A co-analysis framework for exploring multivariate scientific data

Xiangyang He, Yubo Tao *, Qirui Wang, Hai Lin

State Key Lab of CAD&CG, Zhejiang University, China

ARTICLE INFO

ABSTRACT

In complex multivariate data sets, different features usually include diverse associations with different variables, and different variables are associated within different regions. Therefore, exploring the associations between variables and voxels locally becomes necessary to better understand the underlying phenomena. In this paper, we propose a co-analysis framework based on biclusters, which are two subsets of variables and voxels with close scalar-value relationships, to guide the process of visually exploring multivariate data. We first automatically extract all meaningful biclusters, each of which only contains voxels with a similar scalar-value pattern over a subset of variables. These biclusters are organized according to their variable sets, and biclusters in each variable set are further grouped by a similarity metric to reduce redundancy and support diversity during visual exploration. Biclusters are visually represented in coordinated views to facilitate interactive exploration of multivariate data from the similarity between biclusters and the correlation of scalar values with different variables. Experiments on several representative multivariate scientific data sets demonstrate the effectiveness of our framework in exploring local relationships among variables, biclusters and scalar values in the data.

1. Introduction

Scientific simulations often generate data sets with multiple variables for complex physical phenomena. These variables generally include hidden associations, because of their collective applications in the simulation model (Carr and Duke, 2013). For instance, a hurricane is a rapidly rotating storm system characterized by a low-pressure center, strong winds, including heavy rains in climate simulation. However, the heterogeneity and complexity of multivariate data make the extraction of interesting associations, which are typically located only in the subspaces of variables and subsets of voxels, quite challenging. For example, the hurricane eye may possess a strong association with pressure and wind, while the eyewall clouds may be strongly associated with the water vapor and cloud moisture (Liu and Shen, 2016). Therefore, it would be meaningful to extract hidden associations between variables locally and detect local features based on the associated variables.

In multivariate data analysis, a broad variety of techniques have been proposed to explore the associations of variables/voxels in the data. At the voxel level, many clustering algorithms in the data mining field have been used to automatically group associated voxels as features (Tzeng and Ma, 2004; Van Long and Linsen, 2009; Wu et al., 2015). Similarity metrics between voxels are generally defined on the scalar values of all variables. Therefore, it may be challenging to detect features that only depend on a subset of variables, because other unrelated variables may have a negative impact on clustering, owing to the curse of dimensionality. At the variable level, many correlation metrics between variables such as gradient similarity measure (GSIM) (Sauber et al., 2006), Pearson product-moment correlation coefficients (Sukharev et al., 2009), and mutual information (Wang et al., 2011), have been recently proposed. These metrics are usually the average correlation values of all voxels. Because a subset of variables may be strongly associated in a local region, it is desirable to extract these local associations among variables in different local regions rather than the global associations based on all voxels. Furthermore, these methods automatically analyze the similarity of voxels or the correlation of variables independently. Variables and voxels should be analyzed together rather than separately to determine the local associations between them.

Multi-dimensional transfer functions can consider both variables and voxels in the manual classification of features of interest. For example, a feature can be specified by gradually selecting the scalar value intervals of a few variables in the parallel coordinate (Zhao and Kaufman, 2010; Guo et al., 2011) or by specifying Gaussian functions in a scatter plot matrix (Lu and Shen, 2017). In this case, these variables may be associated and their correlated scalar value intervals include the definition of the feature, the voxels of which include a similar scalar-value pattern over these associated variables. In this paper, we call it a bicluster between variables and voxels which denotes two subsets of variables and

voxels with close scalar-value relationships. While manual specification exhibits the advantage of flexibility in obtaining a variety of biclusters, it can be laborious and hinder comprehensive coverage of the data in exploratory analysis. In addition, this specification heavily depends on the domain knowledge and skills of users to determine the associated variables and their correlated scalar values in a meaningful feature. In practice, users must iteratively to refine the specification to search for satisfactory results. When there are many variables in multivariate data, it becomes nearly impossible to find all meaningful features because of the large size of the search space. This introduces the need to automatically find all meaningful biclusters between variables and voxels.

To address these requirements, we propose a co-analysis framework based on biclusters for interactively exploring multivariate data, including bicluster generation, analysis and visual exploration. Our framework first generates all biclusters (local relationships) by simultaneously clustering variables and voxels (Section 4). The scalar values of voxels demonstrate a similar pattern on the variables in a bicluster, including a specific value combination. Because each bicluster is associated with a subset of variables, they can be organized by the variable set to hierarchically explore variables and biclusters. As some of the biclusters may overlap with each other because of the completeness of bicluster generation, biclusters with the same variable set can be grouped based on a similarity metric between biclusters (Section 5). To visually explore biclusters, we design a visual analysis system with four coordinated views to reveal three-faceted relationships in multivariate data (Section 6): variables, biclusters and scalar values. An association matrix between variables and biclusters is designed to facilitate searching for correlated variables. The biclusters in a variable set are displayed through dimensionality reduction to analyze the similarity of biclusters. An enhanced parallel coordinate is used to explore the correlation of scalar values of a group of biclusters or a bicluster. Based on the exploration guideline, we experiment on several multivariate data sets in different domains to demonstrate the effectiveness and usefulness of our co-analysis framework (Section 6).

This paper is an extended version of our conference paper (He et al., 2018a) of *IEEE Scientific Visualization Conference (SciVis)*. In detail, the extensions of our conference paper include the following:

- a detailed description of bicluster generation based on the variance minimization method (Section 4),
- a new correlation measure between variables based on biclusters (Section 5.1),
- an enhanced parallel coordinate for presenting the statistical distribution of associated voxels (Section 6.3), and,
- and two new experiments of the ionization front instability data (Section 6.1) and hurricane Isabel data (Section 7.1).

2. Related work

Multivariate data analysis and visualization, as one of the major challenges associated with scientific visualization, have long been active research topics (He et al., 2018b; Kehrer and Hauser, 2013; Fuchs and Hauser, 2009). In this section, we briefly review previous research related to correlation analysis and interactive classification of multivariate data.

2.1. Correlation analysis

Finding hidden correlations in multivariate data is a common challenge in many computational analysis fields. Many correlation analysis methods have been proposed over the years to explore the relationships between variables and scalar values.

Information theory provides a theoretical framework of measuring the global correlation between variables. Biswas et al. (2013) employed mutual information to measure the informativeness of one variable about the other variable and grouped variables based on mutual information in a graph-based approach. Wang et al. (2011) applied transfer entropy to investigate the causal relationships between variables in time-varying multivariate data, and the correlations between variables were visually encoded in a node-link diagram. These methods consider the entire data, making it challenging to capture local correlations between variables in different regions.

