IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 2, 2018, accepted August 7, 2018, date of publication August 10, 2018, date of current version September 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2864685

Correlation Analysis for Exploring Multivariate Data Sets

LI WANG, XIAOAN TANG, JUNDA ZHANG, AND DONGDONG GUAN
College of Electronic Science, National University of Defense Technology, Changsha 410073, China

Corresponding author: Li Wang (wangli08a@126.com)

ABSTRACT Correlation analysis is of great significance for exploring the multivariate data sets as it helps researchers toward an in-depth understanding of the complex interactions and relationships among variables. In this paper, we propose a correlation analysis method that identifies salient scalars for multivariate data exploration. We exploit specific mutual information metric to measure the information overlap and analyze the relationships between one scalar and other variables. Moreover, we define the information flow and introduce another metric, influence to quantify the associations among scalars of different variables. Furthermore, we integrate these two information metrics and construct a surprise-influence map for users' interaction to identify the salient scalars. By investigating the relationships among these salient scalars, we analyze the correlations among variables. We demonstrate the applicability and effectiveness of our proposed method by applying it to different data sets.

INDEX TERMS Multivariate data, correlation analysis, specific mutual information, information overlap, information flow.

I. INTRODUCTION

Multivariate data sets are widely produced in scientific and engineering domains such as computational fluid dynamics, climate, and aerodynamics, and exploration of these data is of great significance in understanding the intrinsic mechanisms of these physical and simulation phenomena [1]. Generally, there exist complex and heterogeneous interactions among different variables, which make the understanding of the multifaceted data difficult. Hence, the correlations among variables needs to be thoroughly investigated and analyzed to get an in-depth comprehension of the multivariate data sets. Traditional voxel-level analysis methods are tedious and challenging because of high resolution and unprecedented sizes of the data, however, it is more effective to discretize and bin the data into scalar values as the amount of the scalar-level data to be processed largely decreases. By identifying the most representative scalars of variables, researchers could understand the relationships among them and then analyze the correlations among variables for exploring the multivariate data. Nonetheless, it is still difficult to identify the salient scalars without sufficient prior knowledge [2].

In this paper, we propose a correlation analysis method that identifies salient scalars to explore the multivariate data sets. We exploit two information metrics, surprise and influence to help towards in-depth identification of the representative scalars. Since mutual information measures the information overlap among variables, we decompose it into specific mutual information (SMI) to quantify the shared information between one scalar and other variables (i.e. surprise), which facilitates the identification of representative scalars with high surprise. On the other hand, we define the information flow by using conditional probability and quantify the associations among scalars of different variables (i.e. influence) based on an influence-passivity model, which provides another metric to enhance the ability of identification. By integrating the surprise with influence, we construct a surprise-influence map to guide users in the identification of the representative scalars of different variables, and further, we present how to explore the multivariate data by analyzing the interactions among these salient scalars. Experiments demonstrate the applicability and effectiveness of our proposed method.

Our main contributions are threefold:

1) We consider the information overlap and measure the relationships between one scalar and other variables by exploiting SMI, which facilitate the identification of salient scalars.
2) We take the information flow into account and quantify the associations among scalars of different variables, which enhance the ability of identification by integrating surprise.

3) We propose a correlation analysis method in which we explore the correlations among variables by analyzing the relationships among salient scalars.

This paper is organized as follows. In Section 2, we review the related work. In Section 3, we give a brief overview of our method. Section 4 discusses the two information metrics, surprise and influence and proposes the correlation analysis method in detail. Section 5 presents the experimental results on several data sets. In Section 6, we provide the discussion and future work. Section 7 demonstrates the conclusion.

II. RELATED WORK

In this section, we mainly review the works which are directly related to our research: correlation analysis methods, visual exploration applications, and information theory in data analysis.

A. CORRELATION ANALYSIS FOR MULTIVARIATE DATA

Correlation analysis is to comprehend the interactions and underlying relationships among variables. Plenty of previous works have been reviewed in [3] and [4]. Yang *et al.* [5] introduced a Nugget Management System (NMS) for exploration and analysis of multivariate data. Jänicke *et al.* [6] analyzed the multivariate data in a 2D attribute space by transforming the high dimensional data into point clouds. Turkay *et al.* [7] linked the items space and dimensions space and proposed a multidimensional data exploration model. Lee and Shen [8] studied the temporal trend relationships among the variables based on dynamic time warping by their SUBDTW algorithm. Zhang *et al.* [9] arranged the variables into a 2D layout and generated an interactive correlation map with spatial representations for analyzing the relationships of variables. Chen *et al.* [10] devised a sampling-based approach to correlation classification for time-varying multivariate data. Zhang *et al.* [11] analyzed the correlation of time-varying patterns for multivariate data by a dissimilarity-preserving cluster algorithm. Gosink *et al.* [12] studied the variable interactions with a third variable by defining a correlation field as the normalized dot product between two gradient fields from two variables. Romero *et al.* [13] took people in a social network as variables and devised an influence-passivity model to evaluate their influence and passivity. In this paper, we discretize the data into scalars and introduce two information metrics for identifying salient scalars, and then we analyze the interactions among these scalars and correlations among variables.

