



Visual subspace clustering based on dimension relevance



Jiazhi Xia^a, Guang Jiang^a, YuHong Zhang^a, Rui Li^a, Wei Chen^{b,*}

^aSchool of Information Science and Engineering, Central South University, China

^bState Key Laboratory of CAD&CG, Zhejiang University, China

ARTICLE INFO

Article history:

Received 17 September 2016

Revised 3 May 2017

Accepted 20 May 2017

Available online 27 May 2017

Keywords:

High dimensional data

Subspace clustering

Interactive visual analysis

Dimension relevance

ABSTRACT

The proposed work aims at visual subspace clustering and addresses two challenges: an efficient visual subspace clustering workflow and an intuitive visual description of subspace structure. Handling the first challenge is to escape the circular dependency between detecting meaningful subspaces and discovering clusters. We propose a dimension relevance measure to indicate the cluster significance in the corresponding subspace. The dynamic dimension relevance guides the subspace exploring in our visual analysis system. To address the second challenge, we propose hyper-graph and the visualization of it to describe the structure of subspaces. Dimension overlapping between subspaces and data overlapping between clusters are clearly shown with our visual design. Experimental results demonstrate that our approach is intuitive, efficient, and robust in visual subspace clustering.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

High-dimensional (HD) data is usually constructed by a mixture of clusters in different subspaces [1]. Subspace clustering [2], which aims at detecting all the meaningful clusters and subspaces, is essential for HD data analysis. Challenges exist in both subspace analysis algorithm and interactive visualization. On the subspace analysis algorithm side, traditional clustering algorithms can hardly handle subspace clustering task. The major problem is that the full-space Euclidean distance can hardly reveal the intrinsic structure. Clusters exist in appropriate subspaces only and irrelevant dimensions can obscure the clustering patterns. For interactive visualization, the main task is to visually reveal the subspace structure. Usually, subspaces overlap with each other in dimensions, and clusters in different subspaces have overlapping data points [2]. How to describe the underlying structure and construct a desired visual mapping is still an open problem.

Detecting clusters and discovering meaningful subspaces are two inter-dependent tasks [1]. To determine a meaningful subspace, at least some cluster members are identified. On the other hand, to identify the cluster members, the appropriate subspace must be known. To escape from this circular dependency, automatic subspace analysis algorithms either begin with assumption of cluster members in a top-down manner (e.g. PROCLUS [3]) or start by subspace searching in a bottom-up manner (e.g.

CLIQUE [4]). While automatic algorithms often generate redundant and uninterpretable results, subspace visualization is proposed for visually understanding the structure of subspaces [5]. For axis-parallel subspaces, dual space interface is demonstrated to be successful in visual subspace analysis [6–8]. Users can explore the subspaces in the dimension view and investigate dataset in the data view. However, visual subspace clustering faces the same problem to escape from the circular dependency. Although iterative exploring with dual space can result in meaningful discovery sooner or later, it yields heavy trial-and-error overhead in the analysis process.

In this paper, we propose a visual subspace clustering approach with the dual space design. The essential idea is to break the circular dependency by suggesting the relevant dimensions in the subspace exploring process. We follow the line of bottom-up approaches [1] that find dense regions in low dimensional spaces and combine them to form clusters. We define the dimension relevance as the significance of cluster pattern in the subspace expanded by the corresponding dimensions. To be clear, we use the term relevance to distinguish from the term correlation that refers to statistical dependence relationship between variables. We first measure the relevance in each 2-dimensional subspaces. The corresponding layout of dimension view shows the relevance and guides the user to select the 2-dimensional subspace with high relevance. Consequently, the dimensionality of selected subspace increases in a bottom-up manner until no more dimensions can be added into current subspace and preserves the dense pattern. The data view is also updated with the subspace exploration to reveal the cluster pattern in the selected subspace.

* Corresponding author.

E-mail address: chenwei@cad.zju.edu.cn (W. Chen).

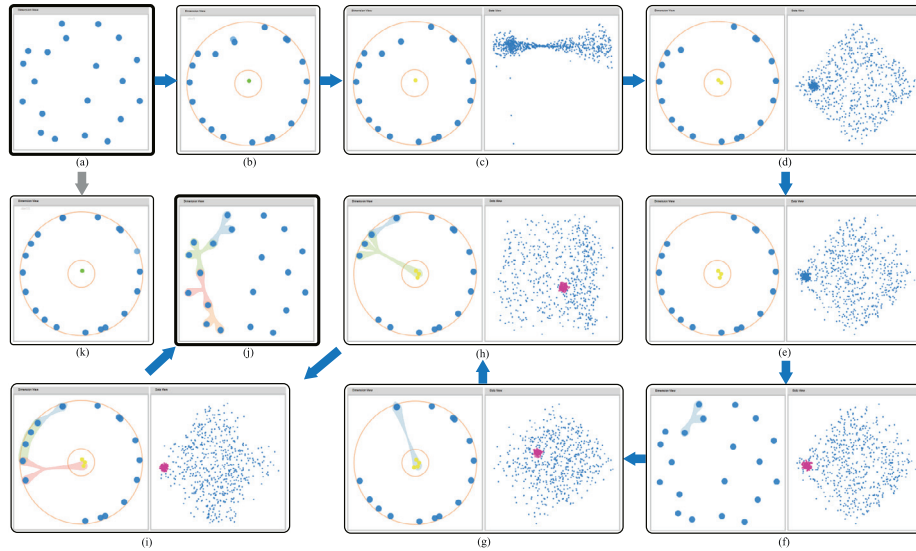


Fig. 1. Visual analysis workflow of the synthetic 20-D dataset. (a) The initial MDS layout in the dimension view; (b) Dynamic radial layout when hovering on dimension d_1 ; (c) The dimension view and data view when selecting d_1 ; (d) The dimension view and data view after adding d_3 ; (e) The dimension view and data view after adding d_5 ; (f) Coloring the dense pattern and save the subspace S_1 ; (g) The dimension view and data view when selecting subspace S_2 ; (h) The dimension view and data view when selecting subspace S_3 ; (i) The dimension view and data view when selecting subspace S_4 ; (j) The dimension view of the exploring result; (k) Dynamic radial layout when hovering on dimension d_{11} .

To summarize, our work has the following contributions:

- A novel dimension relevance measure in the sense of subspace cluster significance;
- A hyper-graph structure and visualization to describe the subspace; and
- A visual analysis approach for interactive subspace clustering.