Many local correlation metrics have been proposed to capture the correlation at each voxel, and the correlation between variables can be measured by the summation of the correlation values of all voxels. Sauber et al. (2006) proposed a gradient similarity measure (GSIM) and a local correlation coefficient to measure the local correlation at each voxel. In addition, they introduced a multifield-graph to present an overview of the correlation between variables. Gosink et al. (2007) derived a correlation field by taking the normalized dot product between two gradient fields from two variables. Jänicke et al. (2007) extended the concept of local statistical complexity to multi-fields to identify spatiotemporal structures that exhibit the same behavior in multivariate data. Sukharev et al. (2009) applied the Pearson product-moment correlation coefficient on temporal curves of voxels to analyze the linear correlation between two variables in time-varying multivariate data. Nagaraj et al. (2011) presented a gradient-based correlation criterion, the norm of a partial derivative matrix, to capture the interactions between multiple scalar fields. The correlation field is visualized to detect regions with high correlation values. In this study, we simultaneously cluster variables and voxels to automatically extract biclusters and employ the biclusters, subsets of voxels instead of all voxels as in previous methods, to better measure the correlation of a subset of variables in local regions.

In addition to the correlation between variables, a specific local relationship between the scalar values in different variables has recently received considerable attention. Biswas et al. (2013) applied the surprise and predictability metrics to measure the variability of one scalar value with respect to another variable. Liu and Shen (2016) considered two-way interactions between the scalar values of two variables as information flow and employed association rules to model these interactions. Because one bicluster includes voxels with a similar scalar-value pattern over a subset of variables, we can directly analyze such local relationships between scalar values in multiple variables.

2.2. Interactive classification

Feature classification is essential for effective exploration of multivariate data. For multivariate non-spatial data, there are many well-studied visualization and interactive exploration techniques that display the distribution and the relationship of data. These techniques include parallel coordinates and scatter plots. Parallel coordinates illustrate information on each dimension, including the correlation between neighborhood axes (Inselberg, 1985; Inselberg and Dimsdale, 1990). Scatter plots display high-dimensional data using dimensionality reduction techniques, such as multidimensional scaling (MDS) (Cox and Cox, 1994) and t-SNE (van der Maaten and Hinton, 2008), because it can be easy to identify and select clusters after projection. Previous classification methods of multivariate data primarily rely on these techniques.

In multivariate data visualization, Zhao and Kaufman (2010) introduced parallel coordinates for multi-dimensional transfer function design, which defines a feature by specifying several scalar value intervals of associated variables. Guo et al. (2011) proposed a novel transfer function design interface combining the parallel

coordinate and MDS plots to facilitate feature specification in multivariate data. Lu and Shen (2017) presented a bottom-up subspace exploration workflow that allows users to interactively design multivariate transfer functions and introduced additional information that guides users in the selection of subspaces, discovering interesting features. While multi-dimensional transfer functions flexibly enable to specify features, it can be time-consuming and challenging to search for all meaningful features in the exploratory analysis. In this study, we automatically extract all meaningful biclusters, and visually explore the similarities of biclusters in the scatter plot as well as the correlation of scalar values in a bicluster in the parallel coordinate.

3. Overview

Biclustering, also known as co-clustering or simultaneously clustering (Hartigan, 1972), addresses this problem by simultaneously clustering both rows (objects) and columns (attributes) in a variety of fields such as DNA micro-array data analysis, text mining, and information retrieval. Biclustering has also been widely used in visualization such as visualizing relevant genes and conditions of gene-expression matrices (Santamaría et al., 2008), reducing visual clutter of a large number of edges (Liu et al., 2017), interpreting the subspace clustering result in high-dimensional data (Liu et al., 2015), and discovering coordinated relations from textual datasets (Sun et al., 2016). In the field of data mining, the biclustering method can effectively extract cohesive objects with a similar scalar-value pattern over a subset of attributes. This study applies biclustering to the analysis of multivariate scientific data. If variables and voxels are considered attributes and objects, respectively, we can extract cohesive voxels with a similar scalar-value pattern over a subset of variables using the biclustering method, i.e., a bicluster in this paper. Therefore, a bicluster is composed of a subspace of variables and a subset of voxels, and these voxels demonstrate a similar scalar-value pattern on these variables, including a specific value combination of these variables. In other words, these variables are locally associated in the space of these voxels, and the corresponding scalar values of these variables are strongly correlated with each other. In this way, a bicluster provides a local association of variables and scalar values in these voxels. Furthermore, the corresponding spatial region of a bicluster can be viewed as a feature of multivariate data.

Based on the concept of biclusters, our co-analysis framework, as shown in Fig. 1, includes bicluster generation, analysis, and visual exploration to guide users in the exploration of various facets of local relationships in multivariate data. We first automatically extract all biclusters from multivariate data based on a biclustering method. Each bicluster contains a local relationship among variables, voxels, and scalar values.

Because the biclustering method is used to generate all possible biclusters, the number of biclusters can be very large, and some of them may overlap with each other. It becomes nearly impossible for users to interactively analyze these biclusters one by one. Therefore, we first hierarchically organize biclusters based on their variable sets, and users can explore biclusters in the context of variable combination. We then design a similarity metric between biclusters to group them to reduce redundancy. This corresponds to the semantic analysis of the three-level analytical tasks of biclusters (Zhao et al., 2018).

For visual analysis of biclusters, we propose a visual analytics system with four coordinated views to interactively analyze local relationships in multivariate data, including the correlation of the variable set, the similarity of biclusters, the correlation of scalar values in biclusters, and the spatial distribution of biclusters. All these views are linked to support the interactive exploration of local relationships in variables, biclusters, and scalar values.

4. Bicluster generation

There are many methods in the category of biclustering/co-clustering, which can be used to generate biclusters. The main difference between them is the clustering strategy. One of the most popular biclustering methods is the variance minimization method (Oghabian et al., 2014), which has been extensively studied under the name of pattern-based clustering. The basic assumption is that an object often exhibits a similar scalar value pattern over several attributes. Therefore, this study employs the variance minimization method to generate biclusters because a local feature/phenomenon in multivariate data may also illustrate similar scalar-value patterns over several variables, i.e., a specific value combination of a subset of variables, including two associated scalar values in two variables (Liu and Shen, 2016).

For example, let us consider ten variables with five voxels in Fig. 2(a). It is obvious that there is no clear pattern among five voxels. However, if we choose two subsets of variables as shown in Figs. 2(b) and 2(c), respectively, the voxels v_1, v_4 and v_5 follow a similar scalar-value pattern, i.e., a coherent pattern, in the variable set {A, C, G, I}, while the voxels v_1, v_2 and v_3 share another coherent pattern in the variable set {B, E, I, J}. The variance minimization method is an effective method for extracting these pattern-based biclusters automatically through the simultaneous analysis of variables and voxels. This can lower the requirement of the domain knowledge and skills of users when performing explorative analysis of multivariate data. Because the phenomena in the data may not be well separated, biclusters do not need to be exclusive. Therefore, this study chooses Maple (Pei et al., 2003), an important algorithm in the variance minimization method as the basis of the co-analysis framework because it can identify overlapped clusters and guarantee completeness of the bicluster search.