B. VISUAL EXPLORATION FOR MULTIVARIATE DATA

Visual exploration is to depict the relationships among variables by visualization technologies. Parallel coordinates plot (PCP) [14]–[16] is widely used for multivariate analysis, in which variables are represented as parallel axes. However, the ordering of the axes in PCP affects the visual exploration process, and inevitable clutter in PCP also poses another problem towards the analysis of relationships between neighboring axes [17]–[19]. Traditional scatter plot [20] is another widely used method for its ability to indicate the trend between two variables, but it is generally difficult to explore the data with huge sizes because of heavy overlap. Scatter plot matrices (SPLOM) [21] extended the traditional scatter plot by simultaneously plotting all pairs of scatter plots for all the variables, but it also tends to be cluttered with the increased number of variables. Guo *et al.* [22] created a novel transfer function design interface by integrating PCP and multidimensional scaling (MDS) plots for visualization of multivariate data. Nagaraj *et al.* [23] proposed a gradient-based measure that reveals the relationships among variables for the purpose of comparative visualization. Sauber *et al.* [24] introduced local correlation coefficients to analyze the relationships among variables. Tatu *et al.* [25] explored the interesting subspaces of high-dimensional multivariate data by proposing an interactive visualization system. In this paper, we construct a scatter plot by mapping scalars to a surprise-influence space to guide users in the identification of salient scalars, and then we visualize the multivariate data for further visual analysis.

C. INFORMATION THEORY IN DATA EXPLORATION

Information theory [26] which can represent the relationships among objects is also widely used in data analysis. Wang *et al.* [27] introduced information entropy and formulated a complete graph to study the causal relationships among the variables of a time-varying multivariate data set. Haidacher *et al.* [28] analyzed multimodal surface similarities by extending mutual information (MI) to multimodal domains. Dutta *et al.* [29] explored the time-varying multivariate data by pointwise MI. Biswas *et al.* [30] measured the information overlap between one scalar and a variable by surprise and predictability metrics, which mainly focused on one-way interactions between two variables and left out information propagation. Viola *et al.* [31] used MI to identify the best visualization view. Bruckner and Möller [32] explored isosurfaces of univariate data by using MI. Feixas *et al.* [33] used specific MI to fuse multimodal data sets. In this paper, we exploit MI metric to measure the information overlap for identifying salient scalars.

III. OVERVIEW

Our main goal is to identify the salient scalars of different variables and analyze how these variables interact with each other for exploring the multivariate data sets. To do this work, we first discretize the variables into scalars and select a reference variable. We quantify the information overlap between one scalar and other variables (i.e. surprise) by introducing the specific mutual information (SMI). To allow for more effective analysis of correlation, we define the information flow and incorporate another information metric named influence to quantify the associations among scalars of different variables. By integrating the surprise with influence, we construct a surprise-influence map to provide an interface for the users to identify the representative scalars of the reference variable, and then the strongly associated scalars of other variables are identified. Further, we visualize these salient scalars together for correlation visualization and analyze the correlations among variables. Fig. 1 is a workflow of our proposed method.

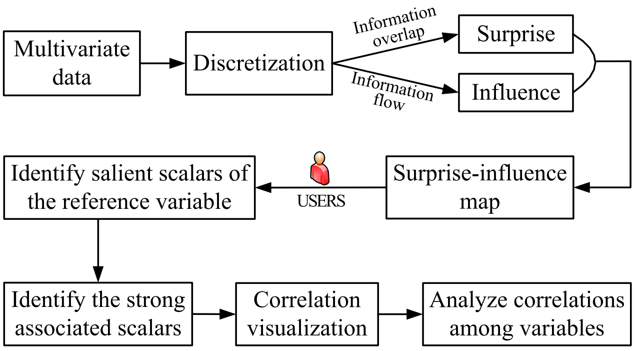

Multivariate data
Discretization
Information overlap
Information flow
Surprise
Influence
Surprise-influence map
USERS
Identify salient scalars of the reference variable
Identify the strong associated scalars
Correlation visualization
Analyze correlations among variables

FIGURE 1. A workflow of our correlation analysis method.

IV. CORRELATION ANALYSIS FOR MULTIVARIATE DATA SETS

Given two variables X and Y, we discretize them into scalars $x_i, i \in [1, M]$ and $y_j, j \in [1, N]$, and then the relationships between X and Y can be reflected in two aspects.

First, from the standpoint of information theory [34], as seen in Fig. 2(a), there exists information overlap between two associated variables, so we could gain some information about Y by observing X. The shared information is represented by joint probability $p(x_i, y_j)$ and quantified by mutual information (MI). The higher the MI is, the more information that X and Y shares.

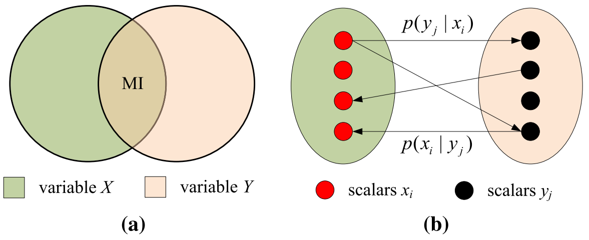


FIGURE 2. A schematic diagram of the information overlap and flow. (a) Information overlap. (b) Information flow, arrows represent the direction.

Second, from the perspective of probability theory, as seen in Fig. 2(b), there exists information flow between two associated scalars x_i and y_j represented by conditional probability $p(y_j|x_i)$ and $p(x_i|y_j)$. Higher $p(y_j|x_i)$ indicates that y_j depends on x_i more heavily and x_i is more likely to infer the existence of y_j, and more information flows from x_i to y_j.

In our paper, we exploit the information overlap as well as the information flow for analyzing the relationships between X and Y. Variables are discretized with histograms, and joint probability $p(x_i, y_j)$ is computed by 2D joint histogram and $p(x_i)$ is computed by 1D histogram, and conditional probability is computed as $p(y_j|x_i) = p(x_i, y_j)/p(x_i)$.