2. Related work

2.1. Subspace clustering

In the data mining community, the so-called subspace clustering algorithms are proposed to detect all the meaningful clusters and corresponding subspaces [1,2]. Generally, two strategies are adopted to escape from the circular dependency between detecting meaningful subspace and discovering clusters. Algorithms which start by assumptions of cluster members are categorized into the top-down type. PROCLUS [3] starts by a set of random potential cluster centers. The neighborhood of each center and the corresponding subspace are updated iteratively. PreDe-Con [9] derives the so-called subspace preference of each point from its ε -neighborhood. To detect arbitrary oriented subspaces, Principal Component Analysis (PCA) is often applied to a local selection of points. PROCLUS [3], 4C [10] and ERic [11] fall into this category. The bottom-up approaches are based on the downward closure property: if a subspace S contains a cluster, then any subspace $T \subseteq S$ must also contains a cluster [12]. CLIQUE [4] start the searching of meaningful subspaces from 1-dimensional subspaces. $k+1$ -dimensional subspaces are checked while the meaningful k -dimensional subspaces are known. Subspaces without dense cluster are pruned in the subspace searching process. SUBCLU [13] applies DBSCAN [14] to subspaces following a similar strategy.

Our approach implicitly follows the philosophy of bottom-up strategy. The significance of cluster patterns in different possible subspaces are visually suggested when exploring the subspaces. Different from the automatic methods, we do not define a concrete model of cluster. Arbitrary-shaped clusters can be detected in the visual analysis process.

2.2. Visual subspace analysis

The automatic subspace clustering methods often generate results which are highly redundant and hard to understand and interpret. Visual analysis provides a mean to address this problem [15]. There is a strong need to visualize subspace clustering results to understand, evaluate, and refine them. Assent et al. [16] visualized the similarity between clusters in terms of the size of clusters and dimensions. Müller et al. developed the Morpheus system [17] to support interactive exploration of subspace clustering results. TripAdvisor [18] provided a navigation framework to explore the subspace clustering results. Tatu et al. [5] developed a visualization and navigation interface to explore the large sets of subspaces found by SURFING [19]. Ferdosi et al. [20] detected relevant subspaces based on the density estimation and morphological operators and proposed visualization toolkits to explore and refine the results.

A few studies are dedicated in interactive subspace analysis. Turkey et al. [6,7] introduced a dual visual analysis framework which favors simultaneously exploring the data space and dimension space. Similarly, Yuan et al. [8] developed a hierarchical approach to iteratively divide the data items and dimensions into subgroups. This paper aims at an interactive visual subspace clustering. Different from existing methods, our approach suggests the dimension relevance in subspace exploring process to reduce the trial-and-error overhead.

2.3. HD data visualization and dimension reduction

HD data visualization has attracted much attention in both data mining and visualization communities. Liu et al. [21] proposed a detailed review on HD data visualization. Conventional techniques, including scatterplot matrix [22], parallel coordinates [23], and Radviz [24], focus on visualize the data points in full dimensions. The uncertainties in HD data visualization also attracts a lot of research interests [25]. When the dimensionality is high, the call for rendering space and the human perception make the visualization to be challenging.

The general solution is to apply dimension reduction (e.g. PCA [26] and classical MDS [27]) or projection [28,29] before vi-

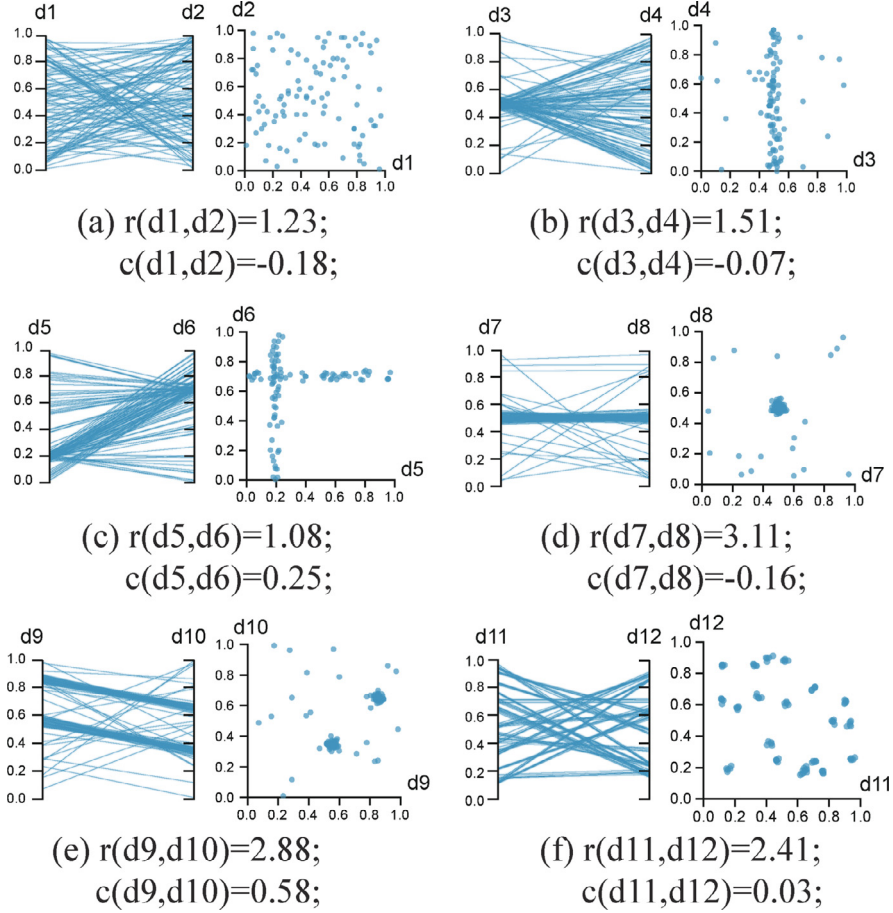


Fig. 2. The relevance r and correlation c between two dimensions. Left: the parallel coordinates of data. Right: the scattered plot of data.

sualization. However, the global dimension reduction do not solve the subspace clustering problem [5]. The clusters may be located in different subspaces. Other than dimensions correlation which is often used in global dimension reduction, we proposed dimension relevance which is defined in the sense of clustering.

3. Dimension relevance

The essential idea of this paper is to measure the relationship among dimensions in the sense of indicating the significance of cluster pattern. The dimension correlation, which refers to statistical dependence relationship between variables, can help to reduce dimensionality. However, as global dimension reduction can not solve the subspace clustering problem [5], dimension correlation is not the choice to indicate the clustering significance [8].

Our solution is inspired by the bottom-up subspace clustering algorithms, such as CLIQUE [4]. Those algorithms model clusters as continuous grids which contain enough data points. They seek the dense grids in a dimensionally increasing order. This idea follows the downward closure property [12]: if a subspace S contains a cluster, then any subspace $T \subseteq S$ must also contains a cluster. In another word, if there is no cluster in a subspace T , then any subspace $S \supseteq T$ can not having cluster.