We first organize all variables (dimensions) $D = \{d_0, d_1, \ldots, d_{M-1}\}$ and voxels $V = \{v_0, v_1, \ldots, v_{N-1}\}$ in a variable-voxel matrix, where M and N are the number of variables and voxels in multivariate data, respectively. Each entry $s_{i,j}$ in the matrix is the scalar value of the jth variable at the ith voxel. The coherence of two voxels v_x and v_y on two variables d_u and d_v is measured by the pScore (Pei et al., 2003) as follows:

$$pScore\left(\begin{bmatrix} s_{x,u} & s_{x,v} \\ s_{y,u} & s_{y,v} \end{bmatrix}\right) = \left\| (s_{x,u} - s_{x,v}) - (s_{y,u} - s_{y,v}) \right\|. \quad (1)$$

The pScore restricts the coherence to a 2×2 matrix and describes the change in scalar values on two variables between two voxels. Clearly, the smaller the pScore, the greater the coherence of the two voxels on two variables. pScore is more rigorous than other pattern definitions and more robust to noises (Kriegel et al., 2009).

Using the pScore, we can define a bicluster (D', V'), $D' \subset D$ and $V' \subset V$, if the pScore of any two voxels $v_x, v_y \in V'$ on any two variables $d_u, d_v \in D'$ is less than or equal to a user-specified tolerance δ. Most biclusters correspond to specific value combinations of voxels in V' on variables in D' within a tolerance, such as the biclusters in Fig. 2. Scientists are more interested in specific value combinations of multi-variables to obtain depth knowledge about the interaction of variables in simulations. Previous methods usually apply data binning to enforce the tolerance before searching for specific value combinations (Liu and Shen, 2016), while we apply the tolerance during the clustering process to generate more complete results. A bicluster is closed if adding any voxels violates the above definition. Therefore, we must only consider closed biclusters in one variable set, and we refer to closed biclusters as biclusters for simplicity.

A depth-first searching algorithm is first used to find biclusters in lower dimensions, and the biclusters are then merged to derive biclusters of higher dimensions. The variable set is iteratively
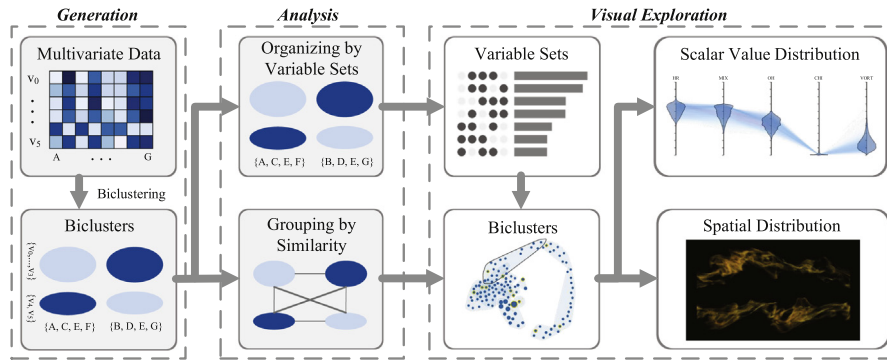
Fig. 1. The co-analysis framework for multivariate data. This example includes eight variables, A–G, and six voxels, v_0-v_5. Our framework first generates all biclusters by simultaneously analyzing variables and voxels. These biclusters are organized by their variable sets and are hierarchically grouped based on a similarity metric in the analysis stage. Four coordinated views are designed to visually explore the local relationships in variables, biclusters, and scalar values.
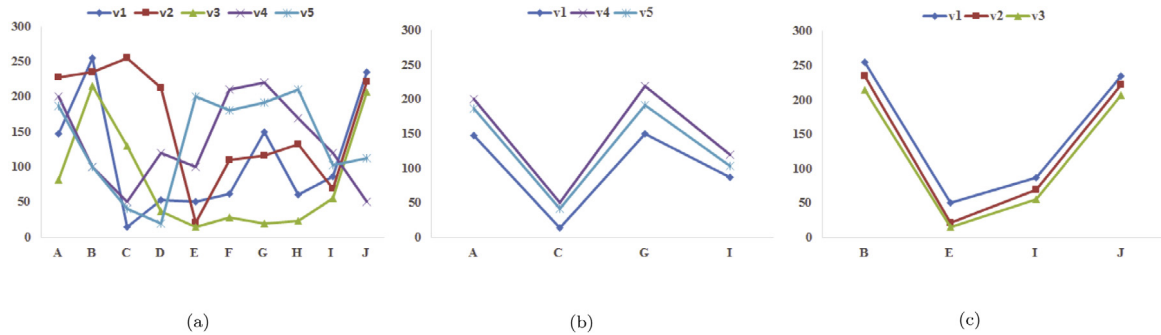

Fig. 2. An illustration of biclusters. (a) A data set with 10 variables and 5 voxels shows no clear patterns. (b) The bicluster ({A, C, G, I}, {v_1, v_4, v_5}) reveals a clear pattern. (c) The bicluster ({B, E, I, J}, {v_1, v_2, v_3} shows another clear pattern.
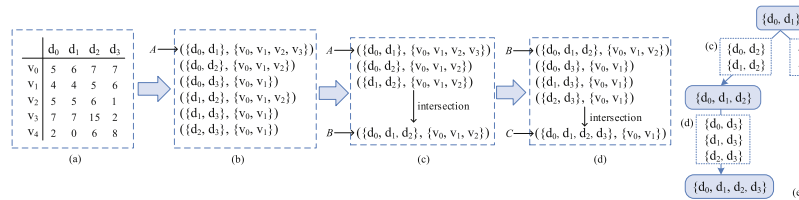

Fig. 3. A running example of bicluster generation. (a) A data matrix with four variables and five voxels. (b–d) The first step is to generate biclusters for all pairs of variables; then the bicluster A is expanded to new biclusters B and C, successively. (e) A variable enumeration tree illustrates the depth-first searching process for the bicluster A. The left subtree corresponds to the searching process (c) and (d), and the searching process of the right subtree is completed after adding the variable d_3, as the variable set {d_0, d_1, d_2, d_3} has been explored in the left subtree.

expanded from two variables to all variables in the manner of multiple trees. If an expanded variable set has been previously explored in the same tree, the searching process can be stopped for this subtree to improve computational efficiency as its bicluster and its children have already been generated.

An example of the searching process with $\delta = 1$ is illustrated in Fig. 3. We systematically enumerate every combination of variables using a variable enumeration tree and a depth-first search. As shown in Fig. 3(b–d), biclusters in two variables are first generated based on the variable-voxel matrix, and then the variable set extends one by one until the number of variables reaches the maximum value. The voxel set of the extended variable set is generated by the intersection of corresponding voxel sets. We repeat this searching process for each bicluster to generate all biclusters using the depth-first searching process illustrated by the variable enumeration tree in Fig. 3(e), and the enumeration procedure is similar to mining frequent closed itemsets.

In multivariate data, the number of voxels is significantly larger than the number of variables. Each variable set includes multiple associated voxel sets, i.e., multiple biclusters with the same variable set. In the exploration of multivariate data, a bicluster may

be statistically insignificant if it contains a small number of voxels, and this can also reduce the searching time of biclusters. Therefore, there are two parameters in the depth-first searching algorithm: δ, the tolerance ($pScore(D', V') \leqslant \delta$), and $minv$, the minimal number of voxels ($|V'| \geqslant minv$).

