

# Dimension Reconstruction for Visual Exploration of Subspace Clusters in High-dimensional Data

Fangfang Zhou<sup>1\*</sup> Juncai Li<sup>2†</sup> Wei Huang<sup>2‡</sup> Ying Zhao<sup>1§</sup>  
Xiaoru Yuan<sup>3¶</sup> Xing Liang<sup>4||</sup> Yang Shi<sup>1\*\*</sup>

<sup>1</sup>School of Information Science and Engineering, Central South University

<sup>2</sup>School of Software, Central South University

<sup>3</sup>Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University

<sup>4</sup>School of Computing, Informatics & Decision Systems Engineering, Arizona State University

## ABSTRACT

Subspace-based analysis has increasingly become the preferred method for clustering high-dimensional data. A visually interactive exploration of subspaces and clusters is a cyclic process. Every meaningful discovery will motivate users to re-search subspaces that can provide improved clustering results and reveal the relationships among clusters that can hardly coexist in the original subspaces. However, the combination of dimensions from the original subspaces is not always effective in finding the expected subspaces. In this study, we present an approach that enables users to reconstruct new dimensions from the data projections of subspaces to preserve interesting cluster information. The reconstructed dimensions are included into an analytical workflow with the original dimensions to help users construct target-oriented subspaces which clearly display informative cluster structures. We also provide a visualization tool that assists users in the exploration of subspace clusters by utilizing dimension reconstruction. Several case studies on synthetic and real-world data sets have been performed to prove the effectiveness of our approach. Lastly, further evaluation of the approach has been conducted via expert reviews.

**Keywords:** High-Dimensional Data, Subspace Clustering, Visual Clustering, User Interaction

## 1 INTRODUCTION

In high-dimensional data, meaningful patterns, such as clusters, are more likely to occur only in locally relevant sub-dimensional spaces. Therefore, subspace-based analysis has increasingly become popular in exploration of high-dimensional data. In the data mining community, a class of so-called subspace clustering algorithms has been proposed to automatically find clusters that exist in different dimensional subspaces[11][13]. Simultaneously, in the visualization community, several pioneering works have introduced visual analysis methods to interpret subspace clustering results[14][10], explore subspaces and clusters interactively[16][20], or fulfill specialized requirements in astronomy[4] and bioinformatics[2].

Subspaces, which are constructed from a high-dimensional data

\*e-mail: zff@csu.edu.cn

†e-mail: dreair@csu.edu.cn

‡e-mail: huangwei\_grace@csu.edu.cn

§e-mail: zhaoying@csu.edu.cn (corresponding author)

¶e-mail: xiaoru.yuan@pku.edu.cn

||e-mail: xliang22@asu.edu

\*\*e-mail: shiyangcsu@126.com

set, may be completely different or partially overlapped. Various cluster structures formed in subspaces provide different views on the same data. The visually interactive exploration of subspaces and clusters is a cyclic process. Discoveries in interesting subspaces may lead to other exploration goals and motivate users to start a new round of subspace analysis. One of the potential goals is to construct new subspaces that combine diverse cluster structures observed in different original subspaces. In these new subspaces, users can observe and compare cluster structures directly, and thus, gain additional insights into the data. However, a simple addition of the dimensions from the original subspaces does not always effectively help users quickly find the target subspaces.

For example, food data with nutrient measurements are plotted in two user-selected subspaces: one subspace contains vitamin dimensions (e.g. vitamin B6 and B12) and the other subspace contains mineral dimensions (e.g. calcium and iron). A cluster, named CluVit, is easily recognizable in the vitamin subspace, but is obscure in the mineral subspace. To gain additional insights into the interrelations between CluVit and the clusters uncovered in the mineral subspace, users expect to construct a new subspace where the cluster structure of the mineral subspace is preserved and CluVit is also distinct. In this case, the mechanical merging of the two original subspaces is likely to obscure all clusters because of the multiplied number of dimensions. Finding a few dimensions related to CluVit in the vitamin subspace may be a little complicated and difficult process. Therefore, helping users efficiently construct goal-driven subspaces becomes a challenging problem.

In this study, we present an approach that enables interactive subspace construction driven by the iteratively refined targets of users. Our approach is based on generating a manageable number of reconstructed dimensions (RDs) to preserve interesting cluster information in subspaces. We define a visual analytical pipeline to introduce these new dimensions into the subspace analysis cycle. In each new round of subspace analysis, the RDs can convey the preserved cluster information to the original subspaces to improve the clustering results of these subspaces. Moreover, the RDs can possibly help users effectively construct subspaces that clearly display merged cluster structures that do not coexist in the original subspaces. To incorporate the knowledge of users and achieve their targets, we introduce the 2D data projections of subspaces as the human-machine interface to intuitively explore subspace clusters and interactively reconstruct dimensions. A visualization tool is specifically designed for our pipeline to help users analyse subspace clusters, reconstruct dimensions, and track clusters in different subspaces.

In summary, we present the following contributions of this study: (1) A method for reconstructing dimensions that preserves interesting cluster information in subspaces; (2) An analytical pipeline and a visualization tool to support the coordinated works of RDs and original dimensions in the interactive exploration of subspace clusters.

## 2 RELATED WORK

In this section, we first summarize interactive subspace analysis in the visualization community. We then discuss a few existing techniques on creating new dimensions for high-dimensional data visualization.

### 2.1 Sub-dimensional Space Analysis and Visualization

In high-dimensional data, many traditional clustering techniques suffer from the “curse of dimensionality”. Meaningful patterns, such as clusters or outliers, sometimes are prominent only in a few sub-dimensional spaces. In the data mining community, subspace clustering[11][13] provides a new path to the analysis of high-dimensional data. It localizes the search for relevant dimensions and uncovers clusters that exist in multiple, possibly overlapped subspaces. However, some new problems are brought out at the same time. One of the typical problems is the difficulty in handling enormous number of interesting subspace candidates, and understanding varied and possibly redundant clustering results. Interactive visualization is an effective solution for this problem, which is already proved by several works to date. Assent et al.[1][9] designed a visual component of WEKA for interactive exploration of clusters found by subspace clustering algorithms. ClustNails[15] is another visualization tool for navigating and interpreting the result of subspace clustering. Tato et al.[14] took advantage of topological and dimensional subspace similarity to deal with redundancy of the full result of SURFING. Nam et al.[10] provided a visualization framework to analyze the clustering result of ENCLUS from overview to detail. Nevertheless, the foregoing works simply focus on the interpretations of automatically generated results of subspace clustering algorithms. What differentiates our work is that we raised an interactive process on constructing subspaces, and this process is driven by the iteratively refined targets of users.