In this paper, our dimension relevance measure is based on the local density of points rather than the dense grids. There are mainly two reasons for this choice. First, the algorithm is simpler to compute a per-point attribute. Second, without explicitly modeling the cluster, it is more flexible to describe dense pattern with various shapes and non-uniform densities. The local density of points obeys the downward closure property too. A point is lo-

cally dense in a subspace means that it is locally dense in all the participating dimensions. Thus, we define the local density in 1-dimensional space first.

Given a dataset DB with size $|DB|$, we define the local density of point p in dimension d_i as the density of its k -Nearest-Neighbors (k -NN) in dimension d_i . We denote its k -NN as $kNN(p, d_i)$. The local density $\rho(p, d_i)$ of a point p in dimension d_i is defined as

$$\rho(p, d_i) = \frac{k-1}{\max(\max(kNN(p, d_i)) - \min(kNN(p, d_i)), \epsilon)}, \quad (1)$$

where ϵ is a small enough number to prevent invalid division. In this paper, we set $\epsilon = 1/|DB|$. $\max(kNN(p, d_i))$ is the maximum value in d_i of $kNN(p, d_i)$, and similarly, $\min(kNN(p, d_i))$ is the minimum value in d_i of $kNN(p, d_i)$.

Following the downward closure property, we define the local density of a point in a subspace as the minimum local density of it in all the participating dimensions. If there is cluster in a subspace, there must be enough number of locally dense points. Thus, we define the dimension relevance $r(d_i, d_j)$ between d_i and d_j as

$$r(d_i, d_j) = \frac{\sum_{p \in DB} \min(\rho(p, d_i), \rho(p, d_j))}{|DB|^2}, \quad (2)$$

For an uniform random distribution with $|DB|$ points in $[0, 1]$, the expectation of local density of a point is $|DB|$. Thus, we normalize $r(d_i, d_j)$ by $|DB|^2$.

Fig. 2 shows six possible cases of pairwise relevance. In Fig. 2(a), the data points follow uniformly random distribution in both dimensions. There is no cluster pattern in d_1 - d_2 space. In Fig. 2(b), there is dense pattern in dimension d_3 only and no cluster pattern presents in d_3 - d_4 space. In Fig. 2(c), there are dense

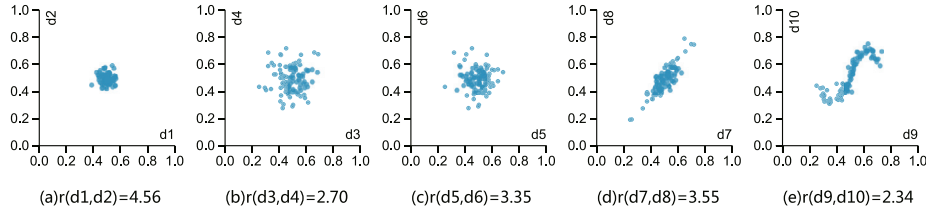


Fig. 3. The relevances of different shape of clusters.

pattern in both dimensions. However, there are hardly any points which is locally dense in both dimensions. As a result, there is still not cluster pattern in d_5 - d_6 space. Fig. 2(d) and (e) show that cluster patterns exist in the 2-dimensional subspace when there are enough common locally dense points in all participating dimensions. An extreme case is shown in Fig. 2(f). All the points are locally dense in both dimensions. However, they are not continuous. As a result, we can hardly find conventional cluster pattern in the d_{11} - d_{12} space. Although we can define a continuity term in the relevance measure, we choose to keep the measure simple for intuition and flexibility. The visual analysis in the data view can help to further identify the cluster pattern.

Because there is no assumption of the shape of cluster in the definition of dimension relevance, it is flexible to capture clusters with different shapes. Fig. 3 shows the dimension relevance of five 2-D clusters with different shapes and densities. All the five clusters are well captured. The relevance between two dimensions can be easily extended to relevance among multiple dimensions. Given a subset of dimensions $\{d_1, d_2, \dots, d_m\}$, the relevance $r(d_1, d_2, \dots, d_m)$ is defined as

$$r(d_1, d_2, \dots, d_m) = \frac{\sum_{p \in DB} \min(\rho(p, d_1), \rho(p, d_2), \dots, \rho(p, d_m))}{|DB|^2}, \quad (3)$$

We define the relevance-based distance between two dimensions as

$$\text{dist}(d_i, d_j) = e^{-(r(d_i, d_j) - 1)}, \quad (4)$$

As the relevance between two dimensions with uniform random distribution are normalized to around 1, the distance between them is about 1, which is the farthest distance.

Similarly, given a subset of dimensions $S = \{d_1, d_2, \dots, d_m\}$ and a dimension d_{m+1} , the distance between d_{m+1} and S is given by

$$\text{dist}(d_{m+1}, S) = e^{-(r(d_1, d_2, \dots, d_m, d_{m+1}) - 1)}. \quad (5)$$

4. Visual subspace clustering

We aim at visually exploring meaningful subspaces and clusters and providing a deep insight into the structure of subspaces. We model the structure of subspace by hyper-graph with additive attributes in the edges. Based on the underlying structure, we explore the subspaces in the dual space interface which contains dimension space and data space. In the dimension space, we visually present the relevance-based distance between dimensions to guide the subspace exploring. Users explore the subspace in a dimension-wise incremental manner. The clustering pattern can be further explored in the data space. With the visual guidance of dimension relevance, the workflow breaks the circular dependency between detecting meaningful subspace and discovering clusters.

In this section, we showcase the visual subspace clustering with the synthetic 20D dataset. This dataset contains 20 dimensions and 800 points. There are three 3-D clusters and one 4-D clusters. The remaining dimensions are uniform distributed noises.

4.1. Subspace structure

The subspace structure is considered to be complicated even though we focus on axis-parallel subspaces. Overlapping occurs in dimensions and data points. Subspaces can overlap each other in dimensions and clusters can have common data points. Modeling and visualizing the subspace structure is critical in subspace exploring.

We use hyper-graph to represent the subspace structure. Given a data set DB with n dimensions, we build a hyper-graph $G = (V, E)$ to represent the subspace structures, where the nodes V represents the dimensions and E refers to the set of edges. An edge in E is a n -element subset of V , and it indicates a meaningful subspace. In this paper, we use the term subspace edge for it. In one meaningful subspace, a data point can belong to one cluster or no cluster. We set the cluster ID of each point to the corresponding edge as the additive attribute.

4.2. Visual analysis system

We propose an integrated system with a suite of visualization components and interaction tools for visual subspace clustering. The interface is shown in Fig. 4.