A. INFORMATION OVERLAP AND SURPRISE

In information theory [34], MI measures the shared or overlapped information among variables, and it also quantifies how much the uncertainty of one variable is reduced after given other variables. For two variables X and Y, MI is defined as:

$$I(X; Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{1}$$

where $p(x_i)$ is the probability of $x_i \in X$, $p(y_j)$ is the probability of $y_j \in Y$ and, $p(x_i, y_j)$ is their joint probability.

Since MI specifies the shared information on average over all (x_i, y_j) combinations, we decompose MI into SMI to measure the shared information between one scalar $x_i \in X$ and the variable Y, which quantifies how much the uncertainty of Y is reduced after given the scalar x_i. In this case, variable X is called the ***reference variable***, which can be selected by domain knowledge or other metrics (e.g. entropy). There are several ways to calculate the SMI and the expression of SMI is not unique, but it must fulfill $\sum\limits_{x_i \in X} p(x_i)I(x_i; Y) = I(X; Y)$. An expression of SMI named ***surprise*** [26] is:

$$I(x_i; Y) = \sum_{y_j \in Y} p(y_j|x_i) \log \frac{p(y_j|x_i)}{p(y_j)} \tag{2}$$

where $p(y_j|x_i)$ is the conditional probability.

The surprise value is strictly nonnegative because it represents the Kullback-Leibler distance [35] between $p(Y)$ and $p(Y|x)$, so it is particularly large when $p(y_j|x_i)$ dominates in regions of X where $p(y_j)$ is small, in that case, the observation of x_i has moved our estimate of y_j towards values that seemed very unlikely prior to the observation x_i [26]. In other words, higher surprise value indicates that some infrequent occurrences y_j have become more probable due to the observation of x_i. On the other hand, higher surprise value means that the x_i shares more information with Y and reduces more uncertainty of Y, since the total shared information between X and Y is a constant $I(X; Y)$, it also denotes that other scalars share less information with Y and reduce less uncertainty of Y. Above all, the scalars $x_i \in X$, for which the surprise value is higher, are more representative of the reference variable X, and we should take these scalars into more account for further analysis.

B. INFORMATION FLOW AND INFLUENCE

In probability theory, variables are not isolated and information is not static. There exists information propagation among variables in a multivariate data set, and each scalar is transmitting as well as receiving information simultaneously. Based on this consideration, we define the ***information flow*** that from scalar x_i to y_j as conditional probability $p(y_j|x_i)$ which measures the information content that y_j accepts from x_i. Then, the acceptance rate [13] of y_j with respect to x_i can be defined as:

$$a_{x_iy_j} = \frac{p(y_j|x_i)}{\sum\limits_{x_k \in X} p(y_j|x_k)} \tag{3}$$

The acceptance rate measures the information content that y_j accepts from x_i normalized by the total information content that y_j accepts from all its associated x_i. Similarly, the acceptance rate of x_i with respect to y_j is:

$$a_{y_jx_i} = \frac{p(x_i|y_j)}{\sum\limits_{y_k \in Y} p(x_i|y_k)} \tag{4}$$

On the other hand, $1 - p(y_j|x_i)$ measures the information content that y_j rejects from x_i, and the rejection rate of y_j with respect to x_i is defined as:

$$r_{x_iy_j} = \frac{1 - p(y_j|x_i)}{\sum\limits_{y_k \in Y} (1 - p(y_k|x_i))} \tag{5}$$

The rejection rate measures the information content that y_j rejects from x_i normalized by the total information content that rejected from x_i by all associated y_j. Similarly, the rejection rate of x_i with respect to y_j is:

$$r_{y_jx_i} = \frac{1 - p(x_i|y_j)}{\sum\limits_{x_k \in X} (1 - p(x_k|y_j))} \tag{6}$$

Reference [13] devised an influence-passivity model (IP model) based on two important conceptions, *influence* and *passivity* to explore the social network. Similarly, we redefine the ***influence*** and ***passivity*** for our multivariate data analysis.

- Influence: influence indicates how actively one scalar interacts with others, and higher influence value represents more active interactions.
- Passivity: passivity indicates how dully one scalar responds to others, and higher passivity value denotes more dull responses.

Next, based on the IP model, we iteratively calculate the influence and passivity values for each scalar as:

$$\begin{cases} I_{x_i} = \sum\limits_{y_j \in Y} a_{x_iy_j} P_{y_j} \\ P_{x_i} = \sum\limits_{y_j \in Y} r_{y_jx_i} I_{y_j} \end{cases} \tag{7}$$

where I_{x_i} and P_{x_i} are the influence and passivity values of $x_i \in X$, respectively.

In (7), we find that the influence value of x_i is high in three cases: first, the amount of scalars $y_j \in Y$ that interacts with x_i is large; second, $a_{x_iy_j}$ is large which indicates that some y_j have highly accepted x_i; third, passivity P_{y_j} is high which means that some dull y_j have responded to x_i and accepted information from x_i. Hence, the scalars $x_i \in X$, for which the influence value is higher, are more representative of the reference variable X, and we should take these scalars into more account for further analysis.

Unlike the surprise which represents the relationships between one scalar and other variables, the influence quantifies the associations among scalars of different variables. By integrating these two metrics, we could identify the representative scalars x_i of the reference variable X in the following part.

C. CORRELATION ANALYSIS BASED ON SURPRISE AND INFLUENCE

1) SURPRISE-INFLUENCE MAP

Based on (2) and (7), each scalar $x_i \in X$ corresponds to two values, the surprise and influence. We map each x_i to the surprise-influence space and construct a scatter plot, in which each point represents one specific scalar x_i, and the x-axis represents the surprise and the y-axis represents the influence. Specifically, in the surprise-influence map, each point is colored by its scalar value which makes it intuitive for us to explore the distribution patterns of these scalars. Compared with traditional scatter plot, this surprise-influence map alleviates overlap and clutter because the number of points is greatly lower than the sizes of the original data sets by discretization.