Compared with automatic subspace clustering in data mining, interactive subspace analysis driven by visualization techniques can better incorporate knowledge and questions of human users. Recently, multiple works have paid attention to interactive subspace exploration. Value and Relation Display (VaR)[18] is an early related work that employs MDS and Jigsaw layout to help users interactively construct subspaces based on the dimensional similarity. Turkay et al.[16] strengthened the interactive power of the dimension projection. They proposed a dual setting of interactions in data projection and in dimension projection. With this approach, users can jointly study the distribution of data items in a user-selected sub-dimensional space as well as the structure of the dimensions. Zhao et al.[21] also allowed users to brush a subset of the data items which can represent both major and minor trends of multi-dimensional data in parallel coordinate. Yuan et al.[20] designed a projection matrix to analyze multiple subspaces generated by a hierarchical manner of subspace exploration. The matrix well integrates the data projection plots and the relevant dimension projection plots into an interface for cross-comparison. A common feature of these interactive explorations of subspaces and clusters are stuck in the original dimension space. However, our work aims to create a few new informative dimensions, and enable users to construct subspaces from the new and original dimensions.

### 2.2 Dimension Partition and Dimension Representative

Only a few works have considered creating new dimensions for high-dimensional data visualization. We discuss them in two parts: dimension partition and dimension representative. Most of the existing research usually treats a dimension as a whole to measure the correlation with clustering results. This may ignore the highly partial correlation among some internal values in one dimension. Sharko et al.[12] described an extension to

Radviz called Vectorized Radviz (VRV) which partitions each dimension into more dimensions on categorical data. Zhou et al.[22] discussed how to apply the idea of VRV to a continuous dataset. Dimension partition provides a more flexible reordering dimension space in Radviz to support the exploration of meaningful clustering results. However, dimension partition often produces more dimensions. This may make the already high-dimensional space worse. Furthermore, the partial correlation may also exist between dimensions[8].

Compared to dimension partition, creating representative dimensions will be more steerable and practical. Each of representative dimensions can represent a subset of original dimensions. Yang et al.[19] suggested to create a representative dimension for each selected dimension cluster. These representative dimensions compose a low dimensional space for gaining a reasonable visualization. However, they did not explicitly point out how to generate representative dimensions. Turkay et al.[17] introduced three different methods to construct representative dimensions. They also provided quantitative measures to refine representative dimensions. These new dimensions can be together with the original dimension for an informed dimension reduction at different levels. Gleicher[6] stressed the importance of expert knowledge in generating representative dimensions. His approach enables experts to craft sets of dimension projection functions that align with user specifications. Therefore, the derived new dimensions can well represent concepts defined by the experts examples.

The concept of dimension reconstruction in our work is a bit similar to dimension representative. However, two aspects can distinguish our work from theirs. RDs are the representations of cluster structures in subspaces interesting to users, whereas representative dimensions are used to represent dimension groups. The second is that RDs are interactively or automatically generated from the data projections of subspaces, while representative dimensions are derived from original dimensions.

## 3 DIMENSION RECONSTRUCTION

In this section, we introduce our core concept, dimension reconstruction, and explain how to interactively reconstruct dimensions. A conceptual illustration of our method is presented in Fig.1, which is composed by three parts. The first part is a direct data analysis by data projection. Data projection techniques (e.g. PCA and MDS) allow a data set with high-dimensional points to be explored by projection in a 2D plane, as shown in Fig.1(a). The most important advantage of this 2D data projection technique is that human users can directly observe and operate data samples from a 2D screen space. This process can be described by Equation 1, where  $n$  denotes the number of original dimensions. As  $n$  increases, the result of direct data projection become meaningless.

$$\{\mathcal{R}^n\} \rightarrow \{\mathcal{R}^2\} \quad (1)$$

Interactive subspace analysis initially groups the original dimensions into multiple sub-dimensional spaces, and then performs data projection on each subspace to help users observe subspace clusters. This process is described by Equation 2, where  $i$  denotes the number of subspaces and  $n_i$  denotes the number of dimensions in the  $i$ th subspace. For example, we have a high-dimensional data set with three categories of data samples, marked by the colors red, green, and blue. Three subspaces of the data set are shown in Fig.1(b). In subspace1, green samples are separable. In subspace2, blue samples are well clustered. In subspace3, green and blue samples are separable, whereas red samples interlace exactly with the boundaries of blue and green samples. Thus, the users next goal of analysis might be to construct a new subspace in which the three categories of samples are well clustered.