4.2.1. The exploring workflow

The exploring process starts in the dimension view (Fig. 1(a)). Users can hover on the dimension points and the layout of dimension points are updated according to the distance to hovered point. In Fig. 1(k), the layout shows that all the remaining dimensions are far from the hovering dimension d_{11} . In Fig. 1(b), two dimensions are close to the hovering dimensions d_1 . Therefore, user can select those dimensions incrementally to build subspaces (Fig. 1(d)–(e)). The cluster pattern in current subspace is shown in the data view (Fig. 1(e)right). If desired cluster pattern is observed, users can save the current subspace as a meaningful subspace. The subspace exploration can be continued by adding dimensions points into or deleting dimensions points from the current editing subset of dimensions. Users can further investigate a meaningful subspace by coloring the cluster members in the data view. The coloring is saved as an additive attribute of the subspace (Fig. 1(f)). Users can also show the coloring of subspace s_i in a different subspace s_j to compare the intrinsic structure of two subspaces (Fig. 5).

4.2.2. Dimension view

The dimension view shows the relevance between dimensions (Fig. 1(a)). Each point in the dimension view represents a dimension. The initial layout of dimension points are given by multidimensional scaling (MDS) based on the distance matrix of dimensions. The relevance-based distances are computed by Eq. (4). Following interactions are provided in this view:

- Hovering on a dimension point or a subspace edge;
- Selecting dimension points or a subspace edge;
- Lasso selecting dimension points.

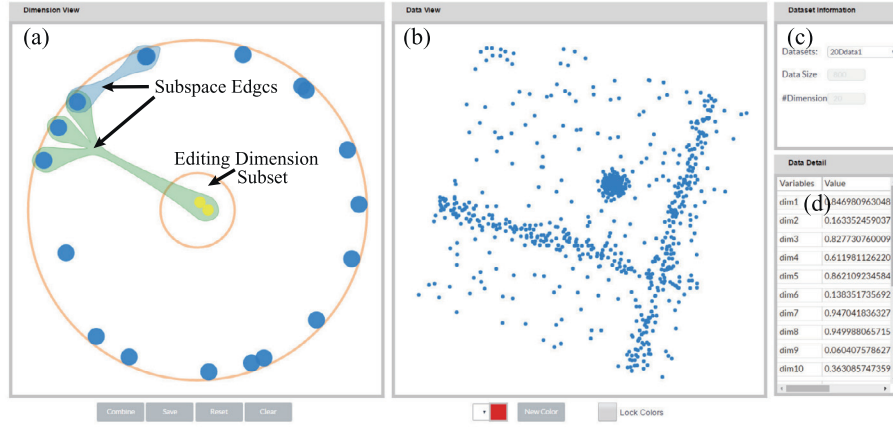


Fig. 4. Interface of the proposed visual analysis system. (a) The dimension view; (b) The data view; (c) The control panel; (d) The data detail panel.

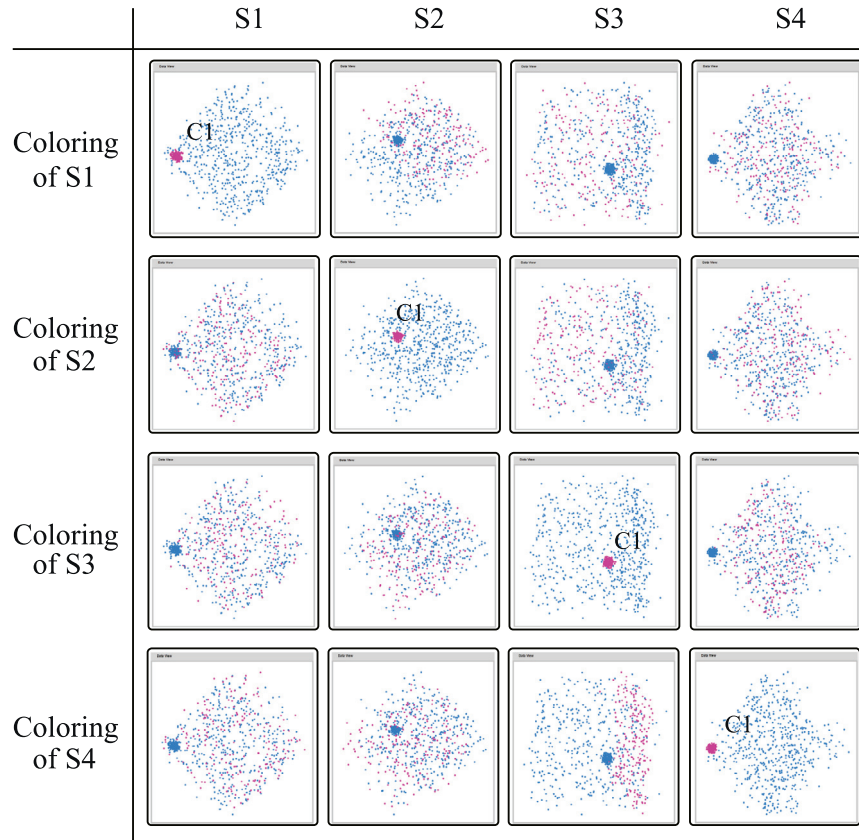


Fig. 5. The comparison of subspace structures of synthetic 2D dataset. Rows: The coloring feature of subspace S_1 , S_2 , S_3 , and S_4 . Columns: The data layout of S_1 , S_2 , S_3 , and S_4 ; The clusters are numbered in the corresponding subspaces only.

A set of operations are provided by the buttons in the bottom of dimension view, including

- Combining dimensions: the editing dimension subset is visually combined into one dimension point;
- Saving subspace: saving the editing dimension subset as a subspace; the clustering information is also saved as the additive attribute of the subspace edge;
- Resetting dimension view: Set the dimension view as the initial MDS layout; and
- Clearing subspaces: clear all the edited subspace information.

The editing dimension subset The selected dimensions and hovered dimension are added into the editing dimension subset.

It is shown in a disk region in the center of the dimension view. The dimension points in the subset are list by the selected order and spirally grows from the center to the outer. The radius of the disk region is set to $0.1 \times w$, where w is the dimension plot width. The width of dimension points is dynamically updated according to the size of list.

