Biclusters Based Visual Exploration of Multivariate Scientific Data

Xiangyang He* Yubo Tao† Qirui Wang‡ Hai Lin§

State Key Laboratory of CAD&CG, Zhejiang University

ABSTRACT

This paper proposes a co-analysis framework based on biclusters, i.e., two subsets of variables and voxels with close scalar-value relationships, to guide the visual exploration process of 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 further grouped by a similarity metric to reduce redundancy and encourage 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 demonstrate the effectiveness of our framework in exploring local relationships among variables, biclusters and scalar values in the data.

Keywords: Multivariate data, bicluster, local association

1 INTRODUCTION

Scientific simulations often generate data sets with multiple variables for complex physical phenomena. These variables generally have hidden associations, since they work collectively in the simulation [3]. For instance, a hurricane is a rapidly rotating storm system characterized by a low-pressure center, strong winds, and along with heavy rain in climate simulation. However, the heterogeneity of multivariate data make it difficult to extract interesting associations, which are typically located only in the subspaces of variables and subsets of voxels. For example, the eyewall clouds may be strongly associated with the water vapor and cloud moisture [8]. Thus, it would be better to extract hidden associations between variables locally and detect local features based on associated variables.

In multivariate data analysis, a broad variety of techniques have been proposed to explore the global relationships between variables [14, 15, 18] and voxels [16, 17, 19] in the data. These methods take all variables or voxels into account. Therefore, it may be difficult to detect features that only depend on a subset of variables, as other unrelated variables may have a negative impact on clustering due to the curse of dimensionality. Since a subset of variables may be strongly associated in a local region, it is desirable to extract these local associations among variables or the similarity of voxels in different local regions rather than analyzing the global associations. To obtain local associations between variables and voxels, they should be analyzed together rather than separately in previous methods. On the other hand, multi-dimensional transfer functions can take both variables and voxels into account to manually classify features of interest. For example, a feature can be specified by gradually selecting scalar value intervals of a few variables in the parallel coordinate [5,20] or specifying Gaussian functions in a scatter plot matrix [9]. In this case, these variables may be associated, and their correlated scalar value intervals provide the definition of the feature, the voxels of which have a similar scalar-value pattern over these associated variables. We name it a bicluster between variables and voxels, that is, two subsets of variables and voxels with close scalar-value relationships. While a manual specification provides flexibility for finding a variety of biclusters, it can be laborious and hinder comprehensive coverage of the data in exploratory analysis. When there are many variables in multivariate data, it becomes nearly impossible to find all meaningful features due to the large size of the search space. This brings a need to find all meaningful biclusters between variables and voxels automatically.

To address these requirements, this paper proposes a co-analysis framework based on biclusters for exploring multivariate data. Our framework first generates all biclusters by clustering variables and voxels simultaneously. Then these biclusters can be organized and grouped by variable sets for hierarchically exploring variables and biclusters based on a similarity metric. In order to visually explore biclusters, we design a visual analysis system to reveal three-faceted relationships: variables, biclusters and scalar values.

2 RELATED WORK

Many correlation analysis methods have been proposed to find hidden correlations in multivariate data and explore the relationships between variables and scalar values over the years. Information theory provides a theoretical framework to measure the global correlation between variables. Biswas et al. [2] employed the mutual information to measure the informativeness of one variable about the other variable and grouped variables based on the mutual information in a graph-based approach. In addition, information theory also can be extended to time-varying fields. For example, Dutta et al. [4] extracted important features using mutual information and its two decompositions in time-varying multivariate data, and multiple time-varying features were encoded into a field to analyze the track and characteristic of these features.

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 correlation values of all voxels. Sauber et al. [14] proposed a gradient similarity measure (GSIM) and a local correlation coefficient to measure the local correlation at each voxel, and introduced the multifield-graph to show an overview of the correlation between variables. Nagaraj et al. [10] presented a gradient-based correlation criterion, the norm of a partial derivative matrix, to capture interactions between multiple scalar fields, and the correlation field can be visualized to detect regions with high correlation values. In this paper, we cluster variables and voxels simultaneously to extract biclusters automatically and employ biclusters, subsets of voxels instead of all voxels in previous methods, to better analyze the features in local regions.

