization

Subsets visualization aims to quickly distinguish subsets and determine the distribution of subsets. In this section, we describe our extended *Bubble Sets* visualization (T4), which is designed to relax visual clutter by enlarging the high-density area or decreasing the edge-crossing area.

Reallocate Intersecting Regions. To relax the visual clutter in the proposed visualization, researchers usually aim to minimize the intersection area through various optimization techniques. However, in some cases, such as the overlapping of various intertwined sets as shown by the pink bubble in Fig. 9a, minimizing the intersection area may increase the number of interactions among all the sets. Furthermore, coupled with the repeating overlay of the semitransparent colors that represent different subsets, the overall visualization tends to have more visual clutter. Thus, we extend the original *Bubble Sets* visualization [10] to satisfy our requirement. The main idea of the original *Bubble Sets* algorithm is to introduce an importance value, which is computed by the "energy" function for each grid cell, in the corresponding sets. After calculating the potential energy values ($E(p)$) for all pixels, virtual lines are applied to connect the elements to ensure that all set members are connected and contained and the items that are not in the set would be excluded.

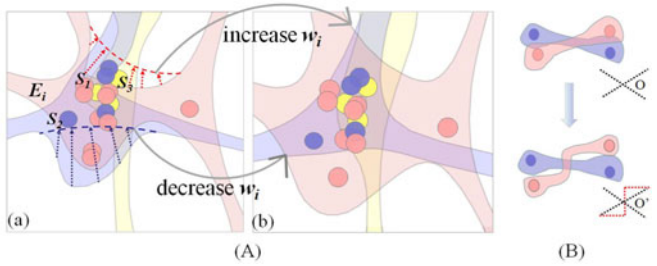


Fig. 9. Extended *Bubble Sets* visualization. The enlarged area of overlap may relax visual clutter (A). Initial visualization of subets (a), results after applying the extended *Bubble Sets* (b). Changing the crossing angle could improve legibility (B).

The idea of our extended algorithm is to increase the overlapping high-density area to relax the visual clutter. Fig. 9A illustrates the process of optimizing the overlapping area. First, we execute the clustering algorithm to identify the areas where severe intersections may occur. If the number of sets in the identified area is larger than or equal to threshold t and the number of points in one set is much larger than those in other sets, then our algorithm would optimize this area (denoted as E_i, which contains sets $S_1, S_2, \ldots, S_k$) by automatically determining the energy function as follows. Second, considering that enlarging the area may mislead users of the number of subsets, our equation must ensure that the relative area of subsets visualization is consistent with the number of points within the subsets. We increase the area of some part of subset S_{max} in certain area E_i to relax the visual clutter to some extent.

$$E(p) = \sum_{i \in S_{pixel}} w_i (R_1 - distance_{i,pixel})^2 / (R_1 - R_0)^2 \tag{5}$$

$$w_i = \begin{cases} \frac{N(S_k)}{\sum_{k \in E_i} N(S_k)} & \forall_i \notin S_{max}\ ,\ i \in E_i\ ,\ S_{max} \subset E_i \\ 1 + \ln \frac{\sum_{k \in E_i} N(S_k)}{N(S_k)} & \forall_i \in S_{max}\ ,\ i \in E_i\ \ S_{max} \subset E_i\ , \\ 1 & i \notin E_i \end{cases} \tag{6}$$

where $distance_{i,pixel}$ represents the euclidean distance to the nearest point on the shape surface, and the values are zero if a pixel is within the corresponding shape. $N(S_k)$ is the number of points in set k, and S_{max} denotes the biggest set in region E_i. R_0 is the distance at which the energy is 1, and R_1 is the distance at which the energy reaches 0. S_{pixel} is the set of influencing points within R_1 of the pixel. Algorithm 1 summarizes the process of optimizing the overlapping area to relax the visual clutter.

Alleviate Edge Crossings Impact. In the original *Bubble Sets* algorithm, virtual edges are created to ensure continuity and accuracy among the subsets. Thus, edge-crossings are often unavoidable even if the detour of the virtual edges successfully diminishes the crossings among bubble subsets.

Another idea of our extended algorithm is to decrease the area of edge-crossing to relax the visual clutter. Fig. 9B illustrates the process of optimizing the crossing area. First, to ensure that the crossings area of the edges is as small as possible, the intersection among edges should form large angles

(e.g., the edges that intersect vertically in Fig. 9B). This could help minimize the area of overlap between the two edges and help enhance the visual continuity of each set itself. We take the intersection of two virtual edges as the center point and add two virtual points on both sides of the virtual edge and then adjust the energy contribution of the two points to be equal to one of the virtual edges.