Star-shaped subspace edge We design a star-shaped hull to visualize the subspace edge of the hyper-graph. Given n points $p_1, p_2, \dots, p_n$, the computation of the star-shaped hull containing them includes the following steps:

Step1: Detect the minimum covering circle of the points, denote the center point of the circle as c (Fig. 6(a));

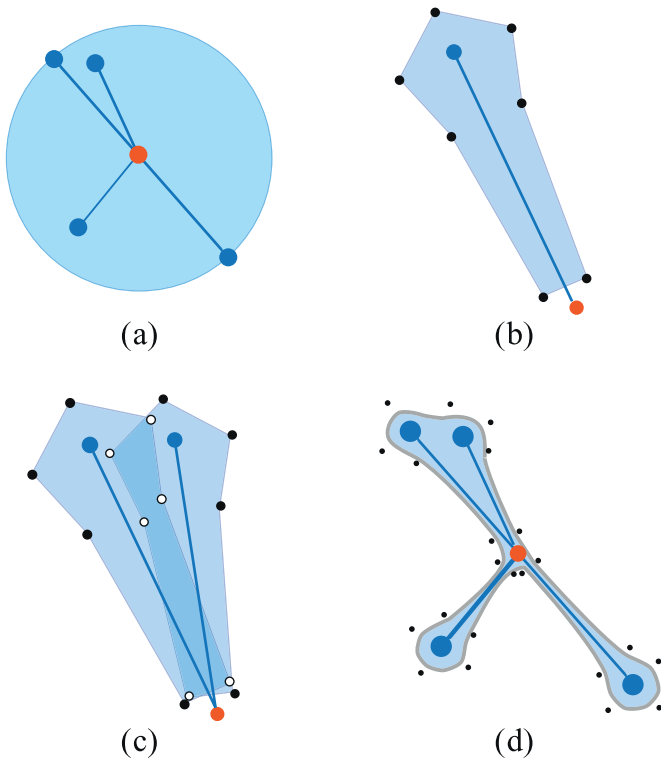


Fig. 6. The star-shaped hull. (a) The minimum covering circle of the points and the center point of the circle (orange). (b) The control points around one branch. (c) Deleting the covered control points. (d) The star-shape hull. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Step2: Connect the n points to c as the skeleton of the star shape;
 Step3: Select 7 points around each skeleton as the control points (Fig. 6(b));
 Step4: For two skeletons which overlap with each other, delete the control points in the overlapped region (Fig. 6(c));
 Step5: Connect the control points in counter-clock-wise direction;
 Step6: Render the closed B-spline with the control points (Fig. 6(d)).

Dynamic layout When there are editing subset of dimensions, the layout of remaining dimensions are updated to indicate the distance to the editing subset. The layout is dynamically updated according to the editing dimension subset. Starting from the initial MDS layout, a remaining dimension point d_i is linearly mapped to the radial line from the center of view to the initial position of d_i . The distance from center is according to $dist(d_i, S)$. Dimension point with $dist = 0$ is mapped to $0.1 \times w$ from the center, which is the edge of center disk. Dimension points with $dist = 1.0$ is mapped to $0.5 \times w$ from the center.

4.2.3. Data view

In the data view, each point represent a data record. Given the editing subspace S in the dimension view, the distance between data records are computed as the Euclidean distance in subspace S . Then, we applied MDS to all the data points using the distance matrix as the input. By preserving the distance in S , the MDS layout reveals the cluster pattern in S . Lasso selection is provided in data view to identify clusters with different colors. The current clusters identifications/colors can be shown in the data view of another subspace by selecting the checkbox “Lock Colors”. This option is useful to compare the intrinsic structure of difference subspaces.

In visual analysis clustering, clusters are considered as dense patterns. Rather than automatic algorithms which always define

the shape, density, and size, we only need to visually and interactively identify the dense patterns which are denser than the distribution of other points. This definition gives more flexibility to describe clusters of various shape, density, and size.

5. Case study

In this section, we evaluate the effectiveness of the proposed system with one synthetic and one real-world datasets.

5.1. Synthetic 12d data

We present a case study on the synthetic 12-D data used in [5,20]. There are 750 data points distributed in four 3-D Gaussian clusters and two 6-D Gaussian clusters. Other dimensions contain uniformly distributed random noise. In this study, we aim to detect the meaningful subspaces and the corresponding clusters.

We start the exploring in the dimension view. Fig. 8(a) shows that there are generally two groups of dimensions. Group #1 contains three dimensions d_2 , d_4 , and d_6 (the name of the dimension is shown when hovering on the dimension point). Group #2 includes six dimensions d_7 , d_8 , d_9 , d_{10} , d_{11} , and d_{12} . The remaining three dimensions d_1 , d_3 , and d_5 are far from each other. The two groups indicate that there are two potentially meaningful subspaces S_1 and S_2 . We further explore the two subspaces in the data view. By selecting dimension subsets, the corresponding clustering pattern is shown in the data view. The data points are visually clustered into four clusters (Fig. 8(f)) and two clusters (Fig. 8(g)) in S_1 and S_2 respectively. Another observation is that the dimensions of S_1 and S_2 are close to each other, although the inter-group distance is not as close as the intra-group distance. Thus, we further investigate the subspace $S_3 = S_1 \cup S_2$ (Fig. 8(d)). Fig. 8(h) shows that there are five dense patterns.

Next, we investigate the structure of the three subspaces in comparative manner. By selecting the checkbox “Lock Colors”, we can show the selected clustering information crossing different subspaces. Fig. 7 shows the three clusterings under S_1 , S_2 , and S_3 respectively. The first row shows that cluster C_2 , C_3 , and C_4 in S_1 belong to the same cluster in S_2 , and C_4 is divided into two clusters in S_3 .

Generally, the exploring of this dataset follows the bottom-up manner. We explore low-dimensional subspaces first. Once clusters in the subspaces are detected, we iteratively add new possible dimensions into current subspace to combine high-dimensional meaningful subspace. We also accelerate this process by selecting a set of dimensions, which are close to each other in the dimension view. As a result, we successfully reveal the subspace clustering structure of the dataset.

5.2. Food data

We performed a case-study on the USDA food composition dataset used in [5,8] for subspace analysis. The dataset is a collection of foods, of which each dimension represents a certain type of nutrient. It contains 722 data points and 18 dimensions. In this study, we would like to explore whether there are clusters and corresponding subspaces in this unknown dataset.

The exploration starts by the initial MDS layout in the dimensions view (Fig. 9(a)). The dimension points shows two dense patterns. The first one contains five dimensions and the second one contains nine dimensions. The density of the first group is higher than the second group. The remaining four dimension points are relatively far from other dimension points.