For interactive feature classfication, Guo et al. [5] proposed a novel transfer function design interface combining the parallel coordinate and MDS plots to facilitate feature specification in multivariate data. Lu and Shen [9] presented a bottom-up subspace exploration workflow to allow users to design multivariate transfer function interactively, and introduced additional information to

*E-mail: xiangyanghe@zju.edu.cn
†E-mail: taoyubo@cad.zju.edu.cn (Corresponding author)
‡E-mail: qiruiw@gmail.com
§E-mail: lin@cad.zju.edu.cn

2018 IEEE Scientific Visualization Conference (SciVis)
21-26 October 2018, Berlin, Germany
978-1-5386-6882-5/18/$31.00 ©2018 IEEE

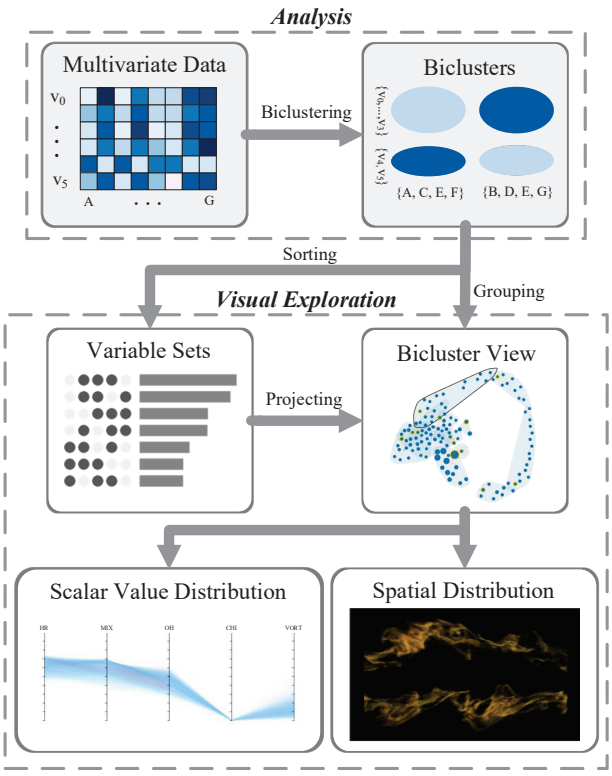

Figure 1: The co-analysis framework. This example contains eight variables, A-G, and six voxels, v_0-v_5. Our framework first generates all biclusters by analyzing variables and voxels simultaneously. These biclusters are grouped hierarchically based on a similarity metric in the analysis stage. Four coordinated views are designed to visually explore the local relationships.

guide users to choose subspaces and discover interesting features. While it is flexible to interactively specify features, it can be time-consuming and challenging to search for all meaningful features. In this paper, we extract all meaningful biclusters automatically, 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 Our Framework

As shown in Fig. 1, our co-analysis framework is based on biclusters to explore the local relationships, since each bicluster contains a local relationship among variables, voxels, and scalar values.

3.1 Bicluster Analysis

A local feature/phenomenon in multivariate data may have a similar scalar-value pattern over several variables, i.e., a bicluster. A bicluster is composed of a subspace of variables and a subset of voxels, and these voxels have a similar scalar-value pattern on these variables, which provides a local association of variables and scalar values in the voxels. In the data mining field, the biclustering method can effectively extract cohesive objects with a similar scalar-value pattern over a subset of attributes.

The variance minimization method [11] is effective in extracting the pattern-based biclusters automatically by analyzing variables and voxels simultaneously. In this paper, we use MaPle [13], one important algorithm in the variance minimization method, as the basis of our co-analysis framework to generate biclusters. These biclusters provides specific value combinations of several variables, and can be used to analyze the interaction of variables in the simulations.

One parameter of MaPle is the number of voxels of a bicluster. A bicluster may be statistically insignificant if it contains a small number of voxels, and this can reduce the searching time of biclusters. The minimal number of voxels for biclusters is specified as 0.2% of the total voxels of the explored volume to capture small features. In most simulations, biclusters corresponding to the background generally have a large number of voxels, and we filtered these less interesting biclusters by the number of voxels (10% of the total voxels) to improve the efficiency of the co-analysis framework.

Since the biclustering method guarantees the completeness of the bicluster search, we acquire all biclusters in multivariate data. Each bicluster is associated with one variable set, and one variable set is generally associated with multiple biclusters. Thus, we first hierarchically organize biclusters based on their variable sets and iteratively expand the variable set from two variables to multiple variables to reduce the complexity of bicluster analysis. In addition, some of the biclusters may overlap with each other, especially biclusters with the same variable set, as a voxel/variable can appear in more than one bicluster. To facilitate visual exploration of biclusters, we then 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. The grouping quality mainly depends on the similarity metric between two biclusters. In this paper, we use the spatial overlap as our similarity metric between two biclusters. If two biclusters have a large spatial overlap, they are more similar to each other. The similarity metric is defined as the Jaccard similarity coefficient:

$$J(A,B) = \frac{|V_A \bigcap V_B|}{|V_A \cup V_B|}, \quad (1)$$

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

With the similarity metric, the agglomerative hierarchical clustering [7] is applied to group 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, is used to compute the distance.