Algorithm 1. Algorithm of Extended *Bubble Sets*

Require: P is a node set and $S_1, \ldots, S_k$ are subsets of P
Ensure: An extended Bubble Sets visualization for $S_1, \ldots, S_k$
1: Perform the DBScan algorithm, and obtain the top hot area $E(E_1, E_2, \ldots, E_i)$
2: **for** each $E_i \in E$ **do**
3: Count the number of points for each subsets $N(S_k)$ in E_i
4: Determine the largest subsets S_{max} in E
5: Add nodes that are both in S_{max} and E_i to M, add remaining nodes to N
6: **for** each $item \in M$ **do**
7: Increase the energy of nodes by increasing the w_i of item
8: **end for**
9: **for** each $item \in N$ **do**
10: Decrease the energy of nodes by decreasing the w_i of item
11: **end for**
12: **end for**
13: **return** Visualization results $V_1, \ldots, V_k$ for $S_1, \ldots, S_k$

5.3 Visual Metaphor of Behavior Changes

The behavior changes of subsets reflect the correlation changes of the members within subsets. The subsets visualization based on the extended *Bubble Sets* visualization could not only help distinguish subsets with different colors, but also provide an intuitive perception of the subsets' convergence or divergence behaviors. Therefore, the visual encoding of the subsets' behavior changes (T3) could be categorized into the following four kinds.

Divergence to Divergence (B1). Multiple subsets are distributed to various subsets across certain subspaces. The behavior of this subset tends to be static, indicating that the added or removed attributes do not increase the similarity of the subsets. The replaced attributes do not strengthen the correlation of these confused members. The remaining attributes share information. Thus, the changed attributes within subspaces play a non-dominant role in the dimensionality reduction results. Otherwise, the coupling phenomenon may exist among the changed attributes.

Divergence to Convergence (B2). Multiple subsets are merged into a single subset across another subspace. The behavior of this subset tends to be dynamic, indicating that the added or removed attributes reinforce the similarity of the subsets. The replaced attributes strengthen the members of multiple subsets to be one group, and the remaining attributes share information. Thus, the dimensionality reduction results are sensitive to the changed attributes, which play a non-dominant role in the dimensionality reduction results.

Convergence to Divergence (B3). One subset is divided into multiple subsets across certain subspaces. The behavior of this subset tends to be dynamic, indicating that the added or removed attributes decrease the dissimilarity of subsets. The replaced attributes strengthen the members of one group to be multiple subsets, and the remaining attributes share information. Thus, the changed attributes may play a domain role in the dimensionality reduction results.

Convergence to Convergence (B4). One subset is still grouped across certain subspaces. The behavior of this subset tends to be static, indicating that the added or removed attributes decrease the similarity of the subsets. The replaced attributes strengthen the correlation of the members of one group, and the remaining attributes share information. Thus, the changed attributes may play a non-dominant role in the dimensionality reduction results. Otherwise, the coupling phenomenon may exist among the changed attributes.

The subsets' behavior tends to be static in B1 and B4, which means that the elements within the subset remains similar across two subspaces. Therefore, the subspaces have no or few influences on the dimensionality reduction results. The subsets in B2 and B3 changed, indicating that the correlation of elements within the subsets changed across two subspaces. Therefore, the subspaces have a significant influence on the dimensionality reduction results. Given that a subspace is composed of a series of dimensions, we can infer that dimension d is the domain dimension. We summarize the characteristics of domain dimensions d as follows:

- *Scenario 1.* If the subspace does not contain dimension d, then the generated dimensional reduction result tends to be mutable.
- *Scenario 2.* If the subspace contains dimension d, then the generated dimensional reduction result tends to be stable.

5.4 User Interactions

EvoSets supports various basic and advanced interactions, including multiple view linking and brushing, to address analytical tasks.

Show Overall Projections First, Followed by Detailed Distribution on Demand. EvoSets presents a visual summary of all the projections generated by different subspaces and their difference. Users can examine the overall projections to locate ones of interest (T2) and perform subsequent exploration tasks, such as projection subtraction. In addition, our system also supports various projections techniques, such as PCA and MDS, and various clustering algorithms, such as K-Means, DBScan, and Birch (T3). Users could freely explore the detailed information based on the above options.

Guidance-Based Subspace Selection. A feature guidance view (T1) could not only offer a summarization of all dimensions, but also allows users to combine subspace of interest. Fig. 1 A shows an example of guidance-based subspace selection. Users could click on dimension buttons to determine which one to select, and the corresponding