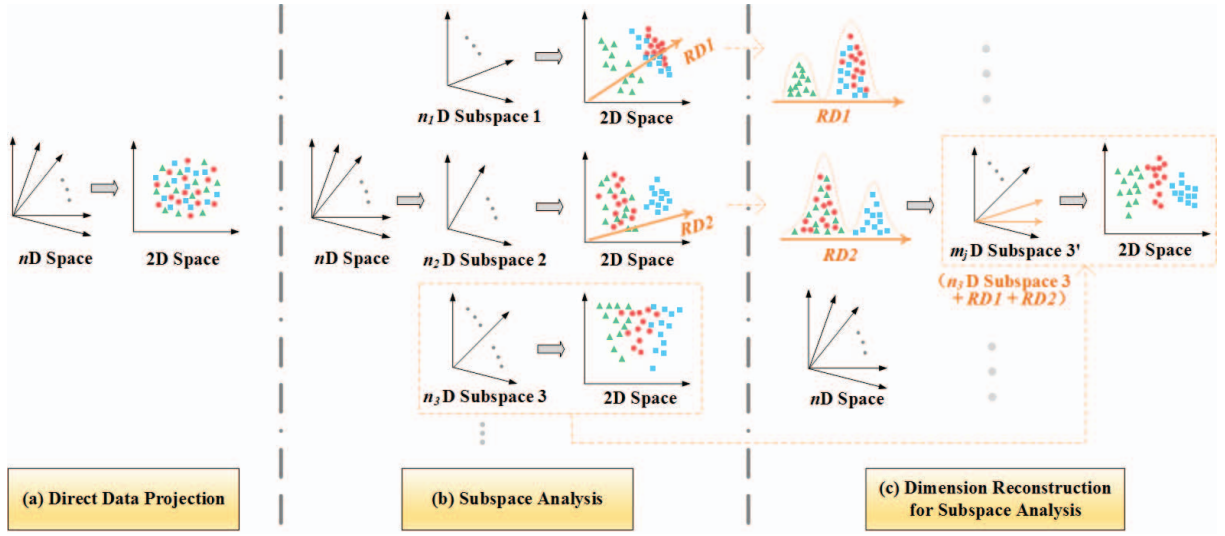


Figure 1: Conceptual illustration of dimension reconstruction for visually interactive subspace analysis.

$$\{\mathfrak{R}^n\} \rightarrow \begin{pmatrix} \mathfrak{R}^{n_1} \\ \mathfrak{R}^{n_2} \\ \vdots \\ \mathfrak{R}^{n_i} \end{pmatrix} \rightarrow \begin{pmatrix} \mathfrak{R}^2 \\ \mathfrak{R}^2 \\ \vdots \\ \mathfrak{R}^2 \end{pmatrix} \quad (2)$$

Our work aims to generate several new dimensions from the 2D data projection of a subspace to preserve interesting cluster information in the subspace. These new dimensions can be used to improve the clustering results of other subspaces. As shown in Fig.1(b), a straight line is interactively constructed as a new dimension (RD1) in the 2D data projection of subspace1, and data points are projected from the data projection onto RD1. Green samples densely concentrate in one region of RD1, and this region is clearly separated from other samples. Therefore, RD1 exhibits good capability of preserving the cluster information of green samples in subspace1. Similarly, RD2 is constructed to preserve the cluster information of blue samples in subspace2. Furthermore, users can construct several such new dimensions in one subspace to maintain the complex cluster structure of this subspace. To sum up, constructing a new dimension goes through two steps. The first step is constructing a subspace from the full original dimensional space, and the second step is constructing this new dimension in the 2D data projection of that subspace. Thus, we name this new dimension “Reconstructed Dimension”, which gives a sense of interactivity and describes a two-step process. In addition, we can also consider this new dimension as “Derived Dimension”.

These RDs, which demonstrate the preservation of interesting information, can be added into existing subspaces. We expect that the old cluster structures will be unbroken and by contrast, the clustering results will be improved. As shown in Fig.1(c). As a result of merging RD1 and RD2 into subspace3, the strong clustering information held by these two RDs leads to a clear representation of clusters in the data projection of subspace3. In fact, subspace3 has become a new subspace with  $n_3 + 2$  dimensions, which demonstrates the significance of these RDs in driving a new round of subspace exploration and helping users effectively construct new subspaces that can carry more meaningful cluster structures. At this point, subspace analysis that uses dimension reconstruction develops to a cyclic and user target-oriented process that can be described by Equation 3, where  $m$  denotes the summation of the original dimensions  $n$  and the

new dimensions  $r$ ,  $i$  denotes the number of subspaces,  $m_i$  denotes the number of dimensions in the  $i$ th subspace, and  $r_j$  denotes the number of new dimensions constructed in the  $j$ th subspace.

$$\{\mathfrak{R}^m\} \rightarrow \begin{pmatrix} \mathfrak{R}^{m_1} \\ \mathfrak{R}^{m_2} \\ \vdots \\ \mathfrak{R}^{m_i} \end{pmatrix} \rightarrow \begin{pmatrix} \mathfrak{R}^2 \\ \mathfrak{R}^2 \\ \vdots \\ \mathfrak{R}^2 \end{pmatrix} \rightarrow \begin{pmatrix} \mathfrak{R}^{r_1} \\ \mathfrak{R}^{r_2} \\ \vdots \\ \mathfrak{R}^{r_j} \end{pmatrix} \quad (3)$$

When the cluster structure in a subspace is sufficiently simple, users can directly reconstruct dimensions through interactive operations on 2D data projection. However, when users encounter a complicated cluster structure in a subspace, (e.g., a single straight line is unable to separate multiple clusters), we need to adopt some inset algorithms to automatically reconstruct the candidates of new dimensions, which are offered to users as options. If users interactively tag each interesting cluster in the 2D data projection of a subspace, we can regard the high-dimensional data as the labeled data set. Thus, the question that follows is how to utilize label information in finding optimal projections that make high-fidelity preservation of interesting information.

Inspired by the idea of Choo et al.[3], which used classical linear discriminant analysis (LDA) for better retaining classification in low-dimensional space, we currently have introduced classic LDA into our analytical pipeline to help users automatically find optimally linear projections on the 2D data projection of a subspace. For two-class problems, such as the samples in Fig.2, LDA leads to an optimal projection line on which the distance between two classes is maximized, whereas the variance of each class is minimized. As shown in Fig.2, if the samples are projected onto RD2, then the samples in the two classes will be mixed near the boundary. If they are projected onto RD1 as determined by LDA, then the two classes will clearly be separable.

In addition, we need to standardize values on RDs. For high-dimensional data, the value spans in different dimensions significantly differ. Z-score standardization is a popular method for comparing a sample to a standard normal deviation. In our method, Z-score standardization is used on both original dimensions and derived RDs to convert diverse value spans into small normalized ranges without destroying their original data distributions.



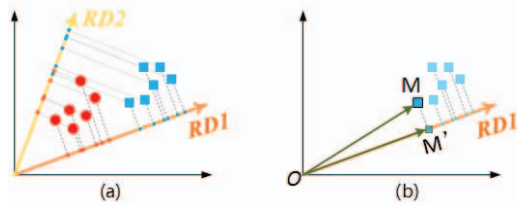


Figure 2: Conceptual illustration of LDA. (a) The two classes are separated well when projected onto RD1. (b)  $M'$  on RD1 is the projected point of  $M$ .