5. Bicluster analysis

Because the biclustering method guarantees the completeness of the bicluster search, we acquire all biclusters in multivariate data. The generated biclusters are not necessarily exclusive, which means that a voxel/variable can appear in more than one bicluster. Consequently, the number of biclusters is generally very large, and some of them are very similar. To facilitate visual exploration of biclusters, it is necessary to organize and group biclusters to reduce redundancy and encourage diversity during visual exploration.

5.1. Bicluster organization

Each bicluster is associated with one variable set, and one variable set is generally associated with multiple biclusters. Because

the number of variable sets, i.e., the variable combination, is much smaller than the number of biclusters, we first organize biclusters based on their variable sets. The variable sets can be organized hierarchically, and we can iteratively expand the variable set from two variables to multiple variables to reduce the complexity of the bicluster analysis.

While previous methods measure the global correlation of multiple variables (Sukharev et al., 2009; Wang et al., 2011), we can measure the local correlation of multiple variables in a variable set by analyzing its associated biclusters. Because the scalar values of voxels are generally linear between two variables in a bicluster, we choose the Pearson correlation coefficient to evaluate the linear correlation between two variables. As the voxels have a similar scalar-value pattern over these variables in the variable set it would be better to use only associated voxels to measure the local correlation of multiple variables in the variable set instead of all voxels in previous methods. The correlation of multiple variables is the minimal absolute value of the Pearson correlation coefficient of each pair of variables in the variable set D', as follows:

$$C(D') = min\left\{ \left| \frac{cov(d_u, d_v)}{\sigma_{d_u}\sigma_{d_v}} \right| : d_u, d_v \in D' \right\}, \qquad (2)$$

where cov is the covariance, and $\sigma_{d_u}/\sigma_{d_v}$ is the standard deviation of the variable d_u/d_v in the voxels of biclusters of the variable set. The correlations of variable sets can help users choose the variable set to explore first.

5.2. Bicluster grouping

Some of the biclusters may overlap with each other, especially biclusters with the same variable set. Therefore, we hierarchically group biclusters with the same variable set hierarchically to yield a smaller set of mutually sufficiently different, yet individually interesting groups of biclusters for interactive exploration.

Grouping quality primarily depends on the similarity metric between two biclusters. Because biclusters to be grouped have the same variable set, and there is one to one mapping between voxels and scalar values for each variable, the similarity metric must only consider voxels in the bicluster. One promising similarity metric is the spatial overlap because the spatial distribution is a more intuitive way to recognize a feature in volume visualization. If two biclusters have more common voxels, i.e., a large spatial overlap, they are more similar to each other. Therefore, the similarity metric is defined as the Jaccard similarity coefficient as follows:

$$J(A, B) = \frac{|V_A \bigcap V_B|}{|V_A \cup V_B|}, \qquad (3)$$

where V_A and V_B are the voxels of two biclusters A and B, respectively.

Using the similarity metric, the agglomerative hierarchical clustering (Han et al., 2011) is applied to group the biclusters. The distance between two biclusters A and B is defined as $d(A, B) = 1 - J(A, B)$. When combining two groups of biclusters, a weighted average linkage criterion, a recursive definition for the distance between two groups, is used to compute the distance between two groups. For each group, one representative bicluster, such as the one with the largest number of voxels, is selected to guide users in the exploration of large or unfamiliar biclusters in multivariate data. Note that the similarity metric and clustering method is a choice of currently available methods that work for multivariate data. It could be replaced by other similarity metrics that are more suited to the specific requirements of the data.

6. Bicluster exploration

With feature subspaces, including their groups and clusters, we design four coordinated views to visually identify, interpret, and compare the local relationships in multivariate data.
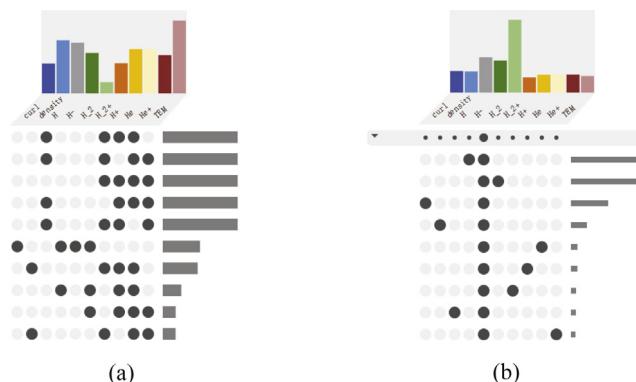

Fig. 4. The association matrix for the ionization front instability data set. (a) The variable sets with at least four variables are sorted by their correlation values. (b) The variable H_2 is drilled down to its children variable sets, and the correlation between H_2 and other variables is shown in the top bar chart.

6.1. Association matrix

The variable sets are inherently hierarchical because of the searching process in bicluster generation. The hierarchical structure of the variable sets is helpful for users to iteratively explore biclusters. We propose an association matrix to display the hierarchical structure of variable sets. The matrix layout is inspired by the combination matrix for the quantitative analysis of sets, their intersections, and aggregates of intersections (Lex et al., 2014). Each column in the association matrix corresponds to a variable of multivariate data, and each row corresponds to a variable set, including the associated biclusters of the variable set. The rows without associated biclusters are hidden by default, but they can be shown on demand during visual exploration. The variable in the variable set is encoded with a filled dark circle, otherwise a light-gray circle, as shown in Fig. 4. Therefores, it is more intuitive to recognize the variables in the variable set in each row, and the names of variables are listed in the top of the matrix. Additional attributes of the variable set can be displayed via the bar chart on the right of each row, and the length of the bar is proportional to the value of the attribute. The matrix layout enables the effective representation of associated data and additional summary statistics, and provides an overview of the relationships between the variable and the bicluster.

As for all matrix-based techniques, sorting is important to ensure an efficient representation of the data. Therefore, we offer various sorting options to analyze the relationship among variables in the local regions and derive a meaningful variable set, i.e., biclusters, for detailed exploration. The sorting attributes primarily include the number of variables in the variable set (the cardinality), the correlation of the variable set, and the number of biclusters in the variable set. We can also use these attributes to filter out less interesting variable sets and reduce the exploration space. For example, the variable sets with at least four variables are sorted by the descending correlation value in Fig. 4(a). The first five variable sets have a high correlation value, and this can guide users to focus on these variable sets first.

We also support drilling down from one variable set to its children variable sets, which is similar to expanding a node in the radial tree to hierarchically explore biclusters. The variables in the expanded variable sets are encoded by smaller dark circles, and other variables are encoded by dark points. The bars associated with expanded rows have a reduced width to distinguish different levels. These children variable sets can be sorted by another attribute for visual comparison. For example, if a user attempts to determine which chemical species are most similar to H_2, he/she

can choose the variable H_2 as the starting variable. As shown in Fig. 4(b), sorted by the descending correlation value, it can be straightforward to obtain the most correlated chemical species H- and H_2+ (the first two rows).