Intuitively, we explore the subspace S_1 , S_2 expanded by the first two groups respectively. One dense pattern is shown in the

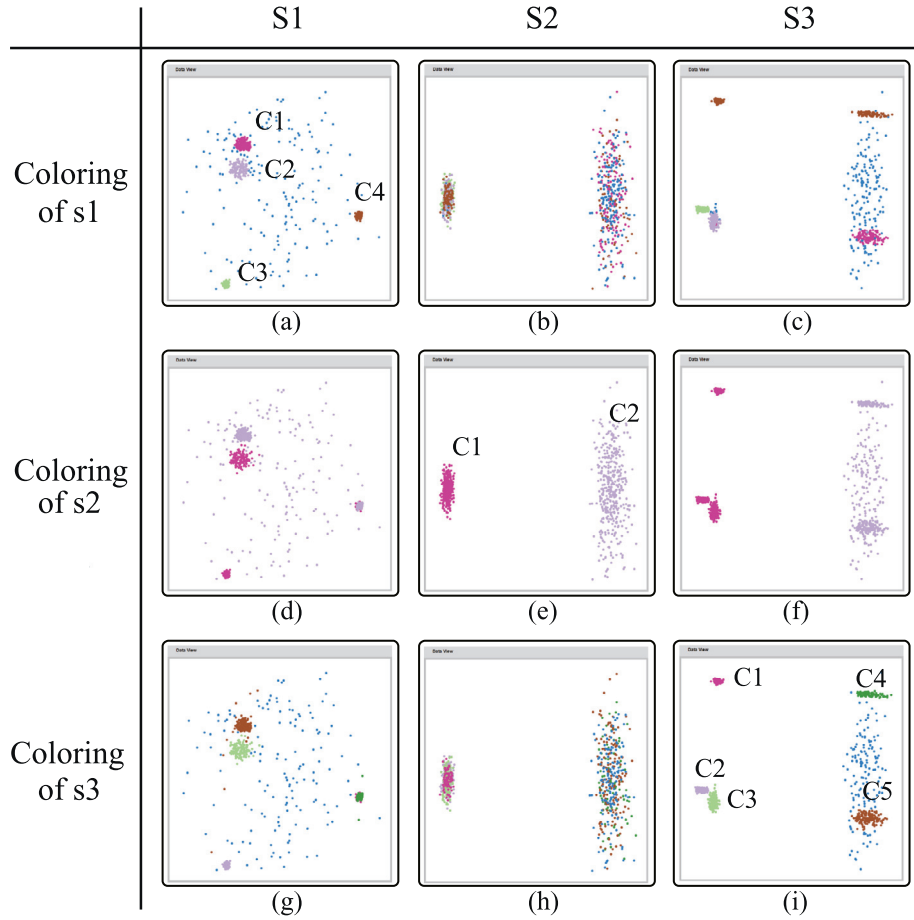


Fig. 7. The comparison of subspace structures of the synthetic 12D dataset. Rows: The coloring feature of subspace S_1 , S_2 , and S_3 . Columns: The data layout of S_1 , S_2 , and S_3 ; The clusters are numbered in the corresponding subspaces only.

data view according to subspace S_1 (Fig. 9(b)). We mark the dense points and save the subspace.

In the data view of S_2 , we can hardly see conventional dense patterns, although there are several locally dense stripes (Fig. 9(c)). We further explore combinations of the dimensions in S_2 and find that the stripe patterns exist in the subspaces containing the dimension Vit_D (Fig. 9(h)). Because all the points are locally dense in Vit_D , Vit_D is close to all dimensions which have locally dense points according to Eq. (4). Thus, we deselect Vit_D from S_2 forming subspace S_3 . One dense pattern is shown in the data view of S_3 (Fig. 9(d)).

Next, we combine the dimensions in S_1 and S_2 as two dimension points respectively. Then, we hover on the dimension in the top left (Fig. 9(e)). The dynamic layout shows that the four dimensions share common locally dense points. We select the four dimensions constructing subspace S_4 (Fig. 9(e)). The data view shows that there are two dense groups.

We also explore the full space S_5 . The dynamic layout in the dimension view shows that there are hardly points which are locally dense in all the dimensions. The data view shows that the data points are generally classified into three classes (Fig. 9(f)), though these three classes are not dense enough.

To further understand the structure of the subspaces, we compare S_5 with S_1 , S_3 , and S_4 (Fig. 10).

In this study, we explore the unknown dataset with the proposed tool, reveal several subspaces which may contain clusters, and compare the structure of these subspaces.

6. Discussion

This section discusses the design philosophy of the proposed visual subspace clustering system. Limitations and possible future work are also discussed.

The design philosophy and workflow Our design philosophy is to provide a visual subspace clustering workflow following the bottom-up strategy. According to the downward closure property [1], the bottom-up searching is guaranteed to be able to find all the meaningful axis-parallel subspaces and clusters. The key of visual subspace clustering is effective branch pruning in the interactive interface.

We proposed dimension relevance to indicate the significance of cluster pattern. This measure is demonstrated to be especially useful to guide the discovering of meaningful subspace. With the guidance, it is intuitive to select close dimension points in the dimension view one by one and form meaningful subspace incrementally. While the irrelevant dimensions are far from the center, they are visually pruned in the exploring process. As a result, users are implicitly guided into a bottom-up workflow, which escapes from the circular dependency between detecting meaningful subspace and detecting clusters.

Dimension relevance vs. correlation In this paper, we use the term relevance to differ from the term correlation which is widely used to describe the statistic dependency relationship between variables. In the term of subspace clustering, neither linear correlation (e.g. Pearson correlation) nor nonlinear correlation can identify the meaningful subspace contains cluster patterns [8]. In con-

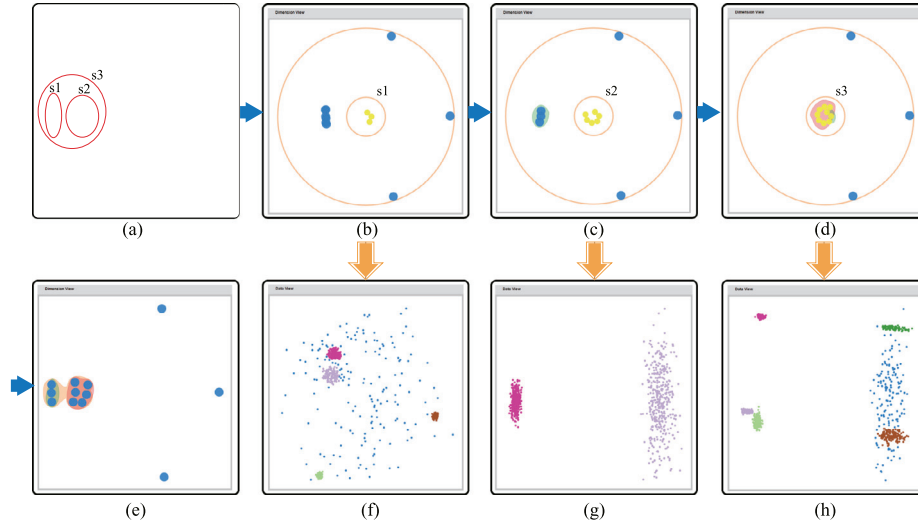


Fig. 8. The workflow of exploring the synthetic 12D dataset. (a) The initial MDS layout in the dimension view; (b) Selecting dimensions $S_1 = \{d_2, d_4, d_6\}$; (c) Selecting dimensions $S_2 = \{d_7, d_8, d_9, d_{10}, d_{11}, d_{12}\}$; (d) Selecting dimensions $S_3 = S_1 \cup S_2$; (e) The subspace edges in the dimension view; (f) The data view of S_1 ; (g) The data view of S_2 ; (h) The data view of S_3 .