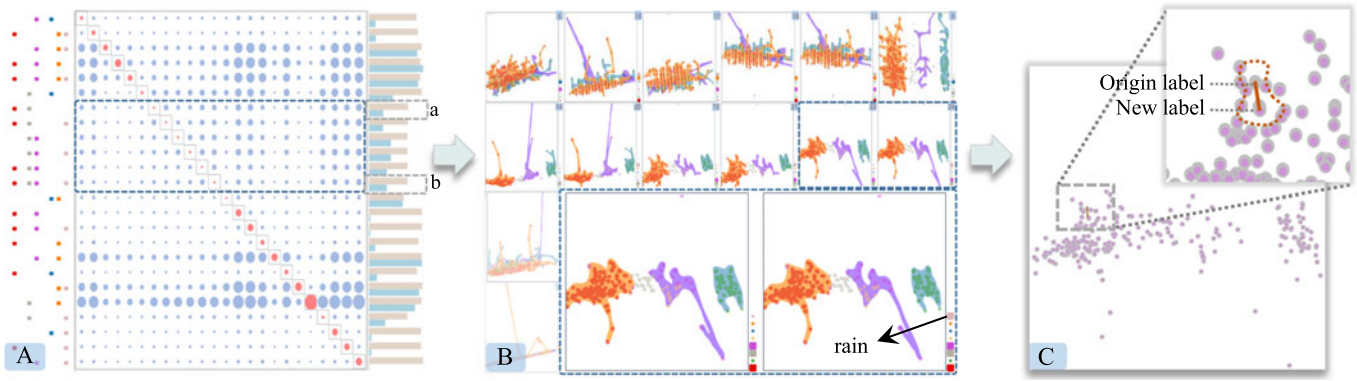


Fig. 10. Correlation of twenty-four subspaces generated in *Pattern Trails* after applying the SURFING algorithm (A). Co-evolution of subsets' behavior changes of convergence or divergence (B). Detailed differences identification of two projections with the layout constraints, where the special transition of points is highlighted with bright lines (C).

projection appears along with the selection to provide timely feedback.

Abnormal Items Identification. A user may want to compare two specific views and examine the difference between them in detail to identify abnormal items (T5). Our system could highlight items based on distance, angle, and clustering constraints through abnormality threshold control using a sliding bar.

Tracking Specific Subsets. We utilize the lasso technique on the scatterplots under subspaces. Users can draw an arbitrary shape on one scatter chart with the mouse. Then, the extended *Bubble Sets* visualization is applied to the subset to show the behavior changes of the subsets in different subspaces.

Reordering Multiple Scatterplots. To help users follow the behavior changes of the subsets when tracking their co-evolution across different subspaces, the system supports reordering projections by dragging projections, which places similar projections near each other for users to easily compare.

6 Case Study

We first employ *EvoSets* to a well-known Forest Fires dataset and discuss our findings in comparison with those using *Pattern Trails* [22]. Then, we apply our system to the vehicle dataset and employ our analysis workflow to find dominant dimensions or subspaces.

6.1 Patterns Finding

The first dataset is a common Forest Fire dataset, which comprises of seven numerical attributes: *Fine Fuel Moisture Code* (*FFMC*), *Duff Moisture Code* (*DMC*), *Drought Code* (*DRC*), *Initial Spread Index* (*ISI*), *Temperature* (*temp*), *Relative Humidity* (*RH*), *Wind*, and *Rain*.

Analysis of Forest Fires Dataset. Our goal is to find the domain attributes on the dimensionality reduction results. First, We qualitatively analyze each attribute of the dataset by specifically combining attributes. The first subspace is composed of all the attributes, and then one attribute is removed iteratively. These combinations of subspaces can be analyzed to reveal the contribution of individual attributes to the dimensionality reduction results.

Second, we obtain multiple dimensionality reduction results, in which we can draw any enclosed shape in a certain dimensionality reduction result to track the behavior changes of subsets in other dimensionality reduction results. As shown in the left part of Fig. 11, we find that the state of the subset in the three figures with black arrows is different from the other projections. For example, the state of the subset wrapped in red bubbles has three divergence and six convergence patterns. In addition, the red and blue subsets are close, indicating that they are more similar than the other dimensionality reduction results. The matrix diagram on the right shows that the dimensions of *DRC* and *RH* are removed in these three figures with black arrows. Therefore, *DRC* and *RH* may play a key role in dimensionality reduction.

Comparison With Pattern Trails. We investigate the changes of dimensionality reduction results across subspaces globally and locally. To demonstrate that our approach performs better than *Pattern Trails* in exploring the differences. We leverage 24 subspaces in *Pattern Trails* to obtain two dozens of dimensionality reduction results.

Globally, the bar chart in Fig. 10A, shows that the metrics (DSC and DC) of each group are similar from points *a* to *b* in the gray box. The corresponding radius of the circle in blue box is small, indicating that these subspaces have a great effect on the dimensionality reduction results in which the changes are not quantified, and users must manually compare the changes among subspaces in *Pattern Trails*. In addition, shown in Fig. 10B shows that compared with the behavior of the subsets within projections in the first row, the behavior of the three subsets (i.e., golden, blue, and purple) are all convergent within the dimensionality reduction results in the second row. Locally, several subspace-pairs are similar, such as the dimensionality reduction results in the blue box in Fig. 10B. To further explore the detailed differences, we click the corresponding dimensionality reduction results. In Fig. 10C, we detect one special data record with a red line whose location offset changed the most. By clicking this point, the parallel coordinate chart shows that it is an outlier in dimension *Rain* in Fig. 1. The comparison of the two subspaces shows that only the results of attribute *Rain* changed in the notable location offset.