By sorting and drilling down, it becomes easy for users to identify variable sets of interest for detailed analysis by choosing the top rows in the matrix. We additionally provide the correlation information between variables based on information theory. This information is calculated on all voxels, instead of the voxels in biclusters, and they are displayed in the bar chart on the top of the matrix. The bar chart illustrates the entropy of each variable by default and displays mutual information when one variable is selected. Therefore, users can select one variable of interest based on their domain knowledge and drill down to its children variable sets to choose a variable set for detailed exploration.

6.2. Bicluster view

When one variable set is chosen in the association matrix, we must analyze and compare its biclusters, especially the similarity between them. Dimensionality reduction methods have been widely used for similarity analysis in 2D. Therefore, we apply MDS (Guo et al., 2012), one of the widely used dimensionality reduction methods, to project the biclusters of the variable set based on the spatial overlap similarity in the bicluster view. The scatter plot provides an overview of the similarity between biclusters, as shown in Fig. 5. Each cycle is a bicluster, and its size is proportional to the number of voxels in the bicluster.

Because biclusters are clustered hierarchically in Section 5.2, approximately 10 groups are selected to illustrate the meaningful features and local correlations and simplify the interactive exploration. The groups must be coherent enough ($d(A, B) \leqslant 0.99$ in our experiments) and the number of groups must not be too large to avoid visual clutter and to help users in the selection and analysis of biclusters. Each group is encoded by a light-blue and convex region, which covers all biclusters in the group. The representative bicluster of each group is highlighted by orange halos to distinguish between them. Because of the projection error, the regions of groups may be overlapped and result in confusion; i.e., it is unclear as to which group do the biclusters in the overlapped area belong to. Therefore, when hovering with the mouse over the region of one group, its biclusters are highlighted to demonstrate the membership. For a correlated variable set, there could be dozens or even hundreds of biclusters, which would result in visual clutter because of the limited projection space. In this case, only the representative biclusters could be displayed for groups, and other biclusters in one group can be shown on demand. Users can select one group or one bicluster to further explore its scalar-value and spatial distribution to identify local correlations and meaningful features.

As biclusters cannot be grouped perfectly, we allow the manual verification and modification of groups. It is difficult to decide from the bicluster view alone if a bicluster belongs to a group because the scalar-value and spatial distributions are more critical in the similarity analysis of biclusters. Owing to hierarchical clustering, we can select one group by clicking on its region to verify its distribution both in the scalar value and space, and split the group (move the next level) into two groups if it is highly diverse in the distribution. Two groups can be merged into one group if they are very similar. Through these refinements, we can better understand the similarity of biclusters and interactively identify meaningful local correlations.

6.3. Scalar-value view

When one group or bicluster is selected, we employ a parallel coordinate to display its scalar-value distribution over its variables in the scalar-value view, as shown in Fig. 5. The axis of each variable in the variable set is moved to the front, or the axes of other variables are hidden to facilitate the correlation analysis between the scalar values and variables. The parallel coordinate usually draws the polylines on the axes directly, making it hard to interpret the density of scalar values. We enhance the parallel coordinate by counting the occurrences of each scalar-value pair between neighbor axes, and using this information, we encode the transparency of the color. Although the transparency makes the density distribution more observable, it is still not easy to visually compare the density for some cases. Therefore, the histogram of the scalar values is rendered on both sides of the axes to further enhance the parallel coordinate. This eases the analysis of the coherent of the scalar values in one variable and the identification of the range of scalar values that includes most voxels.

For a group, the parallel coordinate can be used to verify the similarity of biclusters in the group. If the scalar-value distributions on the axes are all within a small range, biclusters are similar to each other in the group. Otherwise, this group may be split to generate coherent groups. For a bicluster, the parallel coordinate can be used to present the coherent scalar-value pattern and to analyze the manner in which the specific values of the variables interact with each other.

6.4. Spatial view

In addition to the scalar-value distribution of one group or bicluster, the spatial distribution is important for local correlation analysis. The probability of the voxels belonging to a group or bicluster is calculated. The probability volume is visualized by direct volume rendering to display a spatial distribution and analyze the spatial coherence of a group or bicluster.

6.5. Exploration guideline

Our co-analysis framework provides an analysis guideline that enables users to explore various aspects of multivariate data as an overview or in detail, as illustrated in Fig. 5. Given a multivariate data set, the variable set of interest can be obtained by sorting according to application-related attributes, such as the correlation of the variable set and the number of biclusters, or the mutual information between variables in the bar chart. The variable set can be drilled down to identify the variable set most related to a feature/phenomena in the association matrix. Using the selected variable set, the clusters and their biclusters are illustrated in the scatter plot, presenting an overview of their similarities to users. Users can visually explore each cluster or representative bicluster, and analyze the association of scalar values in related variables and the spatial distribution in the spatial view. The clusters can be interactively refined based on the similarity analysis of biclusters. These steps are iteratively performed to verify the local relationships among variables, features, and scalar values.

7. Results

In this section, three representative multivariate data sets in different domains were used to verify the effectiveness and usefulness of our framework in analyzing the local relationships among variables, biclusters, and scalar values. We performed all experiments on an Intel Core i7-7700K 4.20 GHz CPU equipped with an NVIDIA GeForce GTX 1070 GPU. The minimum number of voxels

for biclusters was set at 0.2% of the total voxels of the explored volume to capture small features, such as the hurricane eye. In most simulations, biclusters corresponding to the background generally have a large number of voxels, and we filtered these less interesting biclusters based on the number of voxels (10% of the total voxels).

7.1. Hurricane Isabel data set

The hurricane Isabel data set has been widely used in previous research, and it presents a simulation of a hurricane created by the National Center for Atmospheric Research in the U.S. Ten variables were used in our experiment: PRE, PRECIP, QCLOUD, QGRAUP, QICE, QSNOW, QVAPOR, TC, and VEL, which is the magnitude of the wind speed. The resolution is $250 \times 250 \times 50$, and the 20th time step is chosen in our experiment to explore local relationships and classify the main features of the hurricane, i.e., the hurricane eye and the rainbands. The tolerance δ for generating the biclusters is 20 with coherent scalar-value patterns.

Because we are more interested in local relationships with at least three variables, we first filtered the variable set with only two variables in the association matrix. We sorted the variable sets according to the number of biclusters, which demonstrates the correlation of variables in terms of biclusters. As shown in Fig. 5, the variable set {PRE, QVAPOR, TC} includes the most number of biclusters, and its biclusters are projected on the scatter plot to illustrate the similarity between them. Group C on the right is far away from others, and the spatial view domonstrates that it is the lower part of the hurricane eye. When group A on the left is selected, the upper part of the hurricane eye is presented in the spatial view. Based on these results, we can assume that the groups in the middle of the scatter plot may correspond to the middle part of the hurricane eye. Several groups are overlapped in the middle, and they represent nearly the same feature. To verify the assumption, we interactively merged these groups into one group B and obtained the middle part of the hurricane eye. The scalar values of the three groups are displayed in the scalar-value views at the top of Fig. 5.