3.2 Bicluster Exploration

We propose a visual analytics systems to assist users in interactively exploring biclusters including the association matrix, the bicluster view, the scalar-value view and the spatial view.

Association matrix. We propose an association matrix to display the hierarchical structure of variable sets. Each column in the association matrix corresponds to a variable of multivariate data, and each row corresponds to a 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. 2(a). Additional attributes of the variable set could be displayed and sorted via the bar chart on the right of each row, and the length of the bar is proportional to the value of the attribute, which can guide users to choose interesting variable sets. The sorting attributes mainly contain the number of biclusters and the correlation of the variable set, i.e., the minimal absolute value of Pearson correlation coefficient [15]. We also support drilling down from one variable set to it children variable sets to explore biclusters hierarchically. As shown in Fig. 2(b), 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,

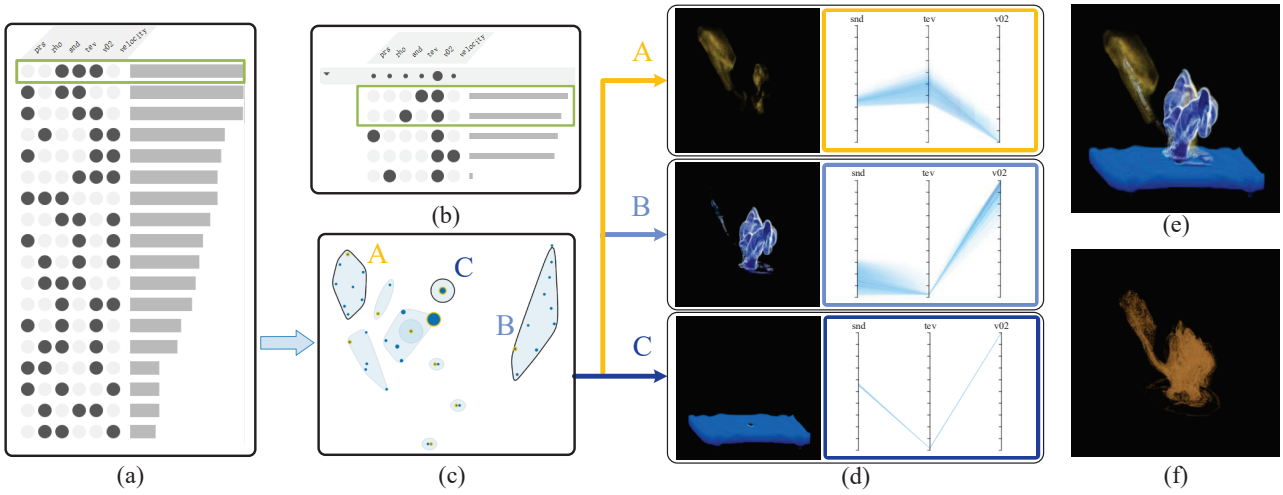

Figure 2: Visual exploration of the local relationships in the deep water impact data set with 27th time step and $150 \times 150 \times 150$ resolution. (a-b) The association matrix is sorted by the number of biclusters and the variable set {snd, tev, v02} is on the top of the list. (c) Corresponding biclusters. (d) The spatial and scalar-value distributions of the 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 the groups *A*, *B* and *C*. (f) The result of GSIM [14] for the correlation of the variable set {snd, tev, v02}.

and these children variable sets can be sorted by another attribute for visual comparison.

Bicluster view. When one variable set is chosen in the association matrix, we need to analyze and compare its biclusters, especially the similarity between them. We apply MDS [6], 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. 1. Each cycle is a bicluster, and its size is proportional to the number of voxels in the bicluster. Each group is encoded by a light-blue and convex region, which covers all biclusters in the group. The representative bicluster of each group, the one with the largest number of voxels, is highlighted by orange halos to distinguish different groups. Due to the projection error, the regions of groups may be overlapped and result in the confusion. Thus, when hovering with the mouse over the region of one group, its biclusters are highlighted to show the membership. Users could select one group or one bicluster by clicking on corresponding region to verify its distribution both in the scalar value and space. Through these refinements, we can better understand the similarity of biclusters and identify meaningful local correlations interactively.