Overall, we identify that the attribute with obvious groups may be the dominant attribute after exploring the

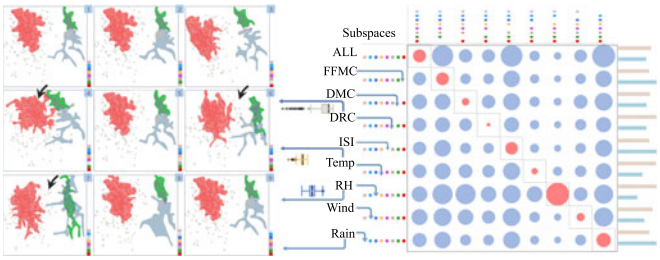


Fig. 11. Sample analysis of the Forest Fires dataset. The left part is the results of dimensionality reduction results with distinguished subspaces. The right part is the change degree of dimensionality reduction results with different subspaces.

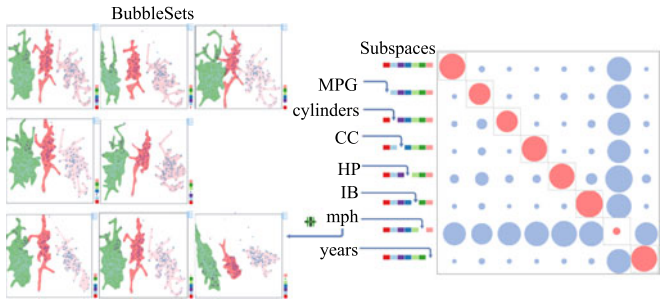


Fig. 13. A sample analysis of the vehicle dataset. The subsets' behavior changes in dimensionality reduction results with distinct subspaces (A). The correlation of dimensionality reduction results with subspaces (B).

distribution of attributes, such as the attribute of *DRC* whose distribution shows more than two clusters. In addition, compared with *Pattern Trails*, our approach could not only quickly observe the correlations among subspaces and track the behavior changes of subsets across subspaces, but also places emphasis on the detailed transition among the dimensionality reduction results.

6.2 Dimensions Perception

We conduct our second experiment on another classical dataset, that is, the vehicle dataset, to validate our method and visual analysis workflow. The data comprises seven dimensions: *economy* (*MPG*), *cylinders*, *displacement* (*CC*), *power* (*HP*), *weight* (*IB*), *mph*, and *year*.

The first subspace is combined with all seven dimensions, and the rest of the subspaces would each remove one different dimension. Fig. 13 shows eight projections, which display the behavior changes of three subsets filled with red, blue, and pink. We then compare the co-evolution of the subsets from the scatterplot. The structural changes of the green area are not obvious, while those of the blue and pink bubble subsets display great degrees of changes. In the dimensionality reduction results, if we remove the *mph* dimension in the subspace, the state of the subsets for the subspaces appears convergent, and the other subspaces exhibit divergence. However, we find identifying the differences when two projection are similar tedious. We explore the changes of the dimensionality reduction results from the overall correlations view. Furthermore, we find the column and row with the largest radius of the circle in the right chart of Fig. 13, which shows that the results of this subspace have a different change. The subspace, which lacks the *mph* dimension, has a different blue circle radius in the corresponding column and row. Therefore, we locate the dominant *mph* dimension with dimensionality reduction results.

For further ranking the domain dimensions of the remaining ones, we repeat the above steps, explore the rest of the dimensions, and obtain seven dimensionality reduction results. We find that the *MPG* dimension is the next dominant dimension. Similarly, the relations chart indicates that *HP* is the next dominant dimension because it has a large circle in the corresponding row and column, indicating its great structure change in dimensionality reduction results. In summary, our visualization technique can sort the dominant dimensions, and the result is *mph–MPG–HP* for the vehicle dataset.

For exploring special points, we click the blue circle of the matrix chart to explore the difference in projections. For example, in the comparison of the two projections in Figs. 12A and 12B, the projection subtraction view helps us