2) IDENTIFICATION OF SALIENT SCALARS

As discussed above, the scalars x_i, for which the surprise value or influence value is higher, are more representative of the reference variable X, and we should pay more attention to them. With this guideline, we interactively select points or regions from the surprise-influence map where the surprise or influence is higher or both are higher, then the corresponding salient scalars x_i are identified. Specifically, considering the importance of domain knowledge in multivariate data exploration, we incorporate it into our method to help users with the identification.

After the representative scalars x_i are identified, the next task is to identify the associated $y_j \in Y$ for correlation analysis. We must realize that only the strongly associated scalars are identified, we can obtain a confident analysis result. Since conditional probability $p(y_j|x_i)$ represents the information content that flows from x_i to y_j, we define the strongly associated y_j as those with higher $p(y_j|x_i)$ than a preset threshold. Higher $p(y_j|x_i)$ means that y_j accepts more information from x_i, which may lead to higher acceptance rate $a_{x_iy_j}$ and influence value I_{x_i} in (7). Furthermore, higher $p(y_j|x_i)$ implicates that it is more likely to generate a higher surprise value in (2). Hence, the strongly associated y_j in turn corroborate that the identified x_i are indeed representative of the reference variable X.

3) CORRELATION VISUALIZATION AND CORRELATION ANALYSIS

We visualize the representative scalars $x_i \in X$ and associated $y_j \in Y$ together for correlation visualization. By exploring the relationships between these identified scalars, we analyze the correlations between variables X and Y. Though the identification of salient scalars is discussed and derived based on two variables, it can be easily extended to multiple variables.

For three variables X, Y, and Z, (2) is expressed as:

$$I(x_i; Y, Z) = I(x_i; Y) + I(x_i; Z|Y) = \sum_{y_j \in Y} \sum_{z_k \in Z} p(y_j, z_k|x_i) \log \frac{p(y_j, z_k|x_i)}{p(y_j, z_k)} \quad (8)$$

and (7) is expressed as

$$\begin{cases} I_{x_i} = \sum_{y_j \in Y} a_{x_i y_j} P_{y_j} + \sum_{z_k \in Z} a_{x_i z_k} P_{z_k} \\ P_{x_i} = \sum_{y_j \in Y} r_{y_j x_i} I_{y_j} + \sum_{z_k \in Z} r_{z_k x_i} I_{z_k} \end{cases} \quad (9)$$

For more variables, and so on.

To sum up, we propose our correlation analysis method and the complete process is presented in **Algorithm 1**.

Algorithm 1 Correlation Analysis for Multivariate Data Sets

```
Input: Multivariate data set, the number of histogram bins.
Output: Salient scalars and correlation results.
Discretize the variables into scalars, and compute probability for each scalar and joint probability for every pair of scalars.
Select the reference variable $X$.
Calculate the surprise value for each $x_i$. (equation 2)
Compute the influence value for each $x_i$. (equation 7)
Construct the surprise-influence map.
Incorporate domain knowledge, interactively identify the representative $x_i$ of the reference variable $X$ with higher surprise or influence value.
Identify the strongly associated scalars of all variables based on conditional probability.
Visualize the identified scalars together for correlation visualization.
Analyze the correlations between or among variables.
```

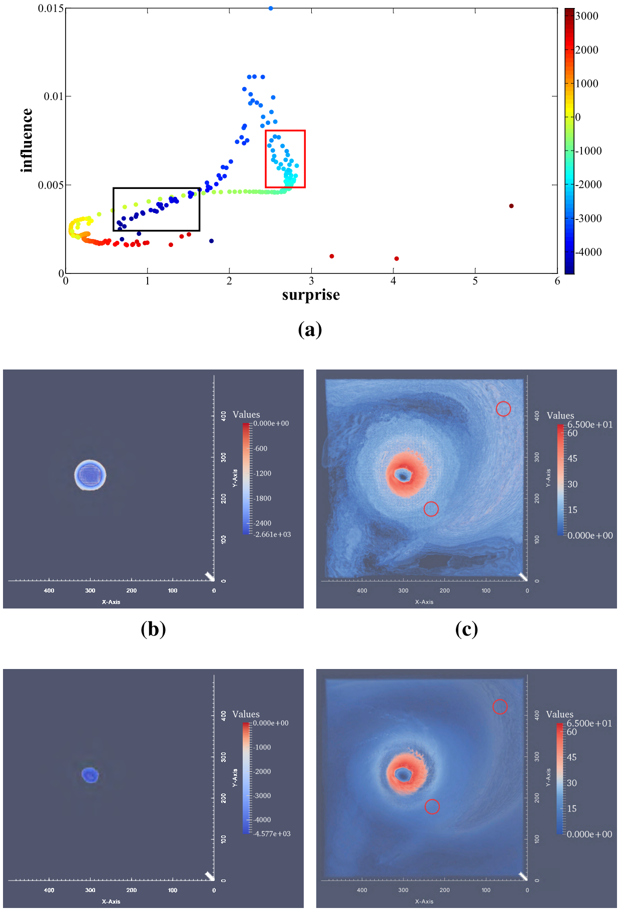
V. EXPERIMENTS

In this section, we demonstrate the applicability and effectiveness of our method on two data sets: Hurricane Isabel data set and Ionization Front Instability data set. Experiments were performed on a Windows 7 desktop computer with an Intel core i5-6500 CPU and 8 GB of RAM. All data sets were discretized with 256 histogram bins for calculating the probabilities, and the threshold for identifying the strongly associated scalars was 5%. The visualizations were generated by using ParaView [36].

