Journal of Visual Languages and Computing 41 (2017) 121–132

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Improved visual correlation analysis for multidimensional data

Yi Zhang, Teng Liu*, Kefei Li, Jiawan Zhang

Tianjin University, China

ARTICLE INFO

Article history:
Received 18 September 2016
Revised 8 February 2017
Accepted 23 March 2017
Available online 4 May 2017

Keywords:
Multidimensional visualization
Correlation analysis
Data distribution feature
Dimension reordering

ABSTRACT

With the era of data explosion coming, multidimensional visualization, as one of the most helpful data analysis technologies, is more frequently applied to the tasks of multidimensional data analysis. Correlation analysis is an efficient technique to reveal the complex relationships existing among the dimensions in multidimensional data. However, for the multidimensional data with complex dimension features,traditional correlation analysis methods are inaccurate and limited. In this paper, we introduce the improved Pearson correlation coefficient and mutual information correlation analysis respectively to detect the dimensions' linear and non-linear correlations. For the linear case,all dimensions are classified into three groups according to their distributions. Then we correspondingly select the appropriate parameters for each group of dimensions to calculate their correlations. For the non-linear case,we cluster the data within each dimension. Then their probability distributions are calculated to analyze the dimensions' correlations and dependencies based on the mutual information correlation analysis. Finally,we use the relationships between dimensions as the criteria for interactive ordering of axes in parallel coordinate displays.

1. Introduction

The rapid development of information technology produces vast amounts of datasets with numerous dimensions and complex structures. These multidimensional datasets offer tremendous opportunities for studying behavioral patterns and predicting future developments. Valuable insight often comes from intricate inter-relationships that exist among data dimensions(or variables). However, for the data with many dimensions and complex structures, it is far from straightforwardly showing the relationships between dimensions in a meaningful and user-interpretable way. Traditionally, low-dimensional representations of high-dimensional spaces [1], obtained by methods such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Self-Organization Map(SOM), etc. are used to interpret their relationships from a macro perspective. Other methods mainly include the scatter plot matrix (SPM) and the parallel coordinate plot (PCP) can basically show some correlations between variables in the multidimensional data.

Correlation analysis is one of the most commonly used methods in multidimensional visualization. It looks for relationships between variables and can indicate whether the variables are related to each other and how strong of the dependency is. Pearson correlation coefficient [2] is a most commonly used method for correlation analysis which is proposed by Pearson in 1895. It is often used in the multidimensional visualization to directly characterize the correlations between two variables by the coefficient *R*. Another method called canonical correlation coefficient [3] is also often applied for the multidimensional visualization. Both of above methods use the correlation coefficient *R* to show the variables' linear correlations. However, they reflect inaccurate relationship when the datasets are not normal distribution. And they are easily influenced by the outliers.

Faced with the multidimensional data with a variety of distributions and structures, traditional linear correlation analysis methods are not efficient to analyze the data relationship. Therefore, we propose an improved method based on the Pearson correlation coefficient in this paper. We think that the calculation for Pearson correlation coefficient should use different parameters according to the datasets with different distributions. We first extract the statistical features of multidimensional data to judge each dimension's distribution. Then all dimensions are classified into three groups according to their distributions. Finally, we select the appropriate parameters to calculate the Pearson correlation coefficient for each group of dimensions.

Correlation coefficient is a good measure when the dimensions are nearly linear distributed. But it appears not suitable for the analysis of non-linear distributed dimensions in the multidi-

* Corresponding author.
E-mail addresses: yizhang@tju.edu.cn (Y. Zhang), 1012606185@qq.com (T. Liu), 351540817@qq.com (K. Li), jwzhang@tju.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.jvlc.2017.03.005
1045-926X/

mensional data. Furthermore, we propose a non-linear correlation analysis method based on mutual information correlation analysis [4] and clustering. We use the information entropy to measure the relationships between variables. It is assumed that the smaller the entropy, the stronger the relationship. On the contrary, the relationship is weaker. This method is not influenced by the distributions of datasets. It has a robustness for the noise points. Firstly, we divide the data within each dimension into some clusters. Then, the probability distribution is given by the frequency of the data points within each cluster. As well as the joint probability distribution between two dimensions and the marginal probability distribution of every single dimension are obtained. Finally, we use the mutual information correlation analysis method to analyze the dimensions' correlations and dependencies.

Since we have achieved the relationships between any two dimensions in the multidimensional data, we can further rearrange the sequence of all dimensions to clearly show the data relationships and meaningful structures. So, we take the correlations and dependencies as the reordering criteria. Besides, the features of dimensions are taken into account for the supplementation of dimension reordering.

Our main contributions can be summarized as follows:

- Calculating the Pearson correlation coefficient based on different parameters according to the dimensions' distributions;
- Proposing a non-linear correlation analysis method based on clustering and mutual information correlation analysis;
- Reordering all dimensions by the criteria of correlations and dependencies.

The remainder of the paper is organized as follows. Related work is discussed in Section 2. Section 3 then describes the framework and working principle of our method, including the linear and non-linear analysis methods. We present the experimental results of the methods in Section 4. Section 5 gives a summary along with plans for future work.

2. Related work

Correlation analysis can effectively help discover the relationships between variables. It can be divided into two categories: linear and non-linear methods. And the linear methods include the correlation analysis between two variables and multiple variables. The Pearson correlation coefficient [5] is a common correlation analysis method which is mainly used in the analysis of two numerical variables. Mcdonnell et al. [6] and Seo and Shneiderman [7] use the Pearson correlation coefficient to calculate the correlation between any two dimensions. Besides, the Spearman coefficient [8] and Q & R coefficient [9] are also used to analyze the correlations between two variables. And they are mainly used for the ordered data and categorical data respectively. The correlation analysis between multiple variables mainly include two methods of partial correlation coefficient [10] and canonical correlation coefficient [3]. Especially, the canonical correlation coefficient are often used for the multidimensional data to analyze the relationship between multiple dimensions. In the paper of Zhou et al. [11], the canonical correlation coefficient is used for the high-dimensional data streams. Our linear correlation analysis method is based on the Pearson correlation coefficient.

As for the non-linear correlation analysis,a commonly used method is the mutual information correlation analysis [12]. It performs the relationships of variables through the information entropy. This non-linear analysis method can detect the relationships between variables with any distributions.And it has a good robustness for noise points. Reshef et al. [4] propose a maximal information coefficient(MIC) to measure the correlations between two variables in the multidimensional data. They mainly divide the scatterplot of two variables into several grids. According to the frequency of points within each sublattice which belongs to any grid to calculate the correlations between two variables. Another non-linear correlation analysis method is based on distance [13]. It takes advantage of characteristic function's distance to measure two random variables' non-linear correlation. This method can detect the non-linear correlations of variables with any distribution. But it is a biased estimation and is susceptible to the number of dimensions. In the paper of Szekely and Rizzo [14], they improve the non-linear method based on distance and obtain the distance coefficient's unbiased estimation. Our non-linear correlation analysis method is based on the mutual information correlation analysis.