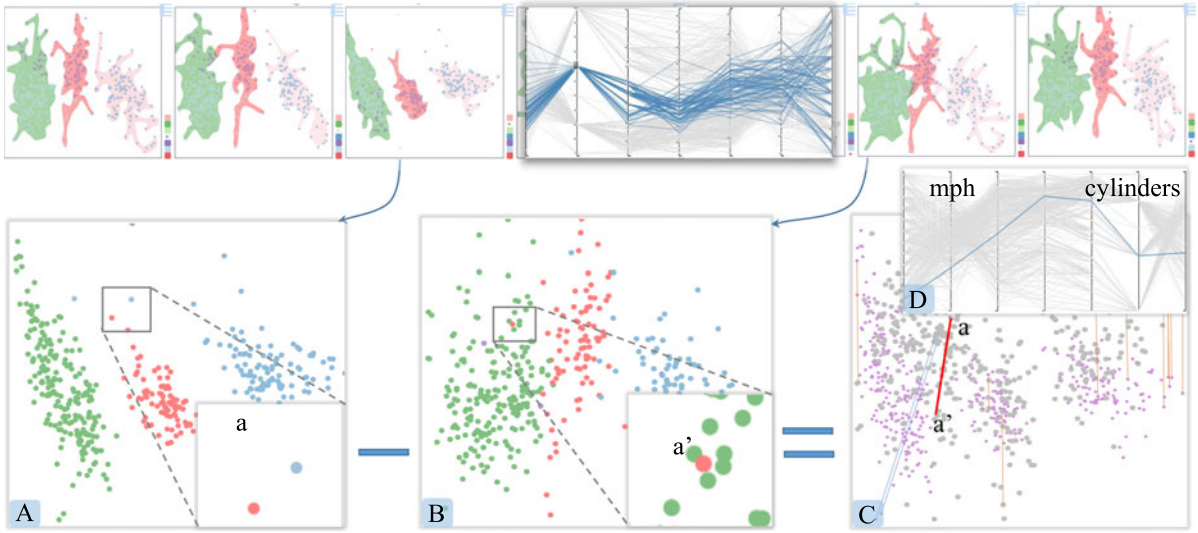


Fig. 12. Detailed dissimilarity comparison. The behavior changes of three subsets in various subspaces of a vehicle dataset (A). The distribution of red subsets in the raw data (B). The position change of point *a* is detected with layout constraints (C).

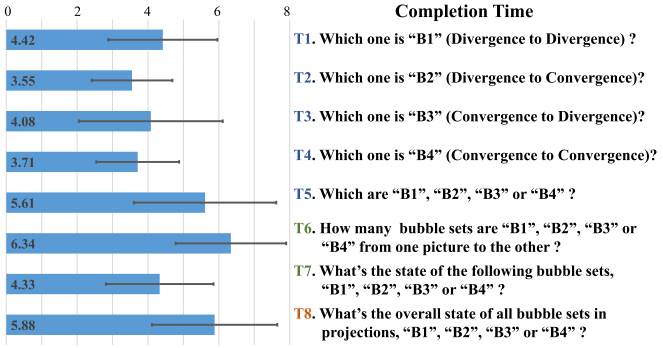


Fig. 14. Evaluation tasks and analysis of time results. Means (in seconds) and standard deviation of completion time are presented.

efficiently locate some special points. Point *a* is at the edge of the red subsets within the former subspace and is then completely mixed with the green subsets within the latter subspace. The former subspace consists of *MPG*, *cylinders*, *CC*, *HP*, *weight*, *IB*, and *year*, and the latter subspace consists of *MPG*, *cylinders*, *CC*, *HP*, *IB*, *mph*, and *year*. Like the previous conclusion, the *mph* is a dominant dimension. Therefore, when the domain dimension is added to the subspace, the cluster label of the points changes into another cluster (*a'*), and the distribution of *a* in the original data is shown in Fig. 12D.

6.3 User Study

We invited 14 participants, who are major in computer science, to evaluate the effectiveness of the visual encoding of subsets' behavior changes. User-feedback are also summarized from 12 users, who have two years of data visualization experience, to evaluate the effectiveness of EvoSets. We first described the visual encoding and user interactions in EvoSets, then demonstrated the patterns that were exposed in case studies. We summarized the user-feedback from three aspects:

Visual Design. Our proposed user interface(Fig. 1) has four major components, including Figs. 1A, 1B, and 1C. All participants can grasp these visual expressions after the description of the system and the training for each visualization technique. Fig. 1D is our proposed new implementation of a visual metaphor of subsets' behavior changes. To evaluate the effectiveness of the visual encoding, 14 participants were asked to complete a series of tasks. Fig. 14 summarizes completion time results of identifying the behavior changes of two pair-wise subsets. The result shows that most tasks are completed in about five seconds. The recognization of B1 tends to take more time than B2, B3, and B4. The standard deviation of T5 (identifying four kinds of subsets' behavior changes simultaneously) tend to be larger, which reveals that the completion times are significantly different. The completion time of T6 and T8 is slower than other tasks, and we speculate that multiple subsets in T6 and T8 result in visual interference. We also found that the accuracy results are all above 85 percent. When identifying the overall state of all bubble sets in projections, users give incorrect results. They all particularly point out subsets' behavior visualization that gives them an intuitive perception of tracking subsets' convergence or divergence behavior changes.

Usability. All users appreciated our system and confirmed the usefulness and effectiveness of subspaces comparison in EvoSets. they all agreed that the system is efficient and effective not only for globally exploring how subspaces affect dimensionality reduction results but also for locally identifying detailed changes between dimensionality reduction results. They all pointed out that the automatic subspaces creation is time-consuming for interactive subspaces generation after understanding the characteristics of all dimensions. One of them commented that "The system allows me to analyze any kinds of multi-dimensional datasets."