A. HURRICANE ISABEL DATA SET

Hurricane Isabel data set [37] was produced by the Weather Research and Forecast (WRF) model, courtesy of NCAR and the U.S. National Science Foundation (NSF). This data set is an atmospheric simulation data with 13 variables and a resolution of 500×500×100. To investigate the relationships among pressure, wind speed, and humidity, we selected the pressure (PRE), wind velocity (VEL), and water vapor (QVA) for exploring the data with PRE as the reference variable.

Fig. 3 analyzed the correlations between PRE and VEL. Fig. 3(a) presented the surprise-influence map, and the red rectangular identified the hurricane eyewall in Fig. 3(b) and the black rectangular identified the hurricane eye in Fig. 3(d). As seen in Fig. 3(c) and 3(e), the global VEL was associated with both hurricane eyewall and eye, and the VEL values around the hurricane center (red regions) were strikingly high. However, a majority of the VEL values in Fig. 3(c) was much higher than Fig. 3(e), particularly in the white regions as shown by red circles. It highlights that the hurricane eye mainly resulted in strong wind around the center, while the hurricane eyewall could generate strong wind in much vaster areas as well as around the center. Hence, we can conclude that the hurricane eyewall is more representative in terms of strong wind, which corroborates the identification that the hurricane eyewall was more salient with higher surprise and influence than the hurricane eye. In addition, we also find that the VEL values were approximately zero at the center of the hurricane, which is consistent with our daily experience that it is clear blue skies at the center of the typhoon.



(a)

(b)

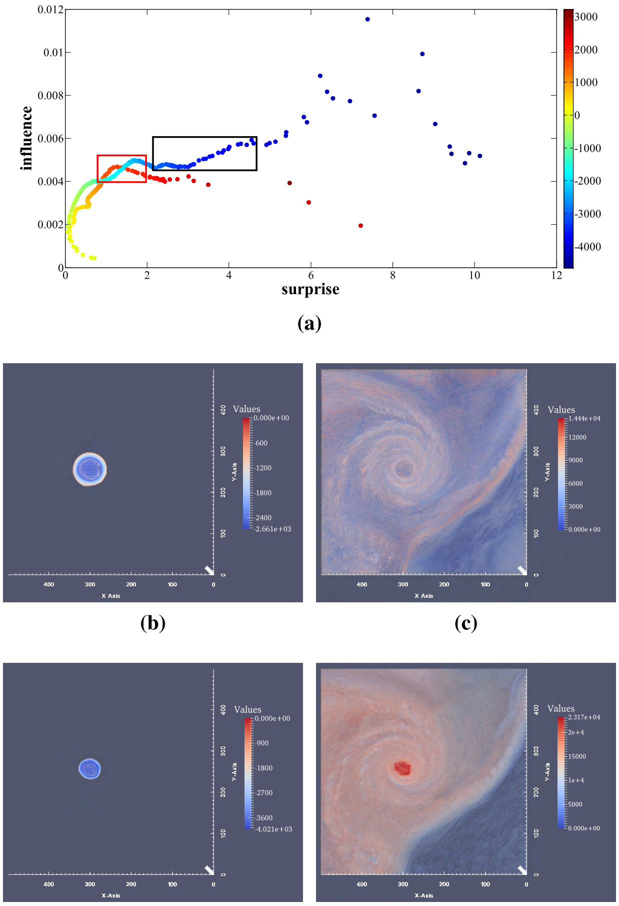
(c)

(d)

(e)

FIGURE 3. Correlation analysis between PRE and VEL. (a) Surprise-influence map, red rectangular identified the hurricane eyewall with higher surprise and influence, black rectangular identified the hurricane eye with lower surprise and influence. (b) The hurricane eyewall with PRE = −2200. (c) VEL associated with the hurricane eyewall. (d) The hurricane eye with PRE = −4600. (e) VEL associated with the hurricane eye.

Fig. 4 analyzed the correlations between PRE and QVA. Fig. 4(a) presented the surprise-influence map, and the red rectangular identified the hurricane eyewall in Fig. 4(b) and the black rectangular identified the hurricane eye in Fig. 4(d). As seen in Fig. 4(c) and 4(e), the global QVA was related to hurricane eyewall and eye, and long bands of rain clouds spiraled inward to the hurricane center. In Fig. 4(c), the QVA values associated with the hurricane eyewall were distributed between 0-1444, while in Fig. 4(e), the QVA values associated with the hurricane eye were 0-2317, and the QVA values in Fig. 4(e) were much higher than that in Fig. 4(c) almost everywhere. Particularly, the QVA values were the highest at the center of the hurricane (red regions). It indicates that compared with the hurricane eyewall, the hurricane eye was more representative in terms of QVA, which corroborated the identification that the hurricane eye was more salient with higher surprise than the hurricane eyewall. In this sense, we conclude that hurricane eye is with strong spiral rainbands.