The second variable set in the association matrix is {PRE, QVAPOR, VEL}. The group on the right indicates the rainbands around the hurricane eye. When we further explored each bicluster in this group, there are three biclusters that correspond to three different rainbands near the hurricane eye. As shown in Fig. 5, the rainbands are gradually away from the hurricane eye, and the scalar values of the pressure are almost the same. However, the water vapor mixing ratio is gradually increased, and the wind speed is gradually decreased. This also agrees with the knowledge of the rainbands around the hurricane eye.

Based on our co-analysis framework, we can quickly identify the variable set related to a local feature/phenomenon in multivariate data. By analyzing biclusters or their groups in the scalar-value view and spatial view, we can find that the variable set {PRE, QVAPOR, TC} is locally associated in the region of the hurricane eye, while the variable set {PRE, QVAPOR, VEL} is more useful in recognizing the rainbands around the hurricane eye.

7.2. Turbulent combustion data set

This data set includes five variables: Heat Release Rate (HR), Mass Fraction of the Hydroxyl Radical (OH), Mixture Fraction (MIX), Scalar Dissipation Rate (CHI), and vorticity (VORT).

We first sorted the variable sets with at least three variables according to correlation of the variable set in the association matrix and selected the first variable set {HR, MIX, OH} to explore its biclusters as shown in Fig. 6(a). It is easy to identify four groups of biclusters that correspond to four parts of the flame in Fig. 6(b), i.e., the outer layer of the flame, the body of the flame, the inner layer of the flame, and the non-combustion region. The spatial distributions and scalar values of the four groups are illustrated in Fig. 6(c–f). It is evident that HR is high in the outer and inner layer of the flame and the non-combustion region, but OH is low, especially in the non-combustion region (nearly zero).

As shown in Fig. 6(a), the correlation value of the first variable set {HR, MIX, OH} is higher than the value of the second variable set {HR, MIX, VORT}. If we measure the correlation between variables based on all voxels (Biswas et al., 2013), instead of the voxels in biclusters, the most correlated variables with an informative variable MIX (the largest entropy) are OH and VORT, i.e., the third variable set. After exploring the third variable set, we find that its biclusters are less interested compared to the first variable set. Therefore, the correlation among variables can be better measured in the associated local regions, i.e., the regions of biclusters.

7.3. Deep water impact data set

This data set has been generated by a 3D simulation of a 250-meter-diameter asteroid impacting into ocean after passing through the atmosphere at an angle of $45°$ in Los Alamos National Laboratory (Patchett and Ahrens, 2018). Six variables were used for the experiment: pressure (prs), density in grams (rho), sound speed (snd), temperature (tev), volume fraction water (v02), and velocity, which is the magnitude of the wind speed.

Domain experts are interested in the effects of the phenomena on natural disasters such as rainfall. Rainfall is related to v02, which represents a fraction of water in the air or water vapor. Therefore, we selected the variable v02 as the starting variable to drill down to its children and further sorted them according to the number of biclusters. As shown in Fig. 7(b), tev and snd are primarily associated with v02. Alternatively, we can also sort the variable sets with at least three variables according to the number of biclusters as shown in Fig. 7(a). The first variable set {snd, tev, v02} is also the one with the most number of biclusters, i.e., more local relationships. The biclusters of the variables set {snd, tev, v02} are projected on the scatter plot in Fig. 7(c). There are several discernible groups, such as three distinguished groups A, B, C, and other groups have less interesting or coherent features. The spatial and scalar-value distributions of the three groups are displayed in Fig. 7(d).

The region with a high temperature in group A is primarily distributed around the asteroid's trajectory. The gravitational potential energy of the asteroid is converted into kinetic energy and the energy for overcoming air resistance. The energy for overcoming air resistance then turns into heat energy, increasing the temperature near the asteroid's trajectory. For group B, it is easy to identify two regions with a high volume fraction of water (v02). One region is above the sea level impacted by the asteroid, and the other is the evacuated channel left by the asteroid's trajectory. For the former, the speed of the asteroid is reduced after impacting into the water, which causes the surrounding water to splash around and leads to an increase in the volume fraction of water above the impact position. A tsunami may occur when the impact is strong enough. For the latter, because of the high-temperature around the asteroid's trajectory, vast amounts of liquid water absorb heat and then change into water vapor, and the water molecules move and spread along the high temperature region. When there is enough water and sufficient suspended particles in a colder stratum, the water condenses together and produces rains if the water's gravity is higher than buoyancy. In addition, H_2O, a greenhouse gas, can absorb the reflected solar radiation of the Earth's surface, so that it may increase the temperature around the area to some extent. Therefore, we conclude that there would be local rainfall with slight warming after the asteroid impacts into the ocean.

We compare our co-analysis framework with gradient similarity measure (GSIM) (Sauber et al., 2006) using the variable set {snd,
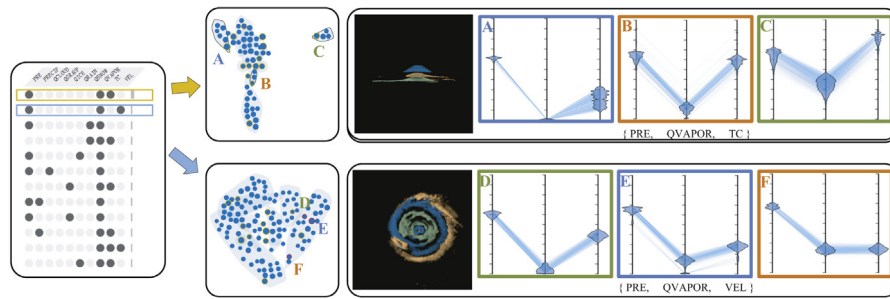
Fig. 5. Visual exploration of the hurricane Isabel data set. Three groups in the first variable set labeled *A*, *B*, and *C* correspond to the upper (blue), middle (yellow), and lower (green) part of the hurricane eye. Their coherent scalar-value distributions are shown on the right of the spatial view. Three biclusters are selected in the second variable set {PRE, QVAPOR, VEL}. They are labeled *D, E* and *F* with red halos, which are the rainbands around the hurricane eye with an increasing distance (green–blue–yellow).
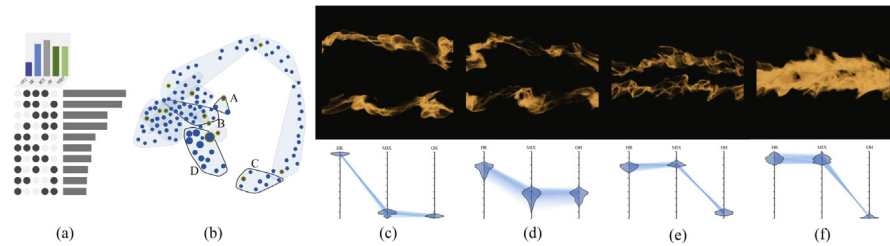