Improvement. The participants provided valuable suggestions on how to strengthen the system. Although the system received positive feedback during the interview from the users, they still raised concerns about scalability (e.g., increasing the number of dimensions to thousands). Another valuable comment from one user is that all subspaces could be presented rather than automatically or interactively selecting some subspaces. Regarding the improvement of the visual encoding, it can be still challenging for users to remember visual encoding, especially for the visual encoding of attributes with increasing dimensions. Two of the participants had difficulty in relating the visual encoding containing subspaces encoding and subsets encoding. Thus all of the users suggest us to further design and simplify the interface.

7 Discussion

Subspaces analysis is one of the common topics and widely used to analyze dimensionality reduction results in various domains. Numerous works have been carried out based on projections. Understanding the meanings and obtaining the interpretation of dimensions on projections could assist analysts in finding implicit patterns and proper dimensionality reduction results for further analysis. As with all empirical works, our work has limitations.

Visual Consistency. We compute and visualize the changes between dimensionality reduction results based on subsets' distribution, and provide the visual metaphor on subsets' behavior changes of convergence or divergence. However, the visual representation still suffer from the deviation of dimensionality reduction results. Although we have relaxed this issue by subsets-based computation, dimensionality reduction results may be significantly distinguished while their visual representation of behavior changes may be fully similar. Therefore, a more effective visualization approach should be designed to display the co-evolution of all dimensionality reduction results and the subsets.

Scalability Issue. The enormous amount of dimensions and subsets makes multiple scatterplots difficult to track and comprehend for both users, raising scalability issue. The possible solutions with respect to our scenario to this issue are as follows. From the perspective of raw data, the hierarchical clustering technique can be leveraged for handling multi-dimensional datasets [19]. From the perspective of automatic algorithm, algorithms [44] to automatically decrease the number of subspace explorations can be applied. From the perspective of manual approach, users are guided in their interactive exploration by sorting

subspaces of interest and using metrics, such as DSC and DC, or other hybrid methods. Here, we leverage the attributes statistics method to guide users in removing, replacing, or adding attributes. In addition, metrics of projection are used to guide users in deleting or retaining any projection they want for further exploration.

Steep Learning Curve. The system provides users with a rich set of features: generating subspaces, comparing projection, and tracking subsets' behavior. Users can explore how subspaces affect dimensionality reduction results. However, remembering series of visual encoding and repeatedly comparing projections across subspaces comes with a steep learning curve. Thus, utilizing the system could still be challenging for ordinary users who are not familiar with dimensionality reduction. More intuitive visualization techniques are desired to encode the overall or subtle changes in dimensionality reduction results across subspaces.

In future work, we would conduct experiments on higher dimensional data and use hierarchical dimensions to relax scalability issues. We would attempt to illustrate all possible combinations of dimensions and present their characteristics and correlations to improve the efficiency of user exploration. In addition, subset memberships are likely to change across subspaces, thus, visual clutter poses challenges with respect to the extended *Bubble Sets*. The "energy" is determined by the density of subsets and the number of elements in the extended *Bubble Sets* algorithm. The "energy" of subsets would be similar when the number of elements of subsets is similar in messing region. Thus, the result of reallocating intersecting regions above may not be ideal. We would also search for a new visual incoding or algorithms to relax visual clutter in tracking the co-evolution of subsets.

8 Conclusion

This paper introduces EvoSets, a visual analysis system for multi-level exploration of how dimensionality reduction results change across various subspaces from two perspectives, globally and locally. For parsing global information, the similarity of the dimensionality reduction results is revealed based on the distribution of subsets. Furthermore, for visually tracking the subsets' behavior changes of convergence or divergence, we extended *Bubble Sets* to visualize how subspaces affect the correlation of subset memberships. For parsing local information, we define three kinds of constraints to filter out special points, considering redundant local detail changes.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1707700, in part by the National Natural Science Foundation of China under Grants 61972356 and 62036009, and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-A2020001.

References

[1] B. Alper, N. H. Riche, G. Ramos, and M. Czerwinski, "Design study of linesets, a novel set visualization technique," *IEEE Trans. Visualization Comput. Graph.*, vol. 17, no. 12, pp. 2259–2267, Dec. 2011.

[2] A. Anand, L. Wilkinson, and T. N. Dang, "Visual pattern discovery using random projections," in *Proc. Visual Anal. Sci. Technol.*, 2013, pp. 43–52.

[3] C. Baumgartner, C. Plant, K. Kailing, H. P. Kriegel, and P. Kröger, "Subspace selection for clustering high-dimensional data," in *Proc. IEEE Int. Conf. Data Mining*, 2004, pp. 11–18.

[4] J. Bernard, N. Wilhelm, M. Scherer, T. May, and T. Schreck, "Timeseriespaths: Projection-based explorative analysis of multivarate time series data," *J. WSCG*, vol. 20, no. 2, pp. 97–106, 2012.