## 4 VISUALIZATION AND INTERACTION

### 4.1 Analytical Workflow

We propose an analytical workflow to support the coordinated works of RDs and original dimensions for the visually interactive exploration of subspace clusters. A number of visualization and interaction techniques are incorporated into the workflow. Shown as Fig.3, the first step is to interactively construct subspaces. To assist users in selecting dimensions to construct subspaces, we provide visual representations of dimension information from overview to detail. In the overview aspect, a dimension projection technique is employed to display the overall relationships among dimensions in a 2D scatterplot. In the detail aspect, a dimensional histogram shows data distribution on each dimension. The second step is to explore meaningful data patterns in the subspaces which are constructed in the previous step. We provide users with multiple data views. Each data view is the 2D data projection of a subspace. In a data view, users can observe cluster structures and adjust the weights of dimensions, and thus, obtain different projection results. Users can also merge subspaces and track user-tagged clusters in different data views. The third step is to reconstruct dimensions. We provide two methods to reconstruct dimensions: manual operation and automated generation. Reconstructing dimensions is allowed in any subspace. RDs are coupled with the original dimensions to start a new round of subspace analysis.

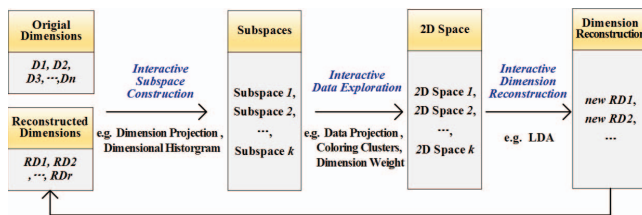


Figure 3: Proposed analytical workflow.

### 4.2 User Interface

A user interface is particularly designed to help users implement our proposed workflow. The interface overview is presented in Fig.4, which includes a data exploration area and a dimension exploration area. A synthetic data set is used in Fig.4. This data set includes 12 dimensions and 750 samples without marking classifications, and has been used in [14][5] for their case studies.

#### 4.2.1 Hierarchical Dimension Projection

The dimension projection plot (Fig.4(a)) is the operating area for interactive subspace construction. The dots and stars in the plot represent original dimensions and RDs respectively. The positions of dots and stars are generated via classical MDS according to pairwise relationships between individual dimensions. In classical

MDS, the dissimilarity transformation is  $D = 1 - P$ , whereas  $D$  is dissimilarity distance matrix,  $P$  is similarity distance matrix computed by Pearson's correlation coefficient. Thus, closely related dimensions are adjacent to one another in the dimension projection. Users are allowed to interactively choose the dimensions in this plot to construct a subspace. After a subspace is established, dimension projection is updated to display dimensions that only exist in the former subspace. Hence, a hierarchical high-dimensional exploration is provided to allow users to drill down the relationships of dimensions and build subspaces in global or local dimensional space. A thumbnail of full-dimension projection is located on the top right of the dimension projection plot. It serves as a navigator to prevent the loss of analysis goals in local explorations. The highlighted points in the navigator are the dimensions used in the current dimension projection. For example, Fig.5(a) shows a MDS projection of 221 dimensions, and 38 dimensions are selected to construct a subspace. Simultaneously, the dimension projection is updated to Fig.5(b), in which 9 dimensions of the 38 dimensions are highlighted.

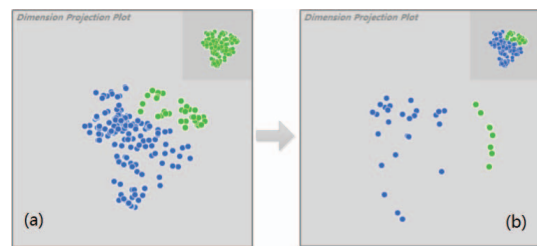


Figure 5: Illustration of hierarchical dimension projection.

#### 4.2.2 Stacked Dimensional Histogram

We present stacked dimensional histogram to provide the distribution information of original data on a dimension to assist users in constructing subspaces. A dimensional histogram plot is a probability distribution histogram that is similar to the gray-scale histogram in digital image processing. For a data set labelled into  $c$  classes, the probability distribution of each class would be stacked in a general histogram with  $c$  colors. When users select an original dimension or a RD from the dimension projection plot, a histogram of that dimension will be added into the histogram list. If users select a data plot, then the histograms of all dimensions in that subspace will be displayed. For example, Fig.4(b) shows all dimensional histograms of 12 dimensions in the synthetic data set, including 6 dimensions with two Gaussian clusters, 3 dimensions containing four multi-modal Gaussian distributions with different means and variances, and 3 dimensions with uniformly distributed random noises.

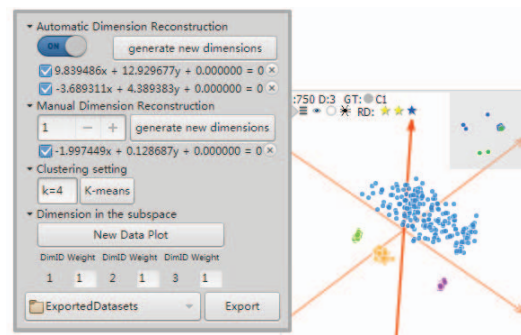


Figure 6: Menu of data plot.