(a)

(b)

(c)

(d)

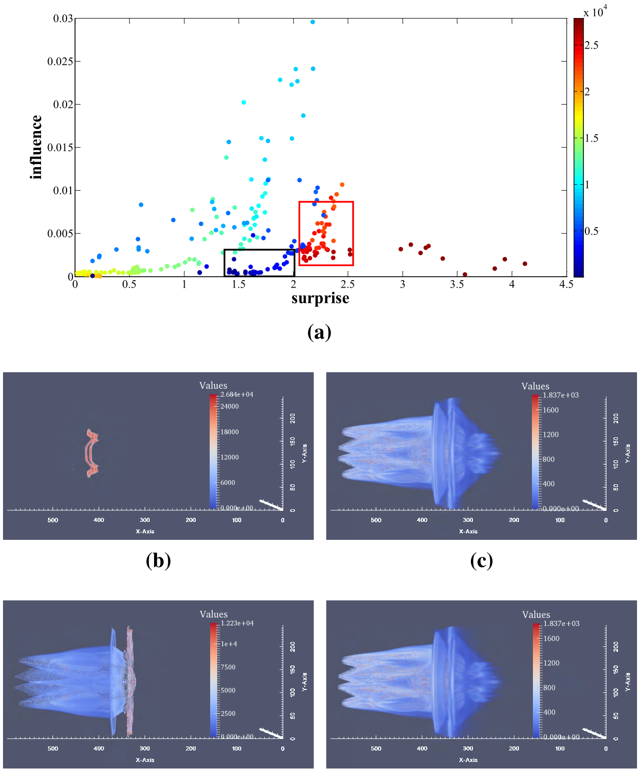
(e)

FIGURE 4. Correlation analysis between PRE and QVA. (a) Surprise-influence map, black rectangular identified the hurricane eye with higher influence and surprise, red rectangular identified the hurricane eyewall with higher influence but lower surprise. (b) The hurricane eyewall with PRE = −2200. (c) QVA associated with the hurricane eyewall. (d) The hurricane eye with PRE = −4000. (e) QVA associated with the hurricane eye.

B. IONIZATION FRONT INSTABILITY DATA SET

Ionization Front Instability data set was produced for exploring the relationships of the ionization front instabilities with the formation of the first stars of the universe [38]. The data set has 13 variables and a resolution of 600×248×248. To investigate that how temperature is related to the turbulence and ionization process, we selected the temperature (TEM), curl magnitude (MAG), and density of hydrogen ion (DoH) for exploring the data with TEM as the reference variable.

Fig. 5 analyzed the correlations between TEM and MAG. Fig. 5(a) presented the surprise-influence map, and the red rectangular identified high TEM in Fig. 5(b) and the black rectangular identified low TEM in Fig. 5(d). As seen in Fig. 5(c) and 5(e), it indicates that both high and low TEM impacted on MAG as the whole tail of the turbulence was related to the identified TEM scalars. Furthermore, compared Fig. 5(b) and 5(d), the high TEM areas interacting with MAG were much smaller than the areas with low TEM, while they related to almost the same results in Fig. 5(c) and 5(e), it denotes that much more scalars of MAG were associated with each high TEM scalar, and hence, high TEM is more representative in terms of MAG which corroborates the identification that the high TEM was more salient with higher



(a)

(b)

(c)

(d)

(e)

FIGURE 5. Correlation analysis between TEM and MAG. (a) Surprise-influence map, red rectangular identified high TEM with higher surprise but lower influence, black rectangular identified low TEM with higher surprise but lower influence. (b) High TEM = 20000. (c) MAG associated with high TEM. (d) Low TEM = 1000. (e) MAG associated with low TEM.

surprise than the low TEM. Then we draw a conclusion that high TEM results in the turbulence.

Fig. 6 analyzed the correlations between TEM and DoH. Fig. 6(a) presented the surprise-influence map, the blue rectangular identified medium TEM in Fig. 6(d) and the black rectangular identified low TEM in Fig. 6(f), specifically, we identified high TEM in Fig. 6(b) with the red rectangular based on our prior knowledge. As seen in Fig. 6(c), 6(e), and 6(g), DoH was greatly affected by high TEM as the hydrogen ions were spread across the whole areas and DoH was particularly high at the head of the turbulence, while medium/low TEM values were mainly related to relatively low DoH in the tail of the turbulence, then we can conclude that high TEM more greatly affects the ionization of hydrogen. In addition, we find that high TEM was not with higher surprise or influence in Fig. 6(a), it indicates that there might be some other variables together with TEM that resulted in high DoH, and we will analyze the multiple-to-one interactions for more in-depth exploration in the future.

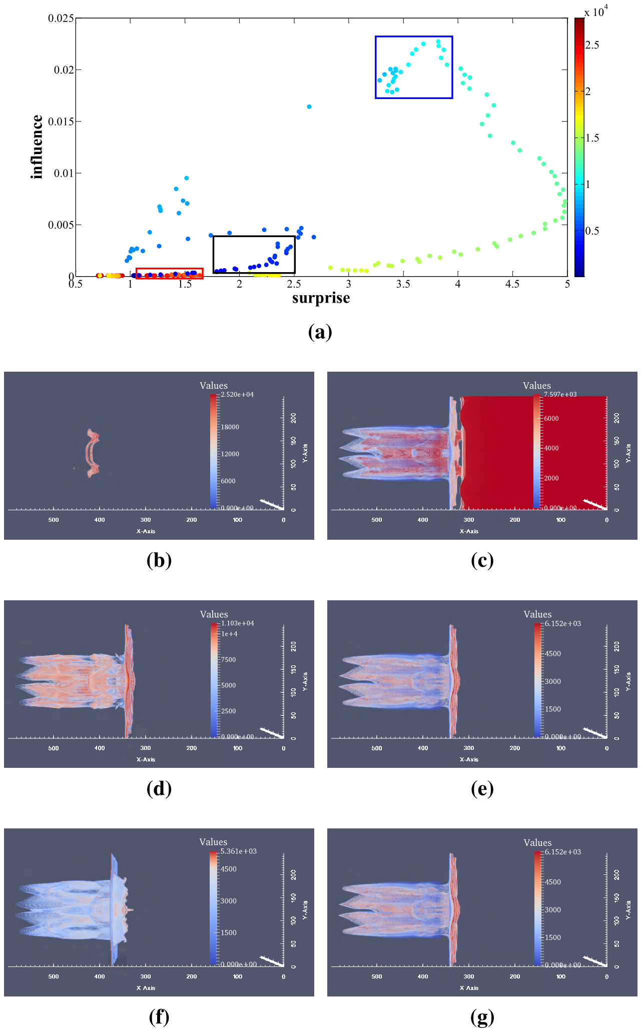


FIGURE 6. Correlation analysis between TEM and DoH. (a) Surprise-influence map, red rectangular identified high TEM by domain knowledge, blue rectangular identified medium TEM with higher surprise and influence, black rectangular identified low TEM with lower surprise and influence. (b) High TEM = 20000. (c) DoH associated with high TEM. (d) Medium TEM = 10000. (e) DoH associated with medium TEM. (f) Low TEM = 3000. (g) DoH associated with low TEM.