[5] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Trans. Visualization Comput. Graph.*, vol. 17, no. 12, pp. 2203–2212, Dec. 2011.

[6] I. Borg and P. J. Groenen, "Modern multidimensional scaling: Theory and applications," *J. Educ. Meas.*, vol. 40, no. 3, pp. 277–280, 2006.

[7] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, "Flowstrates: An approach for visual exploration of temporal origin-destination data," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 971–980, 2011.

[8] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck, "Assisted descriptor selection based on visual comparative data analysis," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 891–900, 2011.

[9] G. D. Cantareira 1 and F. V. Paulovich 2, "A generic model for projection alignment applied to neural network visualization," in *Proc. Eurovis Workshop Visual Anal.*, 2020, pp. 67–71.

[10] C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Trans. Visualization Comput. Graph.*, vol. 15, no. 6, pp. 1009–1016, Nov.–Dec. 2009.

[11] M. Daszykowski and B. Walczak, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Comprehensive Chemometrics*, 1996, pp. 635–654.

[12] A. Erasmo and R. Minghim, "A novel visual approach for enhanced attribute analysis and selection," *Comput. Graph.*, vol. 84, no. 11, pp. 160–172, 2019.

[13] M. Espadoto, R. M. Martins, A. Kerren, N. S. Hirata, and A. C. Telea, "Towards a quantitative survey of dimension reduction techniques," *IEEE Trans. Visualization Comput. Graph.*, vol. 27, no. 3, pp. 1–1, Mar. 2021.

[14] T. Fujiwara, J.-K. Chou, Shilpika, P. Xu, L. Ren, and K.-L. Ma, "An incremental dimensionality reduction method for visualizing streaming multidimensional data," *IEEE Trans. Visualization Comput. Graph.*, vol. 26, no. 1, pp. 418–428, Jan. 2020.

[15] M. Gleicher, "Considerations for visualizing comparison," *IEEE Trans. Visualization Comput. Graph.*, vol. 24, no. 1, pp. 413–423, Jan. 2018.

[16] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford. U.K.: Oxford Univ. Press, 2004.

[17] D. Guo, "Coordinating computational and visual approaches for interactive feature selection and multivariate clustering," *Inf. Visualization*, vol. 2, no. 4, pp. 232–246, 2003.

[18] H.-P. Kriegel, P. Kröger, and Arthur Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1–58, 2009.

[19] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 297–304.

[20] Y. Hu, S. Wu, S. Xia, J. Fu, and W. Chen, "Motion track: Visualizing variations of human motion data," in *Proc. IEEE Pacific Visualization Symp.*, 2010, pp. 153–160.

[21] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475,1985.

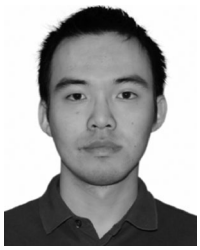
[22] D. Jackle, M. Hund, M. Behrisch, D. A. Keim, and T. Schreck, "Pattern Trails: Visual analysis of pattern transitions in subspaces," in *Proc. IEEE Conf. Visual Anal. Sci. Technol.*, 2017, pp. 1–12.

[23] K. Kailing, H.-P. Kriegel, P. Kröger, and S. Wanka, "Ranking interesting subspaces for clustering high dimensional data," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discov.*, 2003, pp. 241–252.

[24] J. Krause, A. Dasgupta, J. D. Fekete, and E. Bertini, "SeekAView: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces," in *Proc. IEEE Symp. Large Data Anal. Visualization*, 2016, pp. 11–19.

[25] J. Lee and M. Verleysen, "Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods," in *Proc. New Challenges Feature Selection Data Mining Knowl. Discov.*, 2008, pp. 21–35.