Statistical methods [15] can effectively help analyze the dimensions' features such as distribution type, degree of dispersion, the central tendency, outliers, etc. And these features can be used in the analysis of dimensions' similarity and importance,thus are widely adopted in the multidimensional visual analysis. In the paper of Seo and Shneiderman [7], they take the features as the reordering indexes and design an interactive system that allows users to rearrange dimensions according to their preferences. Fernstad et al. [16] take the features as a combined system. It allows users to select different values to filter dimensions. Cagatay et al. [17] and Turkay et al. [18] put the datasets into the dimension space by extracting their features. Then, they iterative filter the important dimensions in the data space and the dimension space. And a similar dimension is generated to replace several dimensions based on the feature similarities. In this paper we use the feature similarity to classify the dimensions into several groups. And we also select several features as the criteria for primary dimension's selection to supplement the dimension reordering.

Dimension reordering and clustering are often used for the multidimensional visual analysis to show the relationships between dimensions more clearly. Besides, it is helpful for us to find some inherit structures in the multidimensional data. Peng et al. [19] propose a method to calculate the clutter between dimensions which is based on the K-means cluster algorithm. Then all dimensions for PCP and SPM are reordered according to the clutter between dimensions. Artero et al. [20] use the similarities of dimensions to rearrange the sequence of all dimensions. Ferdosi and Roerdink [21] reorder the dimensions by the method of subspace clustering. From the results of reordering, we can find some similar data relationships in the same cluster. Zhao and Kaufman [22] mainly analyze the correlation between the adjacent dimensions by the clustering and reordering. Mcdonnell et al. [6] reorder dimensions according to the correlations between dimensions. In this paper, the clustering method is mainly used to calculate the probability distributions. Our reordering methods are based on both the correlations and dependencies between dimensions.

3. Correlation analysis framework

In the paper, we introduce two correlation analysis methods: (1) linear correlation analysis method based on Pearson correlation coefficient and dimension grouping; (2) non-linear correlation analysis method based on mutual information correlation analysis and clustering. And the relationships are taken as the reordering criteria to rearrange all dimensions in the parallel coordinate displays.

3.1. Linear correlation analysis

Pearson correlation coefficient, as one of the common correlation analysis methods, can detect the linear relationships between two variables. The coefficient R cannot only show whether pairs of variables are related and how strong the relationship is, but also

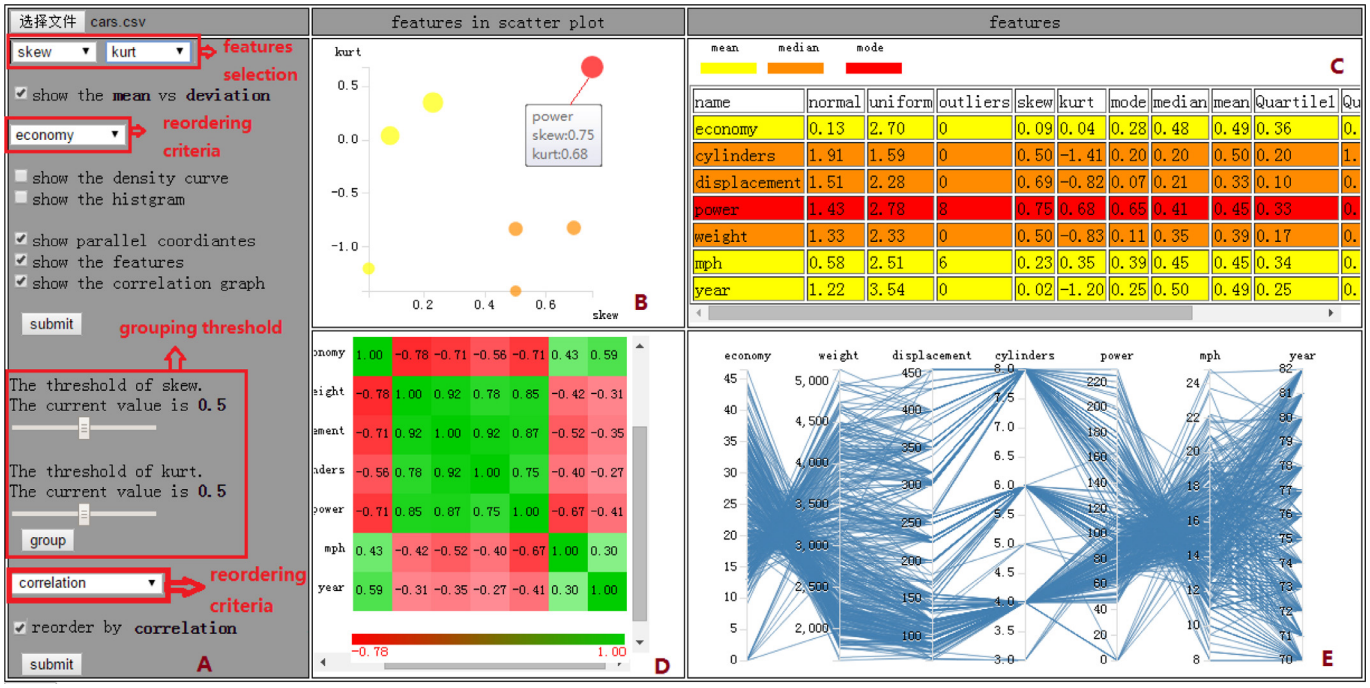

Fig. 1. Linear correlation analysis interactive framework. All dimensions are classified into three groups(B,C) according to the current thresholds of skew and kurt(A). And they are marked by colors. The dimensions in the "yellow" group use mean to calculate Pearson coefficient. The dimensions in the "orange" group use median to calculate Pearson coefficient. The dimensions in the "red" group use mode to calculate Pearson coefficient. Then the corresponding correlation coefficients are obtained(D). And they are color-coded. The green color represent the positive correlation and the red color represent the negative correlation. The darker the color and the stronger the correlation. Finally, all axes in the parallel coordinates are reordered(E) according to the current reordering criteria(A). (Data shown is from the Car dataset (7 dimensions × 406 samples)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

can judge that they are positive related or negative related. However, for the multidimensional data with a variety of distributions, the accuracy of Pearson coefficient will be influenced. In order to make the Pearson coefficient more accurate, we classify all dimensions into three groups according to their distribution features and calculate their correlations use the corresponding parameters. We take three features: normality of distribution, uniformity of distribution and outliers as the criteria for the primary dimension's selection in the dimensions reordering. Then all dimensions can be reordered by their correlations.

3.1.1. Framework overview

Our linear correlation analysis framework (Fig. 1) includes four parts: control panel(A), features view(B,C), correlations view(D) and parallel coordinates view(E). We extract the dimensions' statistical features which include mean, mode, median, deviation, skewness, kurtosis, first Quartile and third Quartile to display in the scatterplot(B) and table(C). And we also calculate three features that are normality of distribution,uniformity of distribution and outliers shown in the table(C) to implement the dimension reordering. The users can choose any two of the statistical features shown in the scatterplot(B) to observe their relationship. Then the users can select appropriate thresholds of skewness and kurtosis(A) to classify all dimensions into three groups(B,C). And they are marked by three colors: yellow, orange and red. Then, in the Pearson coefficient calculation, the parameters of mean, median and mode are selected correspondingly. Next, the linear correlations(D) are obtained by the Pearson correlation coefficient method. And they are mapped by colors. The green color represent the positive correlation and the red color represent the negative correlation. The darker the color and the stronger the correlation. Finally, the users can choose the reordering criteria to rearrange the sequence of dimensions(E).

Table 1
Dimension grouping.