Fig. 6. Visual exploration of the variable set {HR, MIX, OH} in the turbulent combustion data set with 65th time step and 240 × 360 × 60 resolution. (a) The association matrix is sorted by the correlation value. (b) The biclusters of the variable set {HR, MIX, OH}. (c–f) The spatial and scalar-value distributions of groups *A, B, C* and *D*.
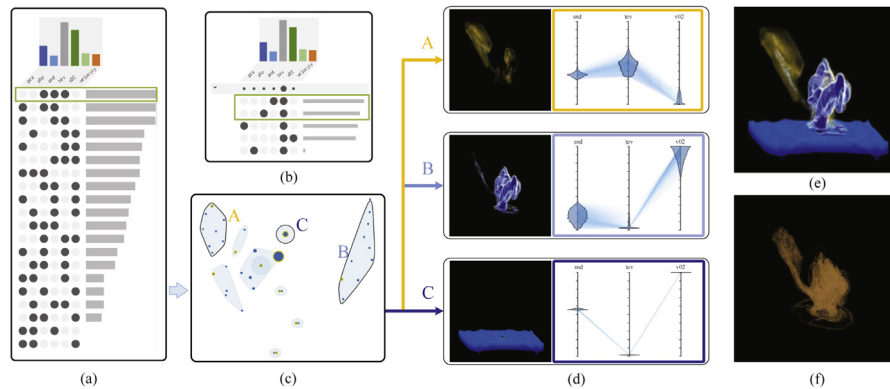

Fig. 7. Visual exploration of the local relationships in the deep water impact data set with 27th time step and 150 × 150 × 150 resolution. (a–b) The association matrix is sorted according to the number of biclusters and the variable set {snd, tev, v02} is on the top of the list. (c) Corresponding biclusters. (d) The spatial and scalar-value distributions of groups *A* (a high temperature, high sound speed, and low water fraction in the air), *B* (a low temperature, low sound speed, and high water fraction in the air), and *C* (a low temperature, high sound speed, and high water fraction in the water). (e) The local features in groups *A*, *B* and *C*. (f) The result of GSIM (Sauber et al., 2006) for the correlation of the variable set {snd, tev, v02}.

tev, v02}. Fig. 7(e) presents the overall spatial distribution of groups *A, B,* and *C*. GSIM can measure the correlation of multiple variables by calculating the similarity among gradients at each voxel, and the result is presented in Fig. 7(f). Overall, the spatial distributions are similar. However, our framework can effectively extract local features with similar scalar-value patterns, i.e., local relationships between variables and voxels, and each local feature includes a specific value combination revealing the local interaction of variables. In contrast, the result of GSIM is a global feature for the three variables, and it is challenging to gain insight into the local associations and their scalar-value distributions such as regions with high water vapor or low temperature.

7.4. Discussion

Our co-analysis framework provides a new perspective for systematically and visually exploring multivariate data. Three experiments demonstrate that our co-analysis framework can help users quickly explore variable sets of interest and discover the local correlations of the scalar values among different variables.

Compared to previous methods in correlation analysis and multi-dimensional transfer functions, our framework clusters variables and voxels *simultaneously* to extract all biclusters with a similar scalar-value pattern *automatically*, and focuses on analyzing the *local* relationships among variables, biclusters, and scalar values. In particular, our framework extends the analysis of value combinations of two variables (Liu and Shen, 2016) to multiple variables such as three variables in our experiments. Although our system supports visual exploration of biclusters of all variables, we focus on the local correlations of 2–4 variables because the feature/phenomenon is generally associated with a subset of variables. Further, there have been many previous methods that use all variables to cluster features (Wu et al., 2015). As shown in Fig. 7, our framework can effectively identify local features with similar scalar-value patterns from multiple variables, which is complementary to previous global correlation analysis (Sauber et al.,

Table 1

The computational time (seconds) of bicluster generation and the number of biclusters for three data sets.

δ	Isabel		Combustion		Asteroid	
	Time	Number	Time	Number	Time	Number
10	4924	6159	451	904	131	985
15	2423	4352	460	1024	72	810
20	1070	3040	472	923	64	756
25	996	2370	482	923	60	636
30	526	1896	212	650	40	539

2006). Compared to interactive classification (Guo et al., 2011), our framework automatically generates all biclusters, groups biclusters to facilitate the exploration of biclusters, and designs coordinated views to identify variable sets of interest without much prior knowledge and to efficiently discover local correlations among variables.

Biclusters are generated in the preprocessing stage before the exploration process. Table 1 present the computational time of bicluster generation and the number of biclusters for three experimented data sets with different tolerances. The computational time ranges from less than 1 min to more than 1 h, and is roughly proportional to the number of biclusters, which depends on the number of variables and the complexity of the volume. The number of biclusters can be very large, especially for 10 variables such as the hurricane data. During visual exploration, users can filter small or large biclusters based on the number of voxels according to their analysis requirements, including the background feature. Because we assume that users have little prior knowledge about the data, biclusters in each combination of variables are extracted from the data. If users have prior knowledge, only variables of interest need to search for biclusters and this would greatly improve computational and analytical efficiency. In addition, the searching process can be accelerated by parallel computation because the expansion of each variable pair is independent.

The biclustering method includes two parameters, the tolerance and the minimal number of voxels. The minimal number of voxels can be set to a very small number, even zero, to determine all biclusters. However, this would increase the computational time and generate many noisy and meaningless biclusters with only a few voxels. In our experiments, the minimal number of voxels is fixed to 0.2% of the total voxels of the explored volume, a relatively low value, to balance these factors. In this case, the resolution of multivariable data has little impact on the number of biclusters, although small biclusters may not be extracted when the resolution is low. The tolerance is related to the data range and meaning of the variables, and determines the maximum difference of scalar values to be considered as the same feature. As shown in Table 1, the number of biclusters decreases with an increase in the tolerance because many small biclusters are merged into large biclusters. In order to restrain the over-segmentation problem, the tolerance can be chosen from 10 to 30 depending on the application (20 by default in our experiments).

A potential limitation of biclustering is that it would generate too many biclusters from multivariate data, which results in a long analysis process. Our framework can group biclusters to reduce the number of biclusters to be analyzed and supports filtering based on the number of voxels for removing too small or too large biclusters, including noisy and background features. As demonstrated in the experiments, the association matrix can sort the variable sets based on the number of biclusters and the correlation of the variable set for exploring interesting variable sets and biclusters. In the future, this issue can be further addressed by recommending meaningful groups or biclusters in different variable sets by automatically evaluating the coherence of scalar values and the spatial distribution.

In addition, the variables of the bicluster with a similar scalar-value pattern may be mathematically correlated, but they may not be correlated in the physical phenomena. Therefore, this requires users to further verify such correlation in the bicluster with domain knowledge and human intelligence during visual exploration.