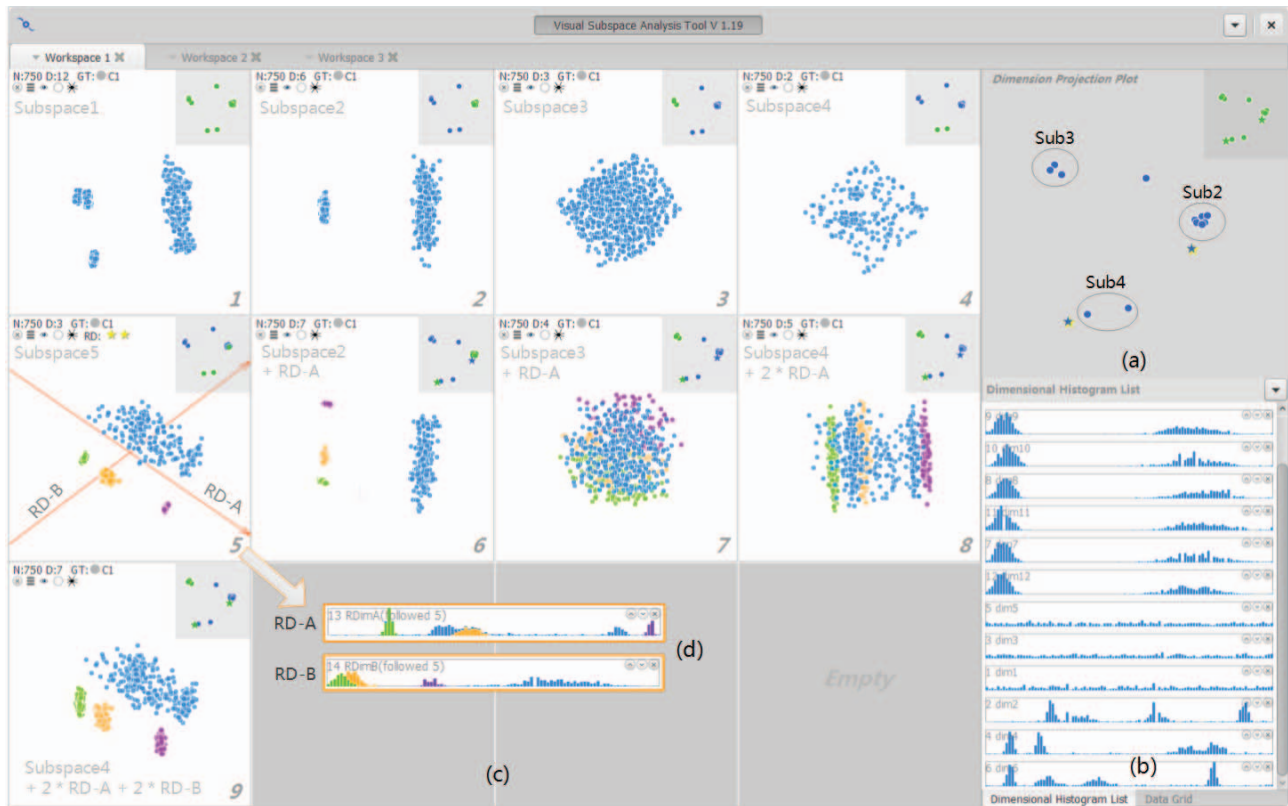


Figure 4: Overview of user interface with (a) dimension projection plot, (b) dimensional histogram list, (c) multi-viewed data exploration area, and (d) stacked dimensional histograms of two new dimensions reconstructed from cluster-tagged subspace5.

#### 4.2.3 Multi-Viewed Data Exploration

**Data Plot.** A data exploration area that involves multiple plots is shown in Fig.4(c). The data plots are arranged according to their creation time. A data plot has three major components: a toolbar, a navigator and a scatterplot. The toolbar contains a pop-up menu and several information tips. The pop-up menu provides four operation areas, namely, automated dimension reconstruction, manual dimension reconstruction, k-means clustering setting, and subspace dimension weight setting. The information tips on the toolbar include the number of data points, the number of subspace dimensions, the glyphs of the ground truth of classifications, and the glyphs of created RDs in this subspace. In Fig.6, the subspace has 3 dimensions and 750 data samples, the ground truth of data classification has only one class, and users have reconstructed two dimensions by automated method and one dimension by manual manipulation. The navigator is a thumbnail of the projection from all dimensions. The scatterplot is the MDS of the data samples, and the distance function of MDS is Euclidean distance. The shape of the data points in the scatterplot corresponds to the glyphs of ground truth.

**Clustering and Cluster Tracking.** Data projection in a data plot can be regarded as a visual clustering result. Users can interactively code their interested clusters by colors. The colors will be maintained in the data plots that will be created later for tracking the distribution of the tagged points. We also offer a k-means clustering method. Users can first cluster non-labelled data by k-means algorithm, then tag the original data by clustering result, and finally apply LDA to the data. As shown in Fig.4, subspace1 includes all 12 dimensions in the synthetic data set. Subspaces 2, 3, and 4 are constructed based on their proximities in the dimension projection plot. Subspace2 includes 6 dimensions

with 2 Gaussian clusters. Subspace3 includes 3 dimensions with random noises. Subspace4 includes 2 dimensions with 4 Gaussian clusters. Subspace5 is composed by Dim2, Dim6, and Dim12, and forms four clear clusters in data plot 5. We mark these four clusters and track them in the following data plots.

**Reconstructing Dimensions.** Users can manually reconstruct dimensions in each data plot through directly drawing lines. Automatic dimension reconstruction is allowed after tagging clusters in a data plot. RDs can be algorithmically built according to the classes coded by users, and are shown as lines on the data plot. The operations of rotation and translation are provided to adjust the parameters of a line. When dimension reconstruction is completed, the dimension projection plot is updated to include the new dimensions. Hence, RDs can be included in the construction of new subspaces. Furthermore, users can adjust the weights of reconstructed dimensions to increase their influences on the clustering effects of new subspaces.

For example, we reconstruct two new dimensions, namely, RD-A and RD-B, by LDA in the well-clustered subspace5. Shown as Fig.4(d), the blue cluster information can be maintained on RD-B, whereas the other three clusters can be preserved on RD-A. Subspace6 is built by adding RD-A into subspace2. Data plot 6 shows that the cluster on the left of subspace2 can be classified into three small clusters. Subspace7 is formed by adding RD-A into subspace3. No improvement occurs in subspace7, because the 3 dimensions in subspace3 are random noises. If we double the weight of RD-A in subspace7, then some patterns can be found on the noises in data plot 8. When we add RD-A and RD-B into subspace3 and then double their weights, the influences from the three noises dimensions are extremely small, and four clusters are clearly observed in subspace9.