Condition	Distribution description	Result
$\|skew\| < s$	Symmetrical distribution	Group "yellow"
$\|skew\| \geq s$, $kurt < k$	Skewed distribution with a flat peak	Group "orange"
$\|skew\| \geq s$, $kurt \geq k$	Skewed distribution with a tip peak	Group "red"

3.1.2. Method and procedure

1. Correlation analysis

Pearson coefficient method is effective only for the normally distributed variables, and is susceptible to outliers. This is because of the parameter mean in Pearson coefficient calculation. Usually, mean is suitable for the analysis of symmetrically distributed datasets and is susceptible to outliers. Mode and median have the similar effect in the data analysis,but they are not easily influenced by extreme values, so are suitable for the analysis of skewed distributed datasets. Especially, mode is suitable for the analysis of datasets with a high peak. Based on above analysis, we classify all dimensions into three groups according to their distribution characteristics and choose corresponding parameters of mean, median and mode to calculate the correlations.

In statistics, the skewness(*skew*) is often used to judge whether a dataset is symmetrical or skewed distribution. And the kurtosis(*kurt*) can intuitively describe the dataset's peak. A dataset is symmetrically distributed, when the value of *skew* is approximately equal to zero. If $|skew| > 0$ it is called a skewed distribution. If $kurt - 3 > 0$, the dataset has a tip peak. So, we use these two features to classify the dimensions into three groups. By setting two thresholds, k for kurtosis and s for skewness, each variable can be classified into one of three distribution groups shown in Table 1.

The calculation of Pearson correlation coefficient (Eq. (1)) in each group use different parameter.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

If a variable belongs to the group "yellow", $\bar{x}$ or $\bar{y}$ represents the parameter mean; If a variable belongs to the group "orange", $\bar{x}$ or $\bar{y}$ represents the parameter median; If a variable belongs to the group "red", $\bar{x}$ or $\bar{y}$ represents the parameter mode.

2. Dimension reordering

According to the correlations between dimensions, all dimensions shown in the parallel coordinates (Fig. 1(E)) can be reordered in a good view. In addition, we also take three features: normality of distribution,uniformity of distribution and outliers to supplement the dimension reordering. They are taken as the criteria for the primary dimension's selection.

(1) Normality

Many statistical analysis methods such as *t*-test, ANOVA are based on the assumption that the dataset is sampled from a normal distribution. In the analysis method of Spearman and coworkers [8,9], the normally distributed dimensions are regarded more important. So, it is useful to know the normality of the dataset. Here, we use the criteria of skewness (s) and kurtosis (k) to judge whether one dimension is normal distribution or not. Since s is 0 and k is 3 for a standard normal distribution, we calculate $|s| + |k-3|$ to measure how the distribution deviates from the normal distribution.

(2) Uniformity

We use the entropy of histogram from Mcdonnell et al. [6] to measure a dimension's uniformity.

$$entropy(H) = -\sum_{i=1}^{t} p_i \log_2^{p_i} \quad (2)$$

where t is the number of bins in the histogram, p_i is the probability that an item belongs to the ith bin. High entropy means that the dimension is close to a uniform distribution. If one dimension is far deviate from the uniform distribution, it sometimes reveals interesting outliers.

(3) Outliers

Generally, the results of data analysis will be bias because of the outliers' presence. For example, the Pearson coefficient can be seriously influenced by the outliers. So, outliers' identification is important. Here we also use the method from Mcdonnell et al. [6] to calculate the outliers. Let a^*IQR as a threshold. Where $a \in [1.5, 3]$ is a constant and IQR represents the difference between the first quartile (Q1) and the third quartile (Q3). An item of value is considered as an outlier if $x > (Q3 + a * IQR)$ or $x < (Q1 - a * IQR)$.

Firstly, we use the three features to select a primary dimension in the first place. The primary dimension can be closest to normal distribution or uniform distribution, and it has the least outliers. Then, a dimension from the rest of dimensions is arranged in the back of it, which guarantees that they have the strongest correlation. And so on, until all dimensions are reordered completely.

3.2. Non-linear correlation analysis

In order to precisely characterize the relationships between dimensions within the multidimensional data. We propose a nonlinear correlation analysis method based on the mutual information correlation analysis and clustering. This method is valid for all variables with any distribution. And it is not influenced by noise points. Besides, it cannot only calculate the variables' pairwise correlations, but also can get their dependencies. Finally, all dimensions can be reordered by their correlations and dependencies.

3.2.1. Framework overview

Our non-linear analysis framework (Fig 2) includes six parts:control panel(A), features view(B), probability distributions view(C), relationships view(D), parallel coordinates view(E) and pairs of dimensions' relationship visualization view(F). The users can select the statistical features to show them in the scatterplot(B). And the points belong to the same dimension are connected by the curve. By observing the curves in scatterplots, the users can detect the distribution similarity of dimensions. Then, the joint probabilities between all clusters are displayed in the map graph(C). Where the elements in the diagonal represent the probability distributions of each dimension. Their probabilities are mapped by colors. A dark color means a high probability. Then, we use the probability distributions to calculate the non-linear correlation and conditional information entropy between dimensions. The users can choose any one of them to show in the map graph(D). The values of relationships between dimensions are mapped by colors. A dark color means a strong relationship. The users can also select any of the elements in the correlation map graph(D) to observe their correlations in the scatterplot(F). Finally, the users can rearrange all dimensions in the parallel coordinates(E) by selecting the reordering criteria from both correlations and dependencies of dimensions. And the lines in PCP are clustered and bundled [23].

3.2.2. Method and procedure

In this non-linear correlation analysis method,we introduce a new method based on clustering to obtain the probability distributions of all variables in the multidimensional data. Then the correlations and dependencies between two variables are obtained by calculating the variables' information entropy based on the probability distributions. According to their correlations and dependencies,all dimensions can be reordered in a good view.

1. Clustering

Clustering can gather the data with similar features together. The data within the same cluster has a certain aggregate characteristic. Here, we take advantage of clustering to support the nonlinear correlation analysis. The probability distributions are given by the frequency of data points within each cluster. An improved K-means algorithm from Chen et al. [24] is selected as our clustering method. This method optimize the selection for initial cluster centers. It selects the data points with high density, large distance and high similarity in the same cluster as the initial cluster centers. Besides, an evaluation function is defined in the paper to evaluate the clustering results, which guarantees an optimal clustering result.

2. Definition of probability distribution

Assume two dimensions d_x and d_y (Fig 3). They all contain *count* samples. And the dimension d_x is divided into three clusters of $c_{x,1}$, $c_{x,2}$ and $c_{x,3}$. The dimension d_y is divided into two clusters of $c_{y,1}$ and $c_{y,2}$. The definitions of their probability distributions as follow.

The probability distribution of dimension d_x:

$$p_x(i) = \frac{num_{x,i}}{count}, i = 1, 2, 3 \quad (3)$$

where $num_{x,i}$ represents the number of data points within cluster $c_{x,i}$.

The probability distribution of dimension d_y:

$$p_y(j) = \frac{num_{y,j}}{count}, j = 1, 2 \quad (4)$$

where $num_{y,j}$ represents the number of data points within cluster $c_{y,j}$.