VI. DISCUSSION AND FUTURE WORK

Our method has only two parameters, the number of histogram bins and the threshold for identifying the strongly associated scalars. Since the threshold is set for visual analysis, we try it with different values until we get a relatively intuitive and succinct visualization result. Hence, we mainly focus on the selection of the number of bins because it affects the calculation of probabilities based on the 1D histogram and 2D joint histogram. Generally, the higher the number of bins is, the more precise probabilities we get, but it is meaningless if we increase it to the data size, and using too many bins even leads to clutter in the surprise-influence map. In this part, we show the results by using 128 and 512 bins. Compared with Fig. 3(a), it indicates that the general patterns of the surprise-influence maps remain the same although there are some differences and 512 bins lead to slight clutter in Fig. 7(b).

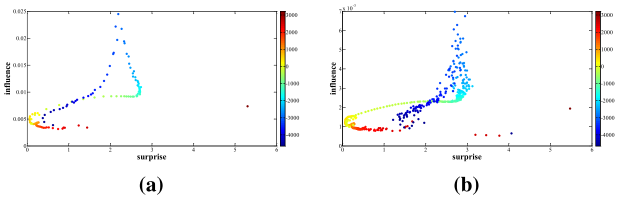


FIGURE 7. Surprise-influence map of PRE and VEL with different histogram bins. (a) 128 bins. (b) 512 bins.

Our method analyzes the correlations in a one-to-multiple way, and we can only analyze how the reference variable affects other variables. However, without sufficient prior knowledge or other guides, the reference variable we select may not be the most representative and significant one that greatly impacts on other variables, such as the TEM in terms of DoH, and hence, we cannot get comprehensive analysis results in these cases. In the future, we will explore the multiple-to-multiple methods to analyze how multiple variables impact on one or more variables.

Furthermore, in a multivariate data, there are some regions of interest (i.e. feature) which are of great significance for further exploration (e.g. feature visualization). However, precise definition of a feature is usually unavailable or we can only obtain a fuzzy description about it, which makes the extraction of such features difficult. Since our method identifies the representative scalars of variables, it provides a preliminary and rough way to extract these features, a typical case is the turbulence that hard to be described and defined with mathematical equations while it is identified by our method, as seen in Fig. 5 (c) and 5(e). In the future, we will apply our method to the domain of feature extraction with better performance.

VII. CONCLUSION

In this paper, we proposed a correlation analysis method that identified salient scalars to explore the multivariate data. We measured the information overlap and analyzed the relationships between scalars and variables by exploiting SMI metric. We defined the information flow and quantified the associations among scalars of different variables by introducing another information metric named influence. We constructed a surprise-influence map and identified the representative scalars of the reference variable, and further, we identified the strongly associated scalars of all other variables. Finally, we analyzed the correlations among variables based on these salient scalars. We demonstrated the applicability and effectiveness of our proposed method by applying it to different data sets.

REFERENCES

[1] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 3, pp. 495–513, Mar. 2013.

[2] H. Janicke, A. Wiebel, G. Scheuermann, and W. Kollmann, "Multifield visualization using local statistical complexity," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1384–1391, Nov. 2007.

[3] P. C. Wong and R. D. Bergeron, "30 years of multidimensional multivariate visualization," in *Scientific Visualization, Overviews, Methodologies, and Techniques*. Washington, DC, USA: IEEE Computer Society, 1997, pp. 3–33.

[4] R. Fuchs and H. Hauser, "Visualization of multi-variate scientific data," *Comput. Graph. Forum*, vol. 28, no. 6, pp. 1670–1690, 2009.

[5] D. Yang, E. A. Rundensteiner, and M. O. Ward, "Analysis guided visual exploration of multivariate data," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol. (VAST)*, Sacramento, CA, USA, Oct. 2007, pp. 83–90.

[6] H. Jänicke, M. Böttinger, and G. Scheuermann, "Brushing of attribute clouds for the visualization of multivariate data," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1459–1466, Nov. 2008.

[7] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions—A dual visual analysis model for high-dimensional data," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2591–2599, Dec. 2011.

[8] T.-Y. Lee and H.-W. Shen, "Visualization and exploration of temporal trend relationships in multivariate time-varying data," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, no. 6, pp. 1359–1366, Sep. 2009.

[9] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 2, pp. 289–303, Feb. 2015.

[10] C.-K. Chen, C. Wang, K.-L. Ma, and A. T. Wittenberg, "Static correlation visualization for large time-varying volume data," in *Proc. IEEE Pacific Vis. Symp.*, Hong Kong, Mar. 2011, pp. 27–34.

[11] H. Zhang, Y. Hou, D. Qu, and Q. Liu, "Correlation visualization of time-varying patterns for multi-variable data," *IEEE Access*, vol. 4, pp. 4669–4677, Aug. 2016.