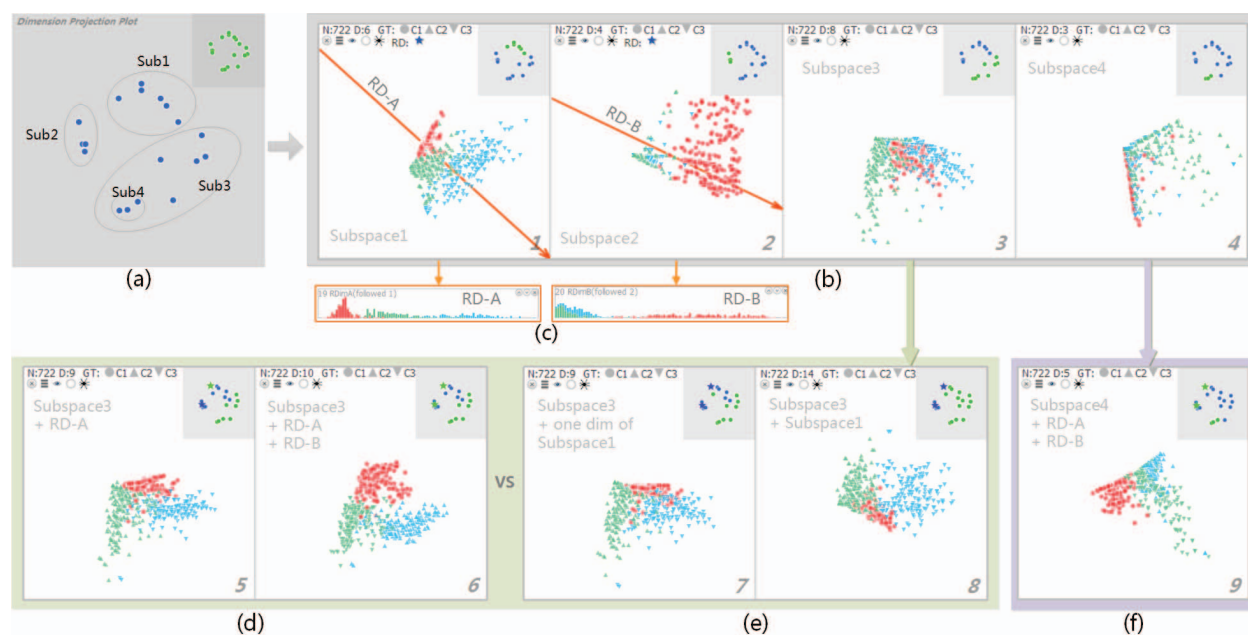


Figure 7: Experiments on the food data set: (a) projection of all dimensions; (b) four subspaces constructed during the first round of analysis; (c) the dimensional histograms of two RDs: namely RD-A from subspace1 and RD-B from subspace2; (d) improved data projections by adding the RDs into subspace3; (e) improved data projections by adding the original dimensions into subspace1 into subspace3. (f) the result of data projection by combining subspace4 with the two RDs.

## 5 CASE STUDY

### 5.1 USDA Food Composition Data

In this case study, we analyze the USDA Food Composition Data, which is a complete collection of raw and processed foods characterized by their composition in terms of nutrients, for demonstrating how our approach improves the clustering results of subspaces. This data set contains 722 samples and 18 dimensions. It was used in [14][20] for their case studies.

In [20], food data samples were divided into three categories. We employ their classification result and interactively construct 4 interesting subspaces mentioned by [20]. The dimension projection of all features are shown in Fig.7(a). Subspace1 consists of six mineral features: Calcium, Carbohydrate, Fiber, Iron, Magnesium, and Manganese. Subspace2 mainly contains measures of vitamin content, including Vit\_B12, Vit\_B6, Vit\_D, and Protein. Subspace3 contains all the remaining dimensions. Subspace4 includes three dimensions in subspace3, namely, Beta\_Carot, Vit\_A and, Vit\_C.

Four data plots in Fig.7(b) present the data projections of the four subspaces. In data plot 1, the three categories of samples are grouped individually and nearly linearly separable but have slightly obscure boundaries. We reconstruct a new dimension, RD-A, to keep the clustering information in subspace1. In data plot 2, although the red samples are not densely clustered, a clear gap exists between the red samples and the other two categories. Thus, we reconstruct RD-B to keep the cluster structure in subspace2. In data plots 3 and 4, the visual clustering results are worse. Fig.7(c) shows the probability distributions of samples and clusters on RD-A and RD-B. In the next round of analysis, we use the preserved cluster information on RD-A and RD-B to improve the data projections of subspaces 3 and 4.

We construct subspace5 by combining all the dimensions of subspace3 with RD-A. By comparing data plot 5 with data plot 3, we observe that a portion of the red samples is separated from the overlapping region. We then add RD-A and RD-B into subspace3, and the three categories are clustered well in Fig.7(d). Similarly,

we add RD-A and RD-B into subspace4, and the visual clustering result in subspace4 is obviously improved, as shown in Fig.7(f).

If we directly add the dimensions of subspace1 into subspace3, would the visual clustering effect be improved? We have conducted two experiments. Firstly, the dimensions of subspace1 are added one by one into subspace3. Data plot 7 shows the best clustering effect in this experiment. Based on the comparison of data plots 5 and 7, adding an original dimension of subspace1 yields a less effective result compared with using RD-A. In the second experiment, all six dimensions in subspace1 are added into subspace3, shown in Data plot 8. Simultaneously adding numerous dimensions results in obvious changes in cluster structure, and the clustering result of the red samples is not that good as shown in data plot 5.

The results of these experiments prove that dimension reconstruction is useful in providing users with an improved perception of subspace clusters. RDs can preserve the cluster information formed by all the dimensions in subspaces, and then are used to improve the clustering results of other original subspaces. This process can also be regarded as a user-target, cluster-oriented process that simplifies dimensions.