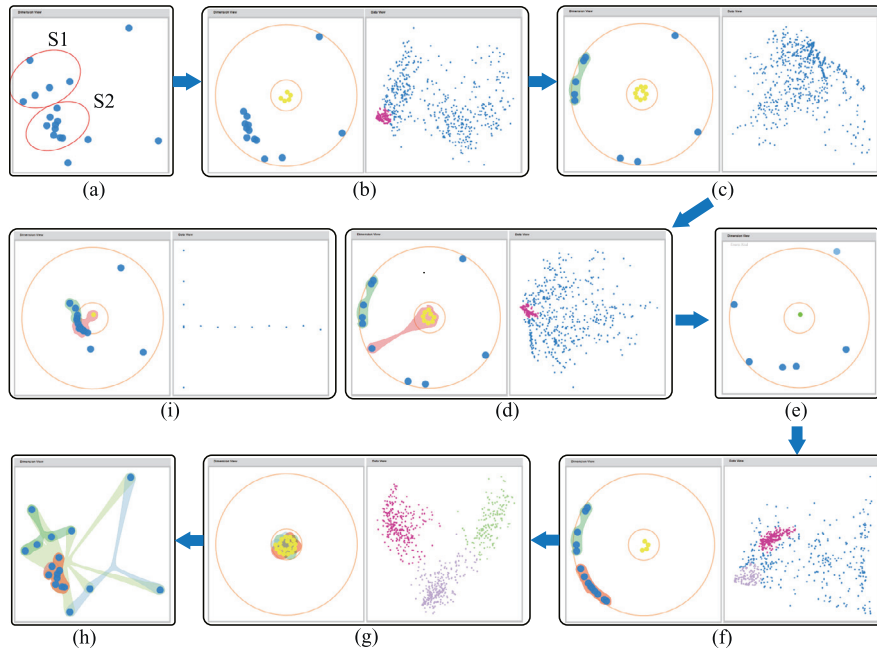


Fig. 9. Exploring the food dataset. (a) The initial MDS layout in the dimension view. (b) Selecting 5 neighboring dimensions forming subspace S_1 ; left: the dimension view; right: the data view.; (c) Selecting 9 neighboring dimensions forming subspace S_2 ; (d) Deleting Vit_D from S_2 resulting S_3 ; (e) Hovering on the dimension in the top left; S_1 and S_3 are combined as two synthetic dimension points; Vit_D is hidden; (f) Selecting the remaining 4 dimensions forming subspace S_4 ; (g) Selecting all dimensions as the full space S_5 ; (h) The dimension view of the exploring result. (i) Selecting dimension Vit_d .

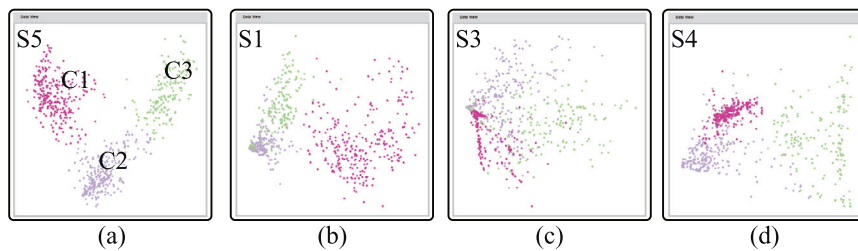


Fig. 10. The comparison of subspace structures of food dataset. (a) The data view of full space S_5 ; (b) The data view of S_1 with the coloring of S_5 ; (c) The data view of S_3 with the coloring of S_5 ; (d) The data view of S_4 with the coloring of S_5 .

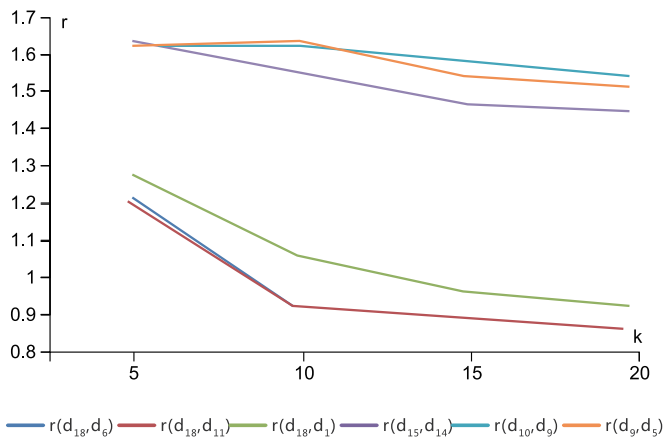


Fig. 11. The relevances of six pairs of dimensions with different k values.

trast, the dimension relevance is designed to reveal the significance of cluster patterns. This measure can be seamlessly integrated into existing subspace analysis systems which rely on the dimension correlation.

When the dimensionality of the dataset is extremely high, the distortion of MDS will be large. The perception burden will also be a challenge to the user. Thus, it is still useful to perform dimension reduction before subspace exploration. A more interesting solution is to consider both correlation and relevance in the visual analysis system. It will enhance the ability of our system to process dataset with extremely high dimensionality. Additionally, it opens new opportunities for correlation analysis. Considering the case that two dimensions are locally correlated on a subset of data, the correlation relationship can be obscured by irrelevant data points. The dimension relevance and subspace clustering suggest a new perspective to seek local correlation.

The k value of k -NN We estimate the local density of a point as the density of k -NN. However, the selection of k is a tradeoff. Small k value can capture fine structure and large k value is able to depict large scale structures. An extremely small k may lead to false positive (Fig. 2(f)). With the 20-D dataset, we test the relevances of a set of dimension pairs with different k values. Fig. 11 shows that the dimension relevance is robust to different k values in a moderate range. In this paper, we choose $k = 9$ to capture the fine structure.