[12] L. Gosink, J. Anderson, W. Bethel, and K. Joy, "Variable interactions in query-driven visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1400–1407, Nov. 2007.

[13] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2011, pp. 33–181.

[14] A. Inselberg, "The plane with parallel coordinates," *Vis. Comput.*, vol. 1, no. 2, pp. 69–91, Aug. 1985.

[15] A. Inselberg, "Multidimensional detective," in *Proc. IEEE Symp. Inf. Vis.*, Phoenix, AZ, USA, Oct. 1997, pp. 100–107.

[16] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proc. 1st IEEE Conf. Vis.*, San Francisco, CA, USA, Oct. 1990, pp. 361–378.

[17] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *J. Consum. Res.*, vol. 34, no. 4, pp. 441–458, 2007.

[18] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.

[19] R. Agrawal, T. Imieliński, and A. N. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[20] P. E. Touchette, R. F. Macdonald, and S. N. Langer, "A scatter plot for identifying stimulus control of problem behavior," *J. Appl. Behav. Anal.*, vol. 18, no. 4, pp. 343–351, 1985.

[21] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1148–1539, Nov./Dec. 2008.

[22] H. Guo, H. Xiao, and X. Yuan, "Scalable multivariate volume visualization and analysis based on dimension projection and parallel coordinates," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 9, pp. 1397–1410, Sep. 2012.

[23] S. Nagaraj, V. Natarajan, and R. S. Nanjundiah, "A gradient-based comparison measure for visual analysis of multifield data," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 1101–1110, 2011.

[24] N. Sauber, H. Theisel, and H.-P. Seidel, "Multifield-graphs: An approach to visualizing correlations in multifield scalar data," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 5, pp. 917–924, Sep./Oct. 2006.

[25] A. Tatu *et al.*, "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *Proc. Vis. Anal. Sci. Technol.*, Seattle, WA, USA, Oct. 2012, pp. 63–72.

[26] M. R. DeWeese and M. Meister, "How to measure the information gained from one symbol," *Netw., Comput. Neural Syst.*, vol. 10, no. 4, pp. 325–340, Nov. 1999.

[27] C. Wang, H. Yu, R. W. Grout, K.-L. Ma, and J. H. Chen, "Analyzing information transfer in time-varying multivariate data," in *Proc. IEEE Pacific Vis. Symp.*, Hong Kong, Mar. 2011, pp. 99–106.

[28] M. Haidacher, S. Bruckner, and M. E. Gröller, "Volume analysis using multimodal surface similarity," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 1969–1978, Dec. 2011.

[29] S. Dutta, X. Liu, A. Biswas, H.-W. Shen, and J.-P. Chen, "Pointwise information guided visual analysis of time-varying multi-fields," in *Proc. SIGGRAPH Asia Symp. Vis.*, Bangkok, Thailand, 2017, pp. 1–17.

[30] A. Biswas, S. Dutta, H.-W. Shen, and J. Woodring, "An information-aware framework for exploring multivariate data sets," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2683–2692, Dec. 2013.

[31] I. Viola, M. Feixas, M. Sbert, and M. E. Gröller, "Importance-driven focus of attention," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 5, pp. 933–940, Sep. 2006.

[32] S. Bruckner and T. Möller, "Isosurface similarity maps," *Comput. Graph. Forum*, vol. 29, no. 3, pp. 773–782, 2010.

[33] M. Feixas *et al.*, "Multimodal data fusion based on mutual information," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 9, pp. 1574–1587, Sep. 2012.

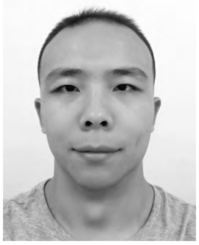
[34] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[35] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, 2007.

[36] U. Ayachit, "The paraview guide: A parallel visualization application," Kitware, New York, NY, USA, Tech. Rep., 2015.

[37] *IEEE Visualization 2004 Contest*. [Online]. Available: http://sciviscontest-staging.ieeevis.org/2004/

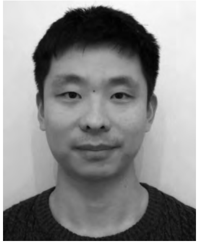
[38] *2008 IEEE Visualization Design Contest*. [Online]. Available: http://sciviscontest-staging.ieeevis.org/2008/



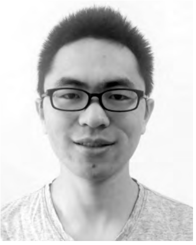
LI WANG received the B.S. degree in information engineering and the M.S. degree in information and communication engineering from the National University of Defense Technology in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree in information and communication engineering with the College of Electronic Science. His research interests include information processing, data analysis, computer graphics, and scientific visualization.

XIAOAN TANG was born in Huainan, Anhui, China, in 1968. He received the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT).

He is currently a Full Professor with the College of Electronic Science, NUDT. His research interests include information fusion, image process, computer graphics, scientific visualization, computer vision, and artificial intelligence.



JUNDA ZHANG received the B.S. degree from the Ocean University of China in 2011 and the M.S. degree in information and communication engineering from the National University of Defense Technology in 2013, where he is currently pursuing the Ph.D. degree in information and communication engineering with the College of Electronic Science. His research interests include image processing, computer graphics, volume rendering, and scientific visualization.



DONGDONG GUAN received the B.S. degree in geomatics engineering from Wuhan University in 2013 and the M.S. degree in photogrammetry and remote sensing from the National University of Defense Technology in 2015, where he is currently pursuing the Ph.D. degree in information and communication engineering with the College of Electronic Science. His research interests include SAR image processing, machine learning, and computation vision in remote sensing applications.