### 5.2 CADASTER Molecule Data

We will now demonstrate how our approach helps users construct subspaces with more meaningful cluster structures by using dimension reconstruction. The Molecule Data is a real-world data set provided by CADASTER Challenge[7]. This data set describes the structural information of molecules. Each sample is a molecule, and each dimension is a SimulationsPlus descriptor (e.g. elements, formal charges and bonds). There are 664 samples and 222 molecular descriptors in this data set. One of the descriptors is the measurement of environmental toxicity  $\log(IGC50 - 1)$  of a molecule. According to the value of the environmental toxicity, all 664 samples are divided into two classes (362 molecules with positive environmental toxicity and 302 molecules with



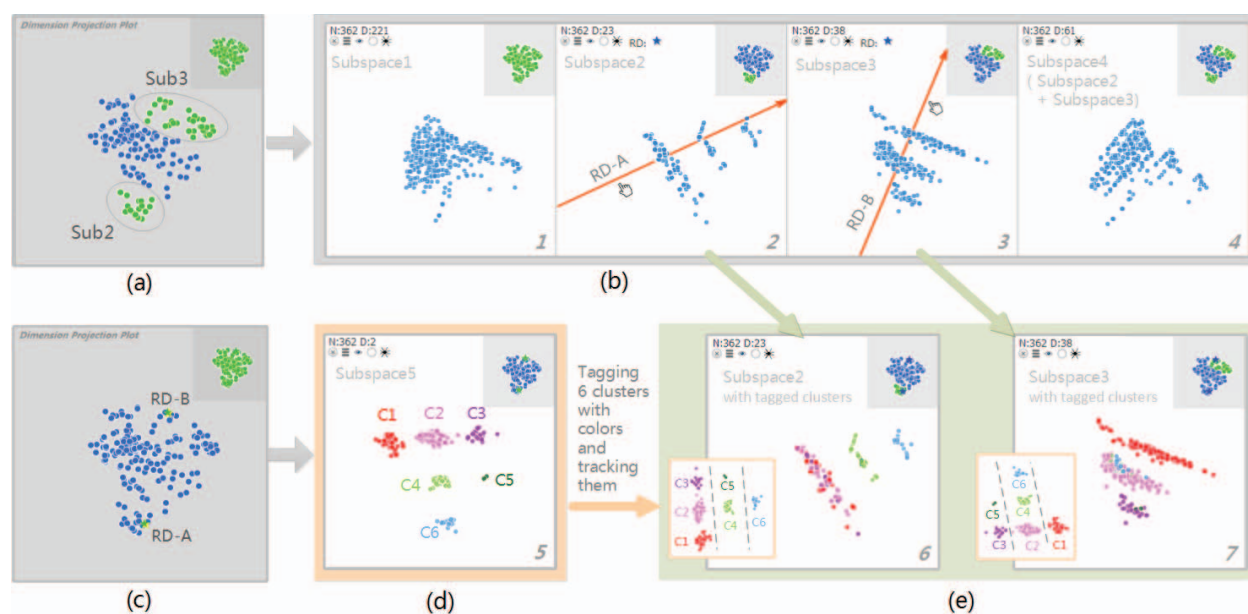


Figure 8: Experiments on the molecule data set: (a) the dimension projection of the molecule data set; (b) four subspaces constructed during the first round of analysis; (c) updated dimension projection after reconstructing two dimensions from subspace2 and subspace 3; (d) 6 clusters formed in subspace5 which is constructed by the two RDs; (e) tagging 6 clusters in subspace5 by colors and tracking them to gain the relationship between two cluster structures formed in subspace2 and subspace3.

non-positive environmental toxicity). In this case, we will explore all the molecules with positive environmental toxicity.

The projection of 221 dimensions (excluding the environmental toxicity) is shown as Fig.8(a). Four subspaces are constructed in Fig.8(b). First, we select all 221 dimensions to construct subspace1. Data plot 1 shows that most of the data concentrates densely, which offers little cluster information. Then we notice that there are two groups of contiguous dimensions on the dimension projection with one group in the lower left corner and the other in the upper right corner. Thus, two subspaces are constructed by selecting the two groups of dimensions: subspace2 with 23 dimensions and subspace3 with 38 dimensions. Data plots 2 and 3 depict that both subspaces 2 and 3 form three strip clusters. We now have a further analysis target, what are the relationships between these two cluster structures in these two different subspaces. We then try to find the answer by merging subspaces 2 and 3. Unfortunately, the two original cluster structures are missed in data plot 4.

We next analyze the relationships between the two cluster structures by using dimension reconstruction. Two dimensions, RD-A and RD-B, are reconstructed from subspaces 2 and 3 respectively. Then, the dimension projection plot is updated to show the two RDs in Fig.8(c). We directly select these two RDs to construct subspace5. Fig.8(d) shows that six clusters with clear boundaries formed in subspace5. We then use different colors to label the six clusters from C1 to C6 individually and observe their distributions in subspaces 2 and 3. As shown in Fig.8(e), the three clusters in subspace2 are grouped as C1, C2, C3, C4, C5, and C6, and three clusters in subspace3 are grouped as C3, C5, C2, C4, C6, and C1. To sum up, we obtain some interesting conclusions: (1) the cluster structures in subspaces 2 and 3 are completely different, while the samples of three clusters of each subspace are interlaced; (2) these 362 molecules can be divided into six categories.

This case explains that dimension reconstruction is useful in exploring the cluster structures formed in different existing subspaces. Dimension reconstruction can provide an intuitive way to help users quickly merge cluster structures that do not coexist in the original subspaces.

## 6 EXPERT REVIEW AND DISCUSSION

To further evaluate our method and the user interface design, we interviewed two experts from University C. The research interest of Expert A is recommendation systems in data mining. Expert B excels in protein structure prediction in bioinformatics. Both of them are considered as the intended users of the approach. During the interview, we first explained the basic concept of our method and demonstrated our interface using the two case studies. After they tried out the tool, we discussed its advantages and limitations. In this section, based on the feedback of Experts A and B, we summarize the advantages and disadvantages of our approach and provide potential methods to improve it.

Both experts recognized that dimension reconstruction is a novel and interesting approach to explore subspace clusters in high dimensional data. Expert A commended, “Dimension reconstruction, as an alternative to dimension reduction, can be useful in real task applications.” Expert B added, “Another advantage of this work is the pipeline incorporates interactivity and user feedback into dimension reconstruction and subspace construction.” Both experts also encouraged us, “Dimension reconstruction may be helpful for the explorative analysis of several data patterns, not only clusters and cluster structures.”