Scalar-value view. When one group or bicluster is selected, we employ the parallel coordinate to display its scalar-value distribution over its variables in the scalar-value view to better analyze the correlation between numerical values, as shown in Fig. 1. 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, as shown in Fig. 2(d).

Spatial view. Besides the scalar-value distribution of one group or bicluster, the spatial distribution is also important for the 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 the spatial distribution.

4 Results

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

4.1 Deep Water Impact Data Set

Six variables of the deep water impact data set [12] were used for the experiment: pressure (prs), density in grams (rho), sound speed (snd), temperature (tev), volume fraction water (v02), and velocity.

Domain experts are interested in the effects of the phenomena on natural disasters, such as the rainfall. The rainfall is related to v02, i.e., the fraction of water in the air or water vapor. Thus, we selected the variable v02 as the starting variable to drill down to its children and further sorted them by the number of biclusters. As shown in Fig. 2(b), tev and snd are most associated with v02. Alternatively, we can also sort the variable sets with at least three variables by the number of biclusters as shown in Fig. 2(a). The first variable set is also {snd, tev, v02} 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. 2(c). There are several discernible groups, such as three distinguished groups *A*, *B*, *C*, and other groups have less interesting or coherent features. Fig. 2(d) shows the spatial and scalar-value distributions of the three groups.

The region with a high temperature in the group *A* is mainly distributed around the asteroid's trajectory. The gravitational potential energy of the asteroid is converted into the kinetic energy and the energy to overcome air resistance. Then the energy overcoming air resistance turns into the heat energy, increasing the temperature near the asteroid's trajectory. For the group *B*, it is easy to identify two regions with a high volume fraction of water (v02). One 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 reduced after impacting into the water, which causes the surrounding water to splash around and leads to an increase of the volume fraction of water above the impact position. A tsunami may occur when the impact is strong enough. For the latter, due to the high-temperature around the asteroid's trajectory, vast amounts of liquid water change into water vapor, and the water molecules move and spread along the high temperature region. When there are 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 its buoyancy. In addition, H_2O, a greenhouse

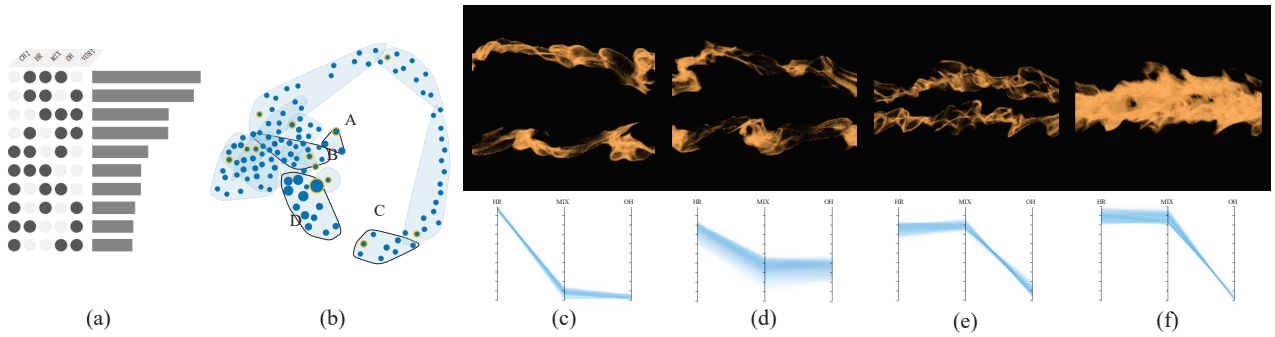

Figure 3: Visual exploration of the variable set {HR, MIX, OH} in the turbulent combustion data set with 65th time step and $240 \times 360 \times 60$ resolution. (a) The association matrix is sorted by the correlation value. (b) The biclusters of the variable set {HR, MIX, OH}. (c-f) The spatial and scalar-value distributions of the groups *A, B, C* and *D*.

gas, can absorb the reflected solar radiation of the Earth's surface, which may increase the temperature around the area to some extent. Therefore, we conclude that there would be a local rainfall with a slight warming after the asteroid impacting into the ocean.

We compare our co-analysis framework with gradient similarity measure (GSIM) [14] by the variable set {snd, tev, v02}. Fig. 2(e) shows the overall spatial distribution of the 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 displayed in Fig. 2(f). Overall, the spatial distributions are similar. However, our framework can effectively extract local features with a similar scalar-value pattern, i.e., local relationships between variables and voxels, and each local feature has 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 fails to gain insights into the local associations and their scalar-value distributions.