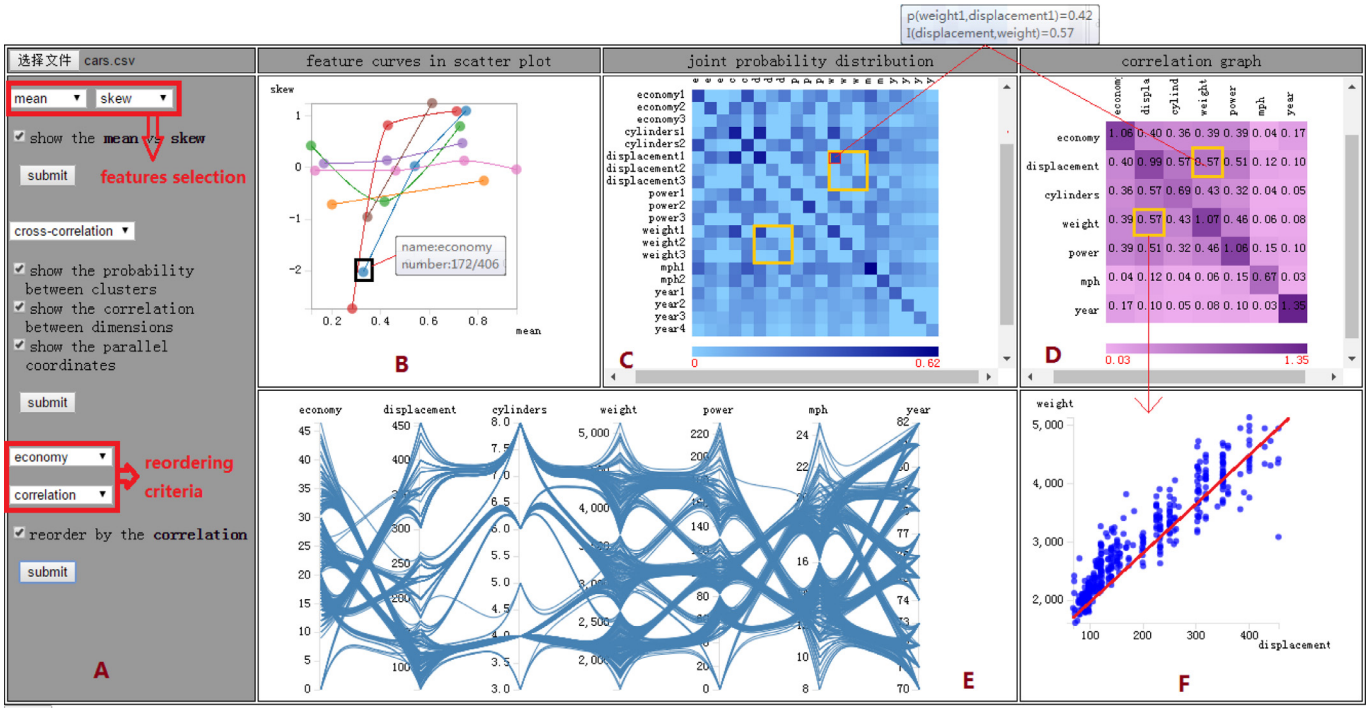

Fig. 2. Non-linear correlation analysis interactive framework. Each of the dimensions is divided into several clusters and their features are shown in the scatterplot(B). Their joint probability distributions(C) and non-linear correlations(D) are also obtained according to the current clustering. And they are mapped by colors. A dark color means a high probability and a strong correlation. An element in the correlation map(D) is clicked, correspondingly, the two dimensions' relationship is shown by the scatterplot(F). It can be seen that the dimensions of "weight" and "displacement" are strongly and positively correlated with each other. An all axes of parallel coordinates are reordered according to the current reordering criteria(A). (Data shown is from the Car dataset (7 dimensions × 406 samples)).

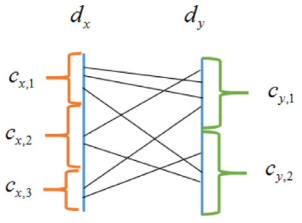

Fig. 3. Definition of probability distribution.

The joint probability distribution of dimension d_x and d_y:

$$p_{xy}(i,j) = p_x(i) \times p(j|i), i = 1,2,3; j = 1,2 \tag{5}$$

$$p(j|i) = \frac{num_{i,j}}{num_{x,i}}, i = 1,2,3; j = 1,2 \tag{6}$$

where $num_{i,j}$ represents the number of data points within both the cluster $c_{x,i}$ and $c_{y,j}$.

3. Mutual information correlation analysis

Let the probability distribution of random variable X be: $P(X = x_i) = p_i, i = 1,2,\ldots,n$. Its information entropy is:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{7}$$

Let the joint probability distribution of random variable (X, Y) be p_{ij}. Their information entropy is:

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log p_{ij} \tag{8}$$

Let the marginal probability distribution of random variables X and Y be $p_{i\bullet}$ and $p_{\bullet j}$. When Y is known the random variable X's conditional entropy is:

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{\bullet j}} \tag{9}$$

Similarly, when X is known the random variable Y's conditional entropy can be defined as:

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{p_{i\bullet}} \tag{10}$$

Finally, let $I(X,Y) = H(X) - H(X|Y)$ or $I(X,Y) = H(Y) - H(Y|X)$ be the criteria to measure the correlation between variables X and Y. The method is called mutual information. A large value of $I(X, Y)$ means a strong correlation between variables X and Y. In addition, we use the conditional information entropy $H(X|Y)$ or $H(Y|X)$ to judge the dependency between variables of X and Y. And $H(X|Y)$ represents the degree of X dependent on Y. $H(Y|X)$ represents the degree of Y dependent on X.

4. Dimension reordering

Finally, all dimensions can be reordered by the correlation and dependency of dimensions. The dimension reordering based on correlation is same as the linear method. As for the dimension reordering based on dependency, we take $depend(i)$ and $depend(j)$ as the criteria (Eq. (11) and Eq. (12)).

$$depend(i) = \sum_{j=1}^{l} H(d_i|d_j), i = 1,2,\ldots,l, i \neq j \tag{11}$$

$$depend(j) = \sum_{i=1}^{l} H(d_i|d_j), j = 1,2,\ldots,l, i \neq j \tag{12}$$

where d_i and d_j represent the dimensions in the multidimensional data. And l represents the number of dimensions. $depend(i)$ represents the degree of variable d_i dependent on the other variables

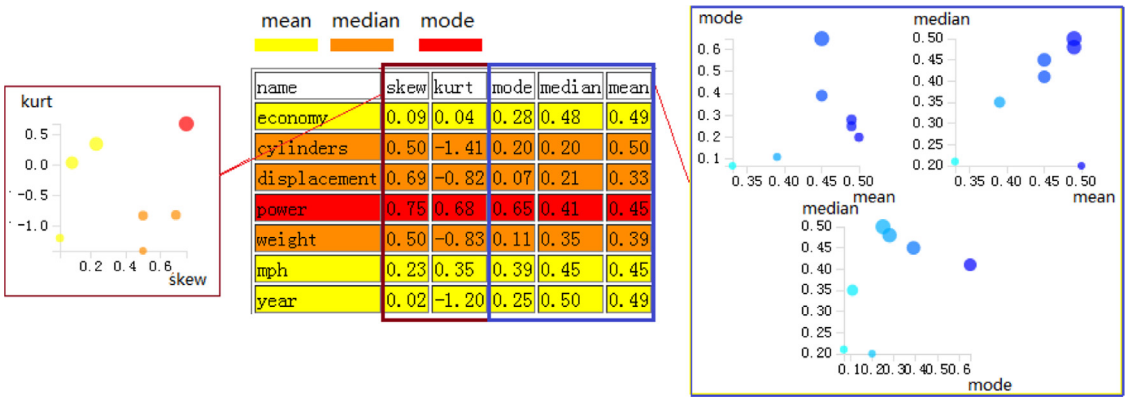

Fig. 4. All dimensions are classified into three groups according to "skew" and "kurt" (left and middle). And they are marked by colors of yellow, orange and red. Correspondingly, the three parameters of mean, median and mode are shown by scatterplots(right). And the values of X-axis are color-coded and the values of Y-axis are mapped by the size of circle. A dark color and a big circle means a high value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