[26] D. J. Lehmann and H. Theisel, "Optimal sets of projections of high-dimensional data," *IEEE Trans. Visualization Comput. Graph.*, vol. 22, no. 1, pp. 609–618, Jan. 2016.
[27] S. Liu, B. Wang, J. J. Thiagarajan, P.-T. Bremer, and V. Pascucci, "Visual exploration of high-dimensional data through subspace analysis and dynamic projections," *Comput. Graph. Forum*, vol. 34, no. 3, pp. 271–280, 2015.
[28] Y. Ma, A. K. Tung, W. Wang, X. Gao, Z. Pan, and W. Chen, "ScatterNet: A deep subjective similarity model for visual analysis of scatterplots," *IEEE Trans. Visualization Comput. Graph.*, vol. 26, no. 3, pp. 1562–1576, Mar. 2020.
[29] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE Laurens," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[30] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer, "Guiding feature subset selection with an interactive visualization," in *Proc. IEEE Conf. Visual Anal. Sci. Technol.*, 2011, pp. 111–120.
[31] W. Meulemans, N. H. Riche, B. Speckmann, B. Alper, and T. Dwyer, "KelpFusion: A hybrid set visualization technique," *IEEE Trans. Visualization Comput. Graph.*, vol. 19, no. 11, pp. 1846–1858, Nov. 2013.
[32] L. Pagliosa, P. Pagliosa, and L. G. Nonato, "Understanding attribute variability in multidimensional projections," in *Proc. SIBGRAPI Conf. Graph., Patterns Images*, 2016, pp. 297–304.
[33] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data," *Comput. Graph. Forum*, vol. 31, no. 3, pp. 1345–1354, 2012.
[34] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM Sigkdd Explorations Newslett.*, vol. 6, no. 1, pp. 90–105, 2004.
[35] D. Pérez, L. Zhang, M. Schaefer, T. Schreck, D. Keim, and I. Díaz, "Interactive feature space extension for multidimensional data projection," *Neurocomputing*, vol. 150, no. 2, pp. 611–626, 2015.
[36] J. Poco *et al.*, "A framework for exploring multidimensional data with 3D projections," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 1111–1120, Jun. 2011.
[37] P. E. Rauber, A. X. Falcão, and A. C. Telea, "Visualizing time-dependent data using dynamic t-SNE," in *Proc. Eurograph. Conf. Visualization: Short Papers*, 2016, p. 73—77.
[38] C. Reinbold, A. Kumpf, and R. Westermann, "Visualizing the stability of 2D point sets from dimensionality reduction techniques," *Comput. Graph. Forum*, vol. 39, no. 1, pp. 333–346, 2020.
[39] D. Sacha *et al.*, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Trans. Visualization Comput. Graph.*, vol. 23, no. 1, pp. 241–250, Jan. 2017.
[40] M. Schaefer *et al.*, "Improving projection-based data analysis by feature space transformations," in *Proc. Visualization Data Anal.*, 2013, pp. 196–210.
[41] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan, "Selecting good views of high-dimensional data using class consistency," *Comput. Graph. Forum*, vol. 28, no. 3, pp. 831–838, 2009.
[42] J. Stahnke, M. Dörk, B. Müller, and A. Thom, "Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions," *IEEE Trans. Visualization Comput. Graph.*, vol. 22, no. 1, pp. 629–638, 2015.
[43] A. Tatu *et al.*, "Automated analytical methods to support visual exploration of high-dimensional data," *IEEE Trans. Visualization Comput. Graph.*, vol. 17, no. 5, pp. 584–597, May 2011.
[44] A. Tatu *et al.*, "Subspace search and visualization to make sense of alternative," in *Proc. IEEE Conf. Visual Anal. Sci. Technol.*, 2012, pp. 63–72.
[45] J. Wang, X. Liu, and H.-W. Shen, "High-dimensional data analysis with subspace comparison using matrix visualization," *Inf. Visualization*, vol. 18, no. 1, pp. 94–109, 2019.
[46] S. Wang and H. Sun, "Measuring overlap-rate for cluster merging in a hierarchical approach to color image segmentation," *Int. J. Fuzzy Syst.*, vol. 6, no. 3, pp. 150–159, 2004.
[47] M. O. Ward and Z. Guo, "Visual exploration of time-series data with shape space projections," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 701–710, 2011.
[48] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987.
[49] K. Xu, L. Zhang, D. Pérez, P. H. Nguyen, and A. Ogilvie-Smith, "Evaluating interactive visualization of multidimensional data projection with feature transformation," *Multimodal Technol. Interaction*, vol. 1, no. 3, pp. 1–20, 2017.
[50] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Trans. Visualization Comput. Graph.*, vol. 19, no. 12, pp. 2625–2633, Dec. 2013.
[51] F. Zhou, J. Li, W. Huang, Y. Zhao, and Y. Shi, "Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data," in *Proc. IEEE Pacific Visualization Symp.*, 2016, pp. 128–135.



Guodao Sun received the BSc degree in computer science and technology from the Zhejiang University of Technology, and the PhD degree in control science and engineering from the Zhejiang University of Technology. He is an associate professor with the College of Computer Science and Technology, Zhejiang University of Technology, HangZhou, China. His main research interests include visual analysis with focus in spatio-temporal, text, and multi-media data.

Sujia Zhu is currently working toward the MS degree with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. Her major research interests include data mining, mult-dimensional data visualization, and visual analytics of multivariate data.

Qi Jiang is currently working toward the MS degree with the College of Control Science and Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include machine learning, information visualization, and visual analytics.

Wang Xia is currently working toward the MS degree with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. Her research interests include computer vision, information visualization, and visual analytics.

Ronghua Liang received the BSc degree in computer science from Hangzhou Dianzi University in 1996 and the PhD degree in computer science from Zhejiang University in 2003. From April 2004 to July 2005, he was a research fellow with the University of Bedfordshire, U.K. and from March 2010 to March 2011, a visiting scholar with the University of California. He is currently a professor of computer science, and the vice dean with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou. His research interests include computer vision, information visualization, and visual analytics.