4.2 Turbulent Combustion Data Set

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

We sorted the variable sets with at least three variables by the 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. 3(a). It is easy to identify four groups of biclusters corresponding to four parts of the flame in Fig. 3(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. Fig. 3(c-f) show the spatial and scalar-value distributions of the four groups.

The mixture fraction variable represents the fraction of fuel and oxidizer and typically indicates where the flame locates when they are in proper proportions. The non-combustion region in Fig. 3(f) has the highest value of MIX, from 0.85 to 1 (pure fuel), while the outer layer of the flame in Fig. 3(c) has the lowest value of MIX, from 0 (pure oxidizer) to 0.1. This agrees with the state that the flame is typically located where the fuel and oxidizer are in stoichiometric proportions, either pure oxidizer or pure fuel will result in the extinction of reaction [1]. The value in the inner layer of the flame in Fig. 3(e) is relatively high, ranging from 0.7 to 0.8. Several other groups correspond to the body of the flame due to the imperfect clustering algorithm.

4.3 Discussion

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 in variables, biclusters, and scalar values.

Biclusters are generated in the preprocessing stage. The computational time for the deep water impact and turbulent combustion data set is 40 seconds and 400 seconds, and the number of generated biclusters is 539 and 923 respectively. The computational time ranges from less than one minute to several minutes, and is roughly proportional to the number of biclusters, which depends on the number of variables and the complexity of the volume.

Compared to previous methods in correlation analysis and multi-dimensional transfer functions, our framework extends the analysis of value combinations of two variables [8] to multiple variables. In our experiments, we only present the results with three variables, since the feature/phenomena is generally associated a subset of variables and our framework supports visual exploration of biclusters of all variables. Besides, our framework can effectively identify local features with a similar scalar-value pattern from multiple variables, which is complementary to previous global correlation analysis [14]. Compared to interactive classification [5], 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 too much prior knowledge and discover local correlations of variables efficiently.

5 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 grouped to reduce the complexity of user interaction, and 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 could 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 the analysis efficiency. We would like to extend our framework to time-varying multivariate data to capture the coherence in the time space.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by the National Key Research & Development Program of China (2017YFB0202203), National Natural Science Foundation of China (61472354 and 61672452), NSFC-Guangdong Joint Fund (U1611263), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] H. Akiba, K.-L. Ma, J. H. Chen, and E. R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Computing in Science and Engg.*, 9(2):76–83, Mar. 2007.

[2] A. Biswas, S. Dutta, H. W. Shen, and J. Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2683–2692, Dec 2013.

[3] H. Carr and D. Duke. Joint contour nets. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1100–1113, 2014.

[4] S. Dutta, X. Liu, A. Biswas, H.-W. Shen, and J.-P. Chen. Pointwise information guided visual analysis of time-varying multi-fields. In *SIGGRAPH Asia 2017 Symposium on Visualization*, p. 17. ACM, 2017.

[5] H. Guo, H. Xiao, and X. Yuan. 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, March 2011.

[6] H. Guo, H. Xiao, and X. Yuan. Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates. *IEEE transactions on visualization and computer graphics*, 18(9):1397–1410, 2012.

[7] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[8] X. Liu and H.-W. Shen. Association analysis for visual exploration of multivariate scientific data sets. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):955–964, Jan 2016.

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

[10] S. Nagaraj, V. Natarajan, and R. S. Nanjundiah. A gradient-based comparison measure for visual analysis of multifield data. *Computer Graphics Forum*, 30(3):1101–1110, 2011.

[11] A. Oghabian, S. Kilpinen, S. Hautaniemi, and E. Czeizler. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one*, 9(3):e90801, 2014.

[12] J. Patchett and J. Ahrens. Optimizing scientist time through in situ visualization and analysis. *IEEE computer graphics and applications*, 38(1):119–127, 2018.

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

[14] N. Sauber, H. Theisel, and H. P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):917–924, Sept 2006.

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

[16] F.-Y. Tzeng and K.-L. Ma. 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, 2004.

[17] T. Van Long and L. Linsen. Multiclustertree: Interactive visual exploration of hierarchical clusters in multidimensional multivariate data. *Computer Graphics Forum*, 28(3):823–830, 2009.

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

[19] F. Wu, G. Chen, J. Huang, Y. Tao, and W. Chen. Easyxplorer: A flexible visual exploration approach for multivariate spatial data. *Computer Graphics Forum*, 34(7):163–172, October 2015.

[20] X. Zhao and A. Kaufman. 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, 2010.