8. Conclusion

In this paper, we proposed a co-analysis framework to guide the visual exploration of the local correlations in multivariate data based on biclusters. The biclustering method is used to automatically generate all biclusters only containing voxels with a similar scalar-value pattern over multiple variables. They are organized and grouped hierarchically to reduce the complexity of user interaction and are visually presented in four coordinated views to facilitate interactive exploration of multivariate data from different facets of multivariate data. Experiments demonstrated that our co-analysis framework can effectively identify the associated variable set related to a local feature/phenomenon, compare the similarity of biclusters, and analyze the correlations of the scalar values of different variables in local regions.

For future work, we plan to recommend meaningful groups or biclusters in different variable sets to further improve analytical efficiency and employ parallel computation to accelerate computational efficiency in bicluster generation. We would also like to extend our co-analysis framework to time-varying multivariate data for capturing the coherence in the time space.

References

Biswas, A., Dutta, S., Shen, H.W., Woodring, J., 2013. An information-aware framework for exploring multivariate data sets. IEEE Trans. Vis. Comput. Graphics 19 (12), 2683–2692.

Carr, H., Duke, D., 2013. Joint contour nets: Computation and properties. In: Proceedings of IEEE Pacific Visualization Symposium (PacificVis) 2013. pp. 161–168.

Cox, T.F., Cox, M., 1994. Multidimensional Scaling, second ed. Chapman Hall, London.

Fuchs, R., Hauser, H., 2009. Visualization of multi-variate scientific data. Comput. Graph. Forum 28 (6), 1670–1690.

Gosink, L., Anderson, J., Bethel, W., Joy, K., 2007. Variable interactions in query-driven visualization. IEEE Trans. Vis. Comput. Graphics 13 (6), 1400–1407.

Guo, H., Xiao, H., Yuan, X., 2011. Multi-dimensional transfer function design based on flexible dimension projection embedded in parallel coordinates. In: Proceedings of IEEE Pacific Visualization Symposium (PacificVis) 2011. pp. 19–26.

Guo, H., Xiao, H., Yuan, X., 2012. Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates. IEEE Trans. Vis. Comput. Graph. 18 (9), 1397–1410.

Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques. Elsevier.

Hartigan, J.A., 1972. Direct clustering of a data matrix. J. Amer. Statist. Assoc. 67 (337), 123–129.

He, X., Tao, Y., Wang, Q., Lin, H., 2018a. Biclusters based visual exploration of multivariate scientific data. In: Proceedings of IEEE Scientific Visualization Conference (SciVis) 2018. pp. 40–45.

He, X., Tao, Y., Wang, Q., Lin, H., 2018b. Multivariate spatial data visualization: a survey. J. Vis. (ChinaVis 2018).

Inselberg, A., 1985. The plane with parallel coordinates. Vis. Comput. 1 (2), 69–91.

Inselberg, A., Dimsdale, B., 1990. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the First IEEE Conference on Visualization 1990. pp. 361–378.

Jänicke, H., Wiebel, A., Scheuermann, G., Kollmann, W., 2007. Multifield visualization using local statistical complexity. IEEE Trans. Vis. Comput. Graphics 13 (6), 1384–1391.

Kehrer, J., Hauser, H., 2013. Visualization and visual analysis of multifaceted scientific data: A survey. IEEE Trans. Vis. Comput. Graphics 19 (3), 495–513.

Kriegel, H.-P., Kröger, P., Zimek, A., 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data 3 (1), 1:1–1:58.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of intersecting sets. IEEE Trans. Vis. Comput. Graphics 20 (12), 1983–1992.

Liu, X., Shen, H.-W., 2016. Association analysis for visual exploration of multivariate scientific data sets. IEEE Trans. Vis. Comput. Graphics 22 (1), 955–964.

Liu, M., Shi, J., Li, Z., Li, C., Zhu, J., Liu, S., 2017. Towards better analysis of deep convolutional neural networks. IEEE Trans. Vis. Comput. Graphics 23 (1), 91–100.

Liu, S., Wang, B., Thiagarajan, J.J., Bremer, P.-T., Pascucci, V., 2015. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. Comput. Graph. Forum 34 (3), 271–280.

Lu, K., Shen, H.W., 2017. Multivariate volumetric data analysis and visualization through bottom-up subspace exploration. In: IEEE Pacific Visualization Symposium (PacificVis) 2017. pp. 141–150.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Nagaraj, S., Natarajan, V., Nanjundiah, R.S., 2011. A gradient-based comparison measure for visual analysis of multifield data. Comput. Graph. Forum 30 (3), 1101–1110.

Oghabian, A., Kilpinen, S., Hautaniemi, S., Czeizler, E., 2014. Biclustering methods: biological relevance and application in gene expression analysis. PLoS One 9 (3), e90801.

Patchett, J., Ahrens, J., 2018. Optimizing scientist time through in situ visualization and analysis. IEEE Comput. Graph. Appl. 38 (1), 119–127.

Pei, J., Zhang, X., Cho, M., Wang, H., Yu, P.S., 2003. Maple: a fast algorithm for maximal pattern-based clustering. In: Proceedings of Third IEEE International Conference on Data Mining (ICDM) 2003. pp. 259–266.

Santamaría, R., Therón, R., Quintales, L., 2008. Bicoverlapper: a tool for bicluster visualization. Bioinformatics 24 (9), 1212–1213.

Sauber, N., Theisel, H., Seidel, H.P., 2006. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. IEEE Trans. Vis. Comput. Graphics 12 (5), 917–924.

Sukharev, J., Wang, C., Ma, K.L., Wittenberg, A.T., 2009. Correlation study of time-varying multivariate climate data sets. In: Proceedings of IEEE Pacific Visualization Symposium (PacificVis) 2009. pp. 161–168.

Sun, M., Mi, P., North, C., Ramakrishnan, N., 2016. BiSet: Semantic edge bundling with biclusters for sensemaking. IEEE Trans. Vis. Comput. Graph. 22 (1), 310–319.

Tzeng, F.-Y., Ma, K.-L., 2004. A cluster-space visual interface for arbitrary dimensional classification of volume data. In: Proceedings of the Sixth Joint Eurographics - IEEE TCVG Conference on Visualization. pp. 17–24.

Van Long, T., Linsen, L., 2009. Multiclustertree: Interactive visual exploration of hierarchical clusters in multidimensional multivariate data. Comput. Graph. Forum 28 (3), 823–830.

Wang, C., Yu, H., Grout, R.W., Ma, K.L., Chen, J.H., 2011. Analyzing information transfer in time-varying multivariate data. In: Proceedings of IEEE Pacific Visualization Symposium (PacificVis) 2011. pp. 99–106.

Wu, F., Chen, G., Huang, J., Tao, Y., Chen, W., 2015. Easyxplorer: A flexible visual exploration approach for multivariate spatial data. Comput. Graph. Forum 34 (7), 163–172.

Zhao, X., Kaufman, A., 2010. Multi-dimensional reduction and transfer function design using parallel coordinates. In: Proceedings of the 8th IEEE/EG International Conference on Volume Graphics, VG'10. pp. 69–76.

Zhao, J., Sun, M., Chen, F., Chiu, P., 2018. BiDots: Visual exploration of weighted biclusters. IEEE Trans. Vis. Comput. Graphics 24 (1), 195–204.