Arbitrary-oriented subspace A common limitation of dual space HD data analysis is that only axis-parallel subspaces can be detected. It seems impossible to detect all the meaningful subspaces in a bottom-up manner due to the numerous directions. A possible solution is to generate new features in which dense patterns are found. Another potential strategy is to integrate the top-down methods with bottom-up approaches. While performing PCA analysis to a subset of data members, new meaningful directions are found and can serve as features in subspace analysis. The difference between the two solutions is whether to generate new feature before subspaces analysis.

7. Conclusion

In this paper, we propose a visual subspace clustering approach for HD dataset. A novel dimension relevance is proposed to identify the significance of cluster pattern. In contrast to dimension correlation, the proposed relevance is cluster-aware and defined on multiple dimensions. We also proposed a hyper-graph structure and the visualization design to describe the structure of subspaces. The dimension overlapping among subspaces and data overlapping

among clusters are clearly shown in our visualization. Case-studies on both real-world and synthetic datasets demonstrated the effectiveness and efficiency of our system.

Acknowledgements

We would like to thank anonymous reviewers for their helpful suggestions and comments. This work is partially supported by National Science Foundation of China(61309009, 61422211), Major Program of National Natural Science Foundation of China (61232012), National 973 Program of China (2015CB352503), Research Fund for the Doctoral Program of Higher Education of China (20130162130001), and Open Project Program of the State Key Lab of CAD&CG (A1710).

References

- [1] H.P. Kriegel, P. Kröger, A. Zimek, Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering, *ACM Trans. Knowl. Discov. Data* 3 (1) (2009) 1:1–1:58.
- [2] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2011) 52–68.
- [3] C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, J.S. Park, Fast algorithms for projected clustering, *SIGMOD Rec.* 28 (2) (1999) 61–72.
- [4] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications, *SIGMOD Rec.* 27 (2) (1998) 94–105.
- [5] A. Tatu, F. Maas, I. Farber, E. Bertini, T. Schreck, T. Seidl, D. Keim, Subspace search and visualization to make sense of alternative clusterings in high-dimensional data, in: *IEEE VAST*, 2012, pp. 63–72.
- [6] C. Turkay, P. Filzmoser, H. Hauser, Brushing dimensions - a dual visual analysis model for high-dimensional data, *IEEE Trans. Vis. Comput. Graph.* 17 (12) (2011) 2591–2599.
- [7] C. Turkay, A. Lundervold, A.J. Lundervold, H. Hauser, Representative factor generation for the interactive visual analysis of high-dimensional data, *IEEE Trans. Vis. Comput. Graph.* 18 (12) (2012) 2621–2630.
- [8] X. Yuan, D. Ren, Z. Wang, C. Guo, Dimension projection matrix/tree: interactive subspace exploration and analysis of high dimensional data, *IEEE Trans. Vis. Comput. Graph.* 19 (12) (2013) 2625–2633.
- [9] C. Böhm, K. Railing, H.P. Kriegel, P. Kröger, Density connected clustering with local subspace preferences, in: *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, 2004a, pp. 27–34.
- [10] C. Böhm, K. Railing, P. Kröger, A. Zimek, Computing clusters of correlation connected objects, in: *SIGMOD '04*, 2004b, pp. 455–466.
- [11] E. Achtert, C. Böhm, H.P. Kriegel, P. Kröger, A. Zimek, On exploring complex relationships of correlation clusters, in: *Scientific and Statistical Database Management, 2007. SSBDM '07. 19th International Conference on*, 2007, pp. 7–7.
- [12] L. Parsons, E. Haque, H. Liu, Subspace clustering for high dimensional data: a review, *SIGKDD Explor. Newsl.* 6 (1) (2004) 90–105.
- [13] H.-P. Kriegel, P. Kröger, A. Zimek, Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering, *ACM Trans. Knowl. Discov. Data* 3 (1) (2009) 1:1–1:58.
- [14] M. Ester, H. Peter Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD 96*, 1996, pp. 226–231.
- [15] G.-D. Sun, Y.-C. Wu, R.-H. Liang, S.-X. Liu, A survey of visual analytics techniques and applications: state-of-the-art research and future challenges, *J. Comput. Sci. Technol.* 28 (5) (2013) 852–867.
- [16] I. Assent, R. Krieger, E. Müller, T. Seidl, Visa: visual subspace clustering analysis, *SIGKDD Explor. Newsl.* 9 (2) (2007) 5–12.
- [17] E. Müller, I. Assent, R. Krieger, T. Jansen, T. Seidl, Morpheus: interactive exploration of subspace clustering, *KDD*, 2008, pp. 1089–1092.
- [18] J.E. Nam, K. Mueller, Tripadvisor^{N-D}: a tourism-inspired high-dimensional space exploration framework with overview and detail, *IEEE Trans. Vis. Comput. Graph.* 19 (2) (2013) 291–305.
- [19] C. Baumgartner, C. Plant, K. Railing, H.P. Kriegel, P. Kröger, Subspace selection for clustering high-dimensional data, in: *ICDM '04*, 2004, pp. 11–18.
- [20] B.J. Ferdosi, H. Buddelmeijer, S. Trager, M.H.F. Wilkinson, J.B.T.M. Roerdink, Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators, in: *IEEE VAST*, 2010, pp. 35–42.
- [21] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, V. Pascucci, Visualizing high-dimensional data: advances in the past decade, in: *Eurographics Conference on Visualization (EuroVis) - STARS*, 2015.
- [22] R.A. Becker, W.S. Cleveland, Brushing scatterplots, *Technometrics* 29 (2) (1987) 127–142.
- [23] A. Inselberg, B. Dimsdale, Parallel coordinates: a tool for visualizing multi-dimensional geometry, in: *IEEE Vis.*, 1990, pp. 361–378.
- [24] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, E. Stanley, Dna visual and analytic data mining, in: *IEEE Vis.*, 1997, pp. 437–441.

- [25] H. Chen, S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, H. Qu, Uncertainty-aware multidimensional ensemble data visualization and exploration, *IEEE Trans. Vis. Comput. Graph.* 21 (9) (2015). 1072–86
- [26] I.T. Jolliffe, *Principal Component Analysis*, Springer, 2002.
- [27] P.C. Wong, R.D. Bergeron, Multivariate visualization using metric scaling, in: *Visualization '97*, Proceedings, 1997, pp. 111–118.
- [28] D.J. Lehmann, H. Theisel, Optimal sets of projections of high-dimensional data, *IEEE Trans. Vis. Comput. Graph.* 22 (1) (2016) 609–618.
- [29] R. Etemadpour, R. Motta, J.G. d. S. Paiva, R. Minghim, M.C.F. de Oliveira, L. Linsen, Perception-based evaluation of projection methods for multidimensional data visualization, *IEEE Trans. Vis. Comput. Graph.* 21 (1) (2015) 81–94.