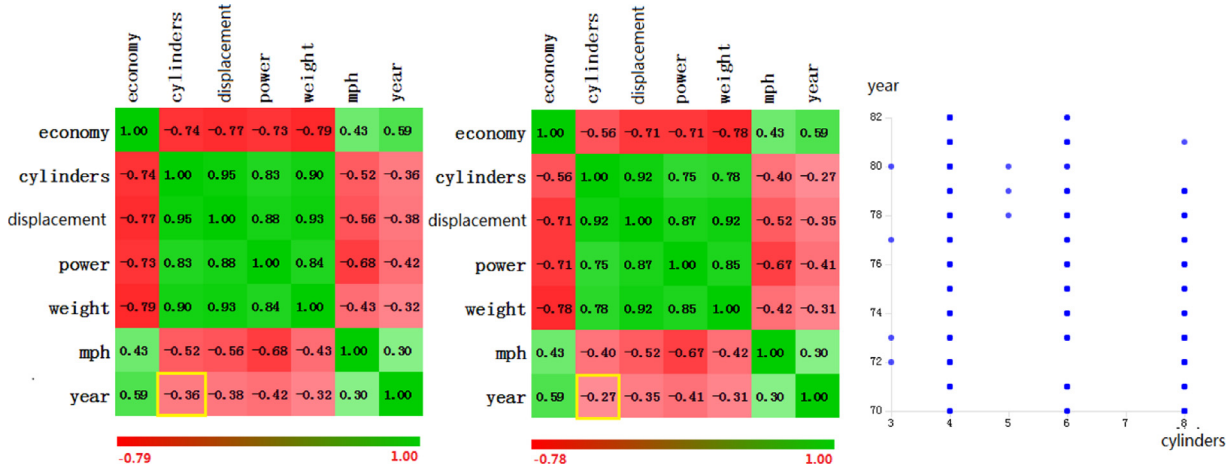

Fig. 5. The correlations between all dimensions are obtained by the traditional (left) and improved (middle) Pearson coefficient. And the dimensions of "cylinders" and "year" are displayed in the scatterplot (right).

of d_j. *depend*(j) represents the degree of variable d_j is dependent on the other variables of d_i. A small value of *depend*(i) or *depend*(j) means a strong dependency in the dataset. So, we rearrange all dimensions according to the values of *depend*(i) or *depend*(j) from small to large.

4. Case study

4.1. Car dataset

The Car dataset is widely used in the multidimensional visualization analysis. It includes 7 dimensions and 406 samples. And there are apparent data relationships between dimensions. So, we take it as one of our cases. First we classify all dimensions into three groups according to the features of skewness and kurtosis (Fig. 4). In the left and middle of Fig. 4, we can see that the group "yellow" contains the dimensions of "economy","mph" and "year". The group "red" only contains the dimension of "power". The other dimensions belong to the group "orange". And the parameters of mean, median, mode that they need in the Pearson coefficient calculation are displayed in the table (middle of Fig. 4) and scatterplots (right of Fig. 4).

Then, we select the corresponding parameters of mean, median and mode for each group of dimensions to calculate the correlations. The results are shown in the middle of Fig. 5. Compared with the correlations obtained by the traditional Pearson coefficient (left of Fig. 5), some of the results from our method are more accurate. For example, in the scatterpot of "year" *vs.* "cylinders" (right of Fig. 5), we can see that they are almost not correlated with each other. So, their correlation coefficient should be small. And the values are −0.36 and −0.27 respectively in the left and middle of Fig. 5. By using our method, some new patterns can be found out.

And, all dimensions are reordered according to the correlations and displayed by the parallel coordinates. The users are allowed to select a primary dimension according to the features: normality of distribution, uniformity of distribution and outliers. In Fig. 6, we can see that all dimensions with three features are ranked by the normality of distribution and shown in the table (Fig. 6A). In the first row of the table is the dimension of "economy". From its density curve (Fig. 6B) we can see that it is closest to the normal distribution. So, the dimension of "economy" is selected as the primary dimension to be ranked in the first place. Next, we can rearrange the other dimensions according to their pairwise correlations (Fig. 6C). The result of reordering is shown in the parallel coordinates (Fig. 6D). Correspondingly, the dimensions in the correlation map (Fig. 6C) are also reordered. It can be seen in the view C and view D of Fig. 6 that the correlations between the adjacent dimensions are strong.

Next, we analyze the dimensions' non-linear correlations and dependencies by the mutual information correlation analysis and clustering. We can see that the correlation graph is shown in Fig. 7 (A) and the conditional information entropy graph is dis-

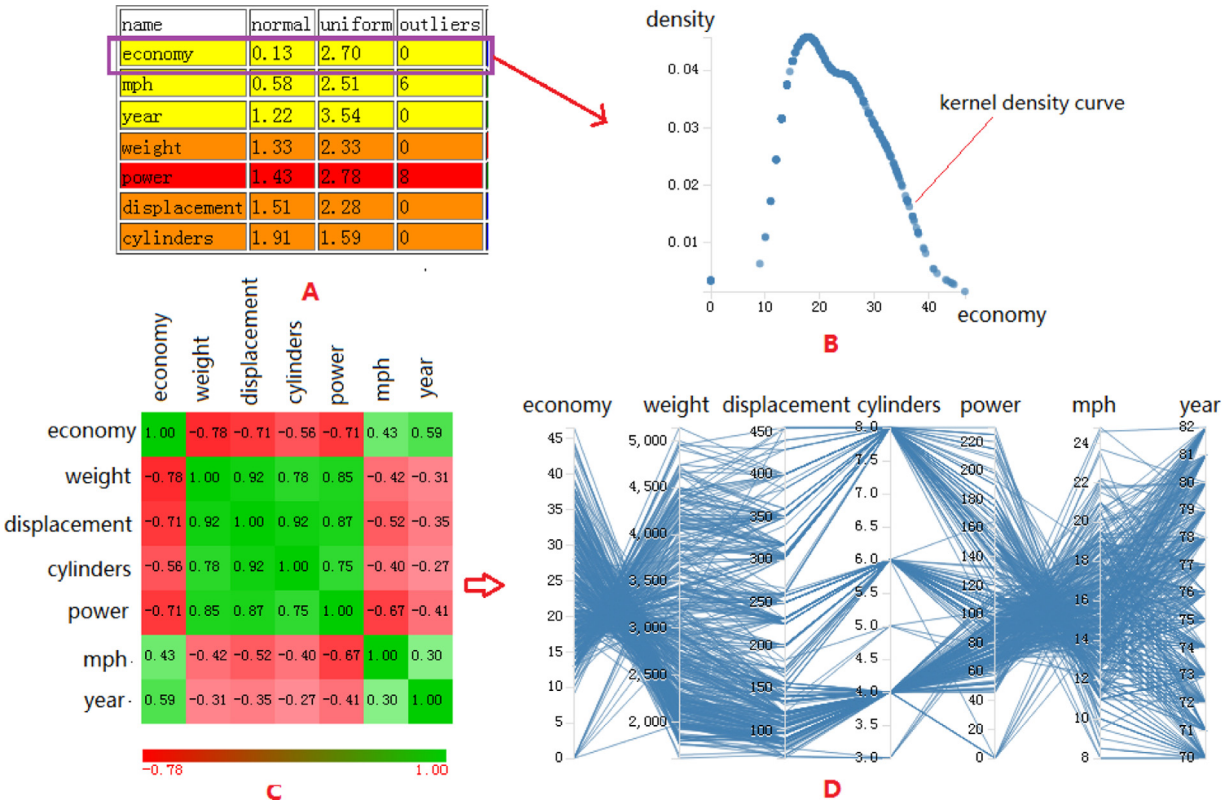

Fig. 6. All dimensions are ranked by the normality of distribution and shown in the table(A). The density curve of "economy"(B). Correlations from our improved Pearson coefficient method are shown in the map graph(C). All dimensions are reordered and shown by the parallel coordinates(D).

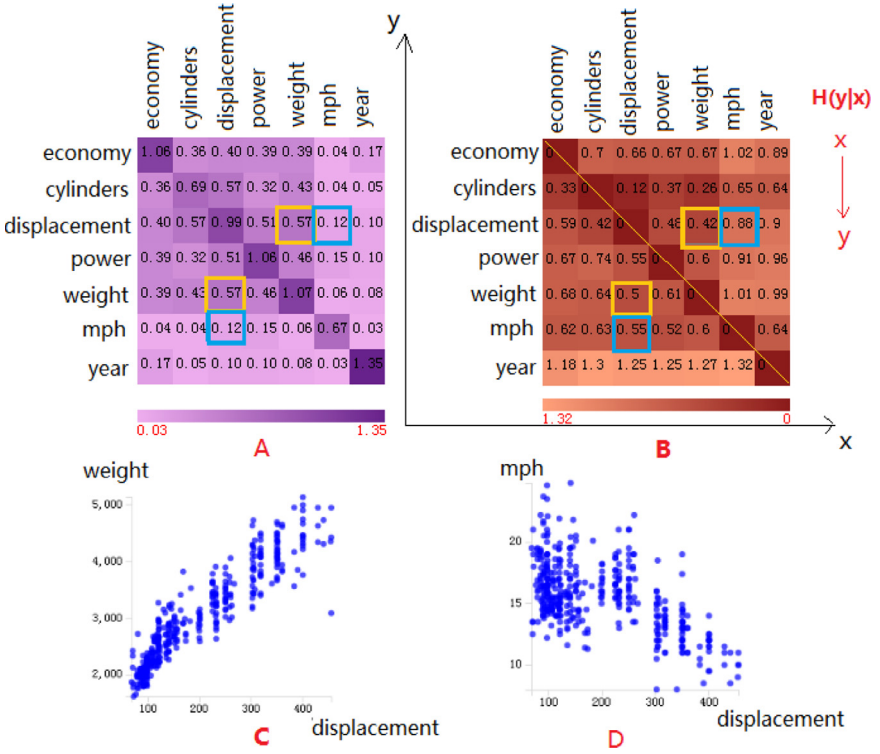

Fig. 7. Correlations(A) and conditional information entropy(B) from non-linear correlation analysis. And they are color-coded. A dark color means a strong correlation and a small conditional information entropy. Scatterplots of "displacement" *vs.* "weight"(C) and "displacement" *vs.* "mph"(D).

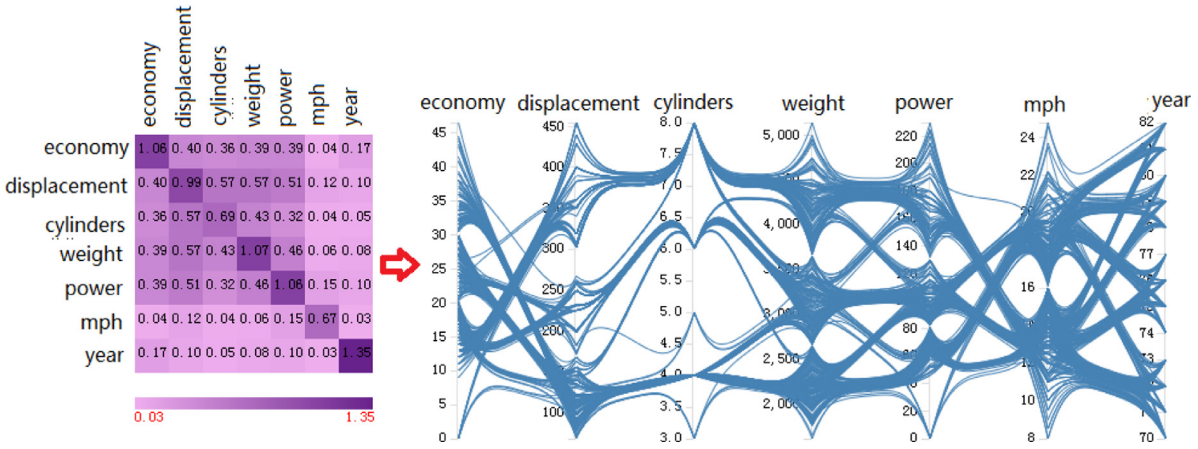

Fig. 8. Reordering by correlation. Non-linear correlations are shown in the map graph (left) and the results of reordering are displayed by the parallel coordinates (right).

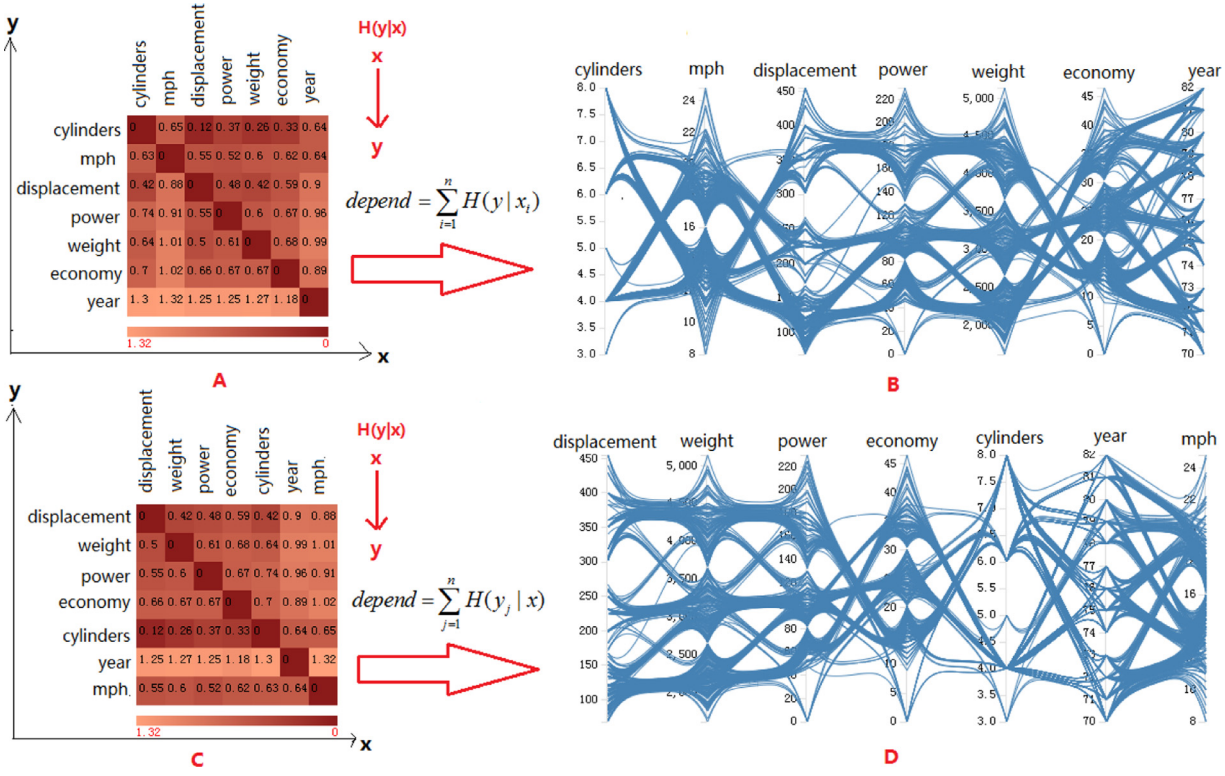

Fig. 9. Reordering by dependency. All dimensions are reordered according to the values of *depend* from small to large. Conditional information entropy are shown in the map graph(A,C) and the results of reordering are displayed by the parallel coordinates(B,D).

played in Fig. 7(B). The conditional information entropy ($H(y|x)$, x represents the dimension in the X-axis and y represents the dimension in the Y-axis) means the degree of variable y dependent on variable x. A large correlation coefficient means a strong correlation, but a large conditional information entropy means a weak dependency. From the scatterplot of "displacement" *vs.* "weight"(Fig. 7C) we can know that they have a strong correlation. So, their correlation coefficient in Fig. 7(A) is large. And their information entropy in Fig. 7(B) are both small. It means that "displacement" and "weight" are strongly dependent on each other. However, From the scatterplot of "displacement" *vs.* "mph" (Fig. 7D) we can know that they have a weak correlation. So, their correlation coefficient in Fig. 7(A) is small. But in Fig. 7(B) we can see that, the information entropy H("*displacement*"|"*mph*") in the triangular is large, but H("*mph*"|"*displacement*") is small in the triangle. It means that "mph" is strongly dependent on "displacement", but "displacement" is weakly dependent on "mph". So, when two variables are strongly dependent on each other, they have a strong correlation. Finally, we take the correlations and dependencies as the reordering criteria to rearrange the dimensions' sequence. It can be seen in Fig. 8 that all dimensions are reordered according to the non-linear correlations by selecting "economy" which is close to a normal distribution as the primary dimension. The reordering result is similar with the result of linear method.

In Fig. 9, we can see that all dimensions are reordered according to the dependencies of dimensions. In view A and view B of Fig. 9, all dimensions are reordered by $depend = \sum_{i=1}^{n} H(y|x_i)$($n$ is the number of dimensions.y represents the dimension in the Y-axis and x represents the dimension in the X-axis). The dimensions in front of Fig. 9(B) represent that they are strongly dependent on the other dimensions. In view C and view D of Fig. 9, all dimen-

Fig. 10. Correlations from traditional Pearson Coefficient(A), correlations from improved Pearson Coefficient(B) and correlations from non-linear analysis method(C). Scatterplots of "CRIM" *vs*. "ZN"(D), "B" *vs*. "ZN"(E) and "TAX" *vs*. "ZN"(F).

sions are reordered by $depend = \sum_{j=1}^{n} H(y_j|x)$. The dimensions in front of Fig. 9(D) represent that they are strongly dependent on the other dimensions. From these two parallel coordinates shown in view B and view D of Fig. 9 we can know that the dimension of "year" is an independent variable which is unimportant in the Car dataset. Because it is weakly dependent on the other dimensions and is weakly dependented by the other dimensions. And the dimensions of "cylinders" and "mph" are dependent variables. Because they are strongly dependent on the other dimensions, but are weakly dependented by the other dimensions. As for the dimensions of "displacement", "weight" and "power", they are strongly dependent on the other dimensions and are strongly dependented by the other dimensions. So, these three dimensions are important in the Car dataset.

4.2. House dataset

The well known 'Boston Neighborhood Housing Prices' dataset [25] is study as another case. This dataset contains information gathered by the U.S Census Service to understand the relation between housing prices and other factors in the area of Boston, Massachusetts. It consists of 506 samples and 14 dimensions. We use 12 of the 14 dimensions to analyze. They include: 'per capita crime rate by town'(CRIM), 'proportion of residential land zoned for lots over 25,000 sq.ft.'(ZN), 'proportion of non-retail business acres per town'(INDUS), 'nitric oxides concentration (parts per 10 million)'(NOX), 'average number of rooms per dwelling'(RM), 'proportion of owner-occupied units built prior to1940' (AGE), 'weighted distances to five Boston employment centres' (DIS), 'full-value property-tax rate per $10,000' (TAX), 'pupil-teacher ratio by town' (PTRATIO), '$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town'(B), '% lower status of the population' (LSTAT) and 'Median value of owner-occupied homes in $1000's' (MEDV).

We use the linear and non-linear correlation analysis methods to calculate the correlations between dimensions of House dataset. The results are shown in Fig. 10. Fig. 10(A) are the results of traditional Pearson coefficient. Fig. 10(B) are the results of our improved Pearson coefficient. And Fig. 10(C) are the results of our non-linear analysis method. It can be seen that the correlation between the dimension of "CRIM" and "ZN" are respectively −0.20, 0.00, 0.01 in Fig. 10 (A), (B), (C). In the scatterplot of "ZN" *vs* "CRIM" (Fig. 10D), we can know that they are almost not related with each other. There is only one point to connect them, so their correlation coefficient should be very small. Obviously, the results of our improved Pearson coefficient and non-linear analysis method are better. And the non-linear method is more precisely. In the scatterplot of "ZN" *vs* "B" (Fig. 10E), we can see it clearly that the dimension of "B" is weakly and negatively related with the dimension of "ZN". So their correlation coefficient should be a small and negative value. And they are respectively 0.18,−0.02,0.04 in Fig. 10 (A),(B),(C). So, we can see that the values in Fig. 10 (B) and (C) are more appropriate. And Fig. 10 (B) is better. For the correlation between the dimension "ZN" and "TAX" (Fig. 10F), we can also find that their correlations obtained by our improved Pearson coefficient and non-linear analysis method are more precisely.

Our non-linear correlation analysis method can also be used to analyze the dependencies of the variables. It can be seen that Fig. 11(A) is the correlation graph and Fig. 11(B) is the conditional information entropy ($H(y|x)$) graph. From the scatterplot of "AGE" *vs* "CRIM" (Fig. 11C) we can know that they are almost not correlated with each other. So the correlation coefficient in Fig. 11(A) is 0.02 which is very small. However, from Fig. 11(B) we can see that the entropy of H("*CRIM*""*AGE*") is small and the entropy of H("*AGE*""*CRIM*")is very large. It means that the dimension of "AGE" is almost not dependent on the dimension of "CRIM". But the dimension of "CRIM" is strongly dependent on the dimension of "AGE". Next, let see the scatterplot of "MEDV" *vs* "LSTAT" (Fig. 11D), they are weakly correlated with each other. So the correlation coefficient in Fig. 11(A) is 0.4 which is not too small. And their entropy are both not too large. That is to say that

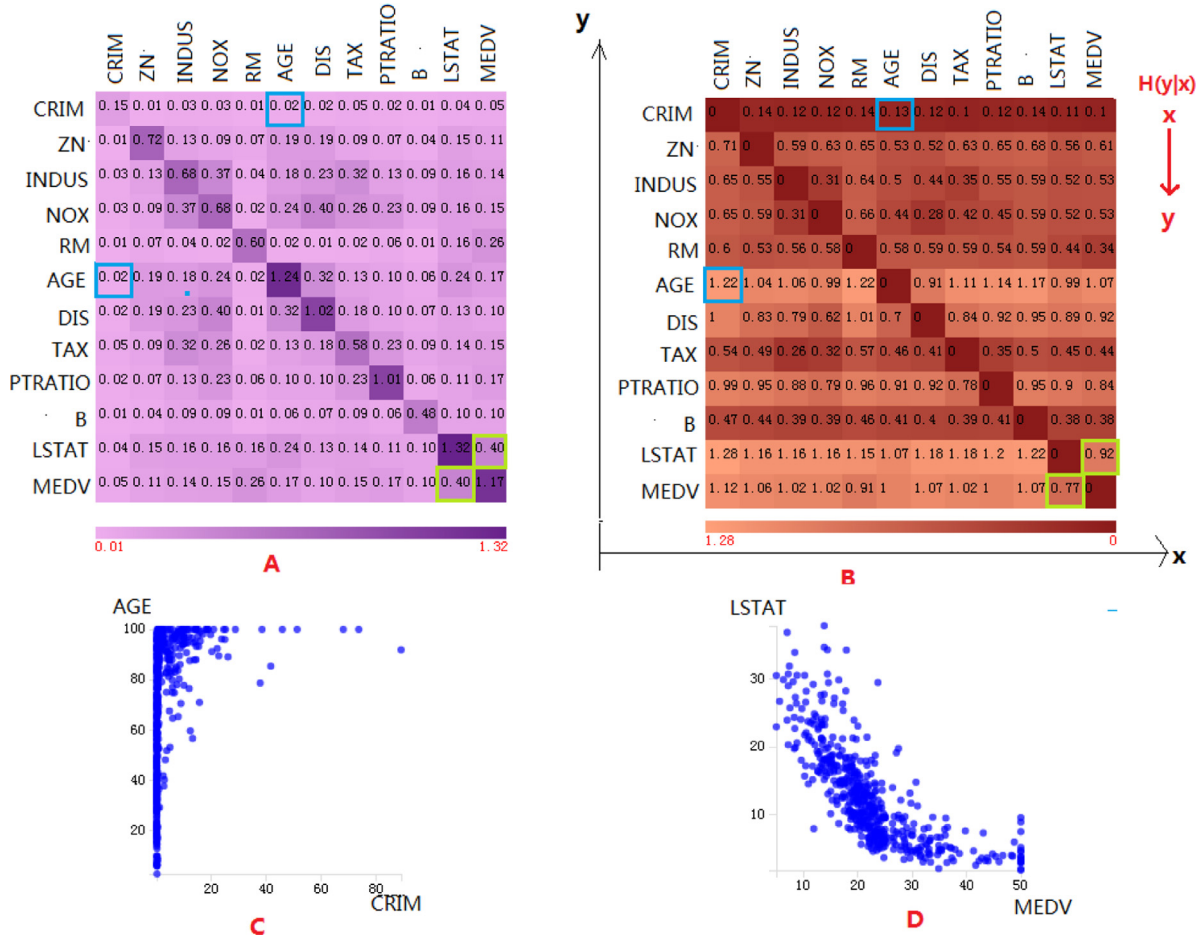

Fig. 11. Correlations(A) and conditional information entropy(B) from the non-linear analysis method. Scatter plots of "AGE" *vs* "CRIM"(C) and "LSTAT" *vs* "MEDV"(D).

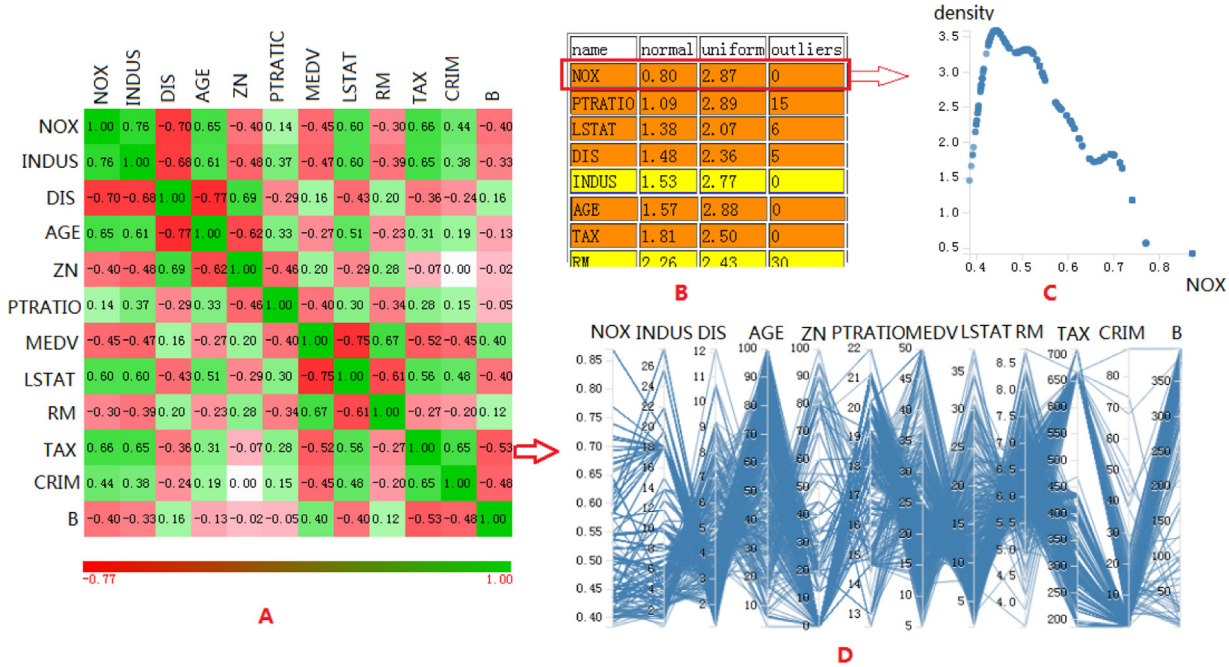

Fig. 12. Reordering by correlation. Correlations from improved Pearson coefficient(A). Features are displayed in the table(B). The density curve of "NOX"(C). All dimensions are shown by the parallel coordinates(D).