The discussion of limitations mainly focused on the issues of usability and scalability. The first usability suggestion is to incorporate more guidance information in the explorative analysis of subspaces and clusters. Though a number of visualizations and interactions are applied in our tool to support each analytical step, this is a little complicated and tedious in a way. Additionally, some strongly correlated dimensions are used for selecting subspace in the two cases. This is not always a rational way because a meaningful subspace may be composed of very unrelated dimensions. In fact, the idea of dimension reconstruction is more suitable in the situation that users have already got some interesting subspaces, such as the beginning subspaces in our cases. And what’s more, our intended users would need to have a good general understanding of the concepts of clustering and projection.

Another usability problem is that the current tool depends on linear methods. Expert B pointed out, “The classic MDS limits dimension and data projections to a reflection of linear relationships.” Expert A pointed out that “typical LDA assumes that clusters follow Gaussian or convex distribution.” With regard to this limitation, we aim to solve it from two perspectives. On the one hand, diverse nonlinear techniques can be considered, such as spectral clustering and kernel-based clustering. On the other hand, we plan to make the interactive dimension reconstruction more flexible, such as using stepwise lines or curves.

Furthermore, the issue of possible data distortion should be noticed. In 2D space, visual entities effectively support the use of high cognitive abilities of human users. However, there is a risk that data are distorted in the space conversion from  $n$ -D to 2D. Expert A commented, “Its important to make users more aware of possible data distortion caused by 2D projections, because this may affect the conclusions drawn from the data.” In addition, RDs do not have specific physical meanings in the real world like what original dimensions have. We should help users explore the links between RDs and original dimensions. One possible method to achieve this is to use parallel coordinate plots.

Both experts agreed that this approach is able to deal with truly high dimensional data sets, but they also raised some concerns of scalability when the input consists of huge amounts of data items or dimensions. Firstly, the dimension and data projection is a time-consuming task. This issue can be improved by GPU-based parallel computing or a density map with an appropriate resolution. Secondly, the current tool cannot efficiently present high numbers of data items, dimensions or subspaces at a time. We plan to provide users with hierarchical navigators or heat-map-like depictions.

## 7 CONCLUSION

This paper presents an approach called Dimension Reconstruction for visually exploring subspace clusters in high-dimensional data. Our motivation is to help users improve visual clustering results of subspaces by taking advantage of various cluster structures in subspaces. To reach this goal, new dimensions are interactively or algorithmically reconstructed from the 2D data projections of subspaces for cluster information preservation. After adding the appropriate RDs into other original subspaces, the preserved cluster information will be well exposed, which will improve the clustering results of these subspaces. More importantly, dimension reconstruction plays a significant role in the interactive subspace analysis cycle. As the adoption of RDs, it will be easy for users to construct new subspaces that have informatively merged cluster structures. An analytical workflow and a visualization tool are provided in this approach to support the co-work of RDs and original dimensions. In the case study, we demonstrated the effectiveness of our approach through the analysis of two real-world data sets. We also discussed the limitations of our approach and proposed some potential methods to improve it.

## ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61103108, 61170204 and 61402540. (video demo for this paper: <https://youtu.be/dYaInHzviLs>)

## REFERENCES

- [1] I. Assent, R. Krieger, E. Müller, and T. Seidl. Visa: visual subspace clustering analysis. *ACM SIGKDD Explorations Newsletter*, 9(2):5–12, August 2007.
- [2] S. Barlowe, J. Yang, D. J. Jacobs, D. R. Livesay, J. Alsakran, Y. Zhao, D. Verma, and J. Mottonen. A visual analytics approach to exploring protein flexibility subspaces. In *Proceeding of IEEE Pacific Visualization Symposium (PacificVis)*, pages 193–200, 2013.
- [3] J. Choo, H. Lee, J. Kihm, and H. Park. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 27–34, 2010.
- [4] B. J. Ferdosi, H. Buddelmeijer, S. Trager, M. H. F. Wilkinson, and J. B. T. M. Roerdink. Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 35–42, 2010.
- [5] B. J. Ferdosi and J. B. T. M. Roerdink. Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Computer Graphics Forum*, 30(3):1121–1130, 2011.
- [6] M. Gleicher. Explainers: Expert explorations with crafted projections. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2042–2051, 2013.
- [7] <http://www.cadaster.eu/node/65>. Cadaster challenge dataset.
- [8] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 111–120, 2011.
- [9] E. Müller, I. Assent, R. Krieger, T. Jansen, and T. Seidl. Morpheus: interactive exploration of subspace clustering. In *Proceedings of the 14th ACM SIGKDD*, pages 1089–1092, August 2008.
- [10] J. E. Nam and K. Mueller. Tripadvisor-n-d: A tourism-inspired high-dimensional space exploration framework with overview and detail. *IEEE Transactions on Visualization and Computer Graphics*, 19(2):291–305, February 2013.
- [11] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, August 2004.
- [12] J. Sharko, G. Grinstein, and K. A. Marx. Vectorized radviz and its application to multiple cluster datasets. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1444–1427, 2008.
- [13] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong. A survey on enhanced subspace clustering. *Data mining and knowledge discovery*, 26(3):332–397, 2013.
- [14] A. Tatu, F. Maas, I. Farber, E. Bertini, T. Schreck, T. Seidl, and D. Keim. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 63–72, 2012.
- [15] A. Tatu, L. Zhang, E. Bertini, T. Schreck, D. K. S. Bremm, and T. Landesberger. Clustnails: Visual analysis of subspace clusters. *Tsinghua Science and Technology*, 17(4):419–428, 2012.
- [16] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions: a dual visual analysis model for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, December 2011.
- [17] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, 2012.
- [18] J. Yang, D. Hubball, M. Ward, E. Rundensteiner, and W. Ribarsky. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *IEEE Transactions on Visualization and Computer Graphics*, 13(3):494–507, 2007.
- [19] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on Data visualisation 2003(VisSym'03)*, pages 19–28, 2003.
- [20] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, December 2013.
- [21] G. Zhao, W. James, and L. R. S. Markov chain driven multi-dimensional pattern analysis with parallel coordinates. In *Proceedings of Vision, Modeling, and Visualization (VMV)*, pages 191–198, 2012.
- [22] F. Zhou, W. Huang, J. Li, Y. Huang, Y. Shi, and Y. Zhao. Extending dimensions in radviz based on mean shift. In *Proceeding of IEEE Pacific Visualization Symposium (PacificVis)*, pages 111–115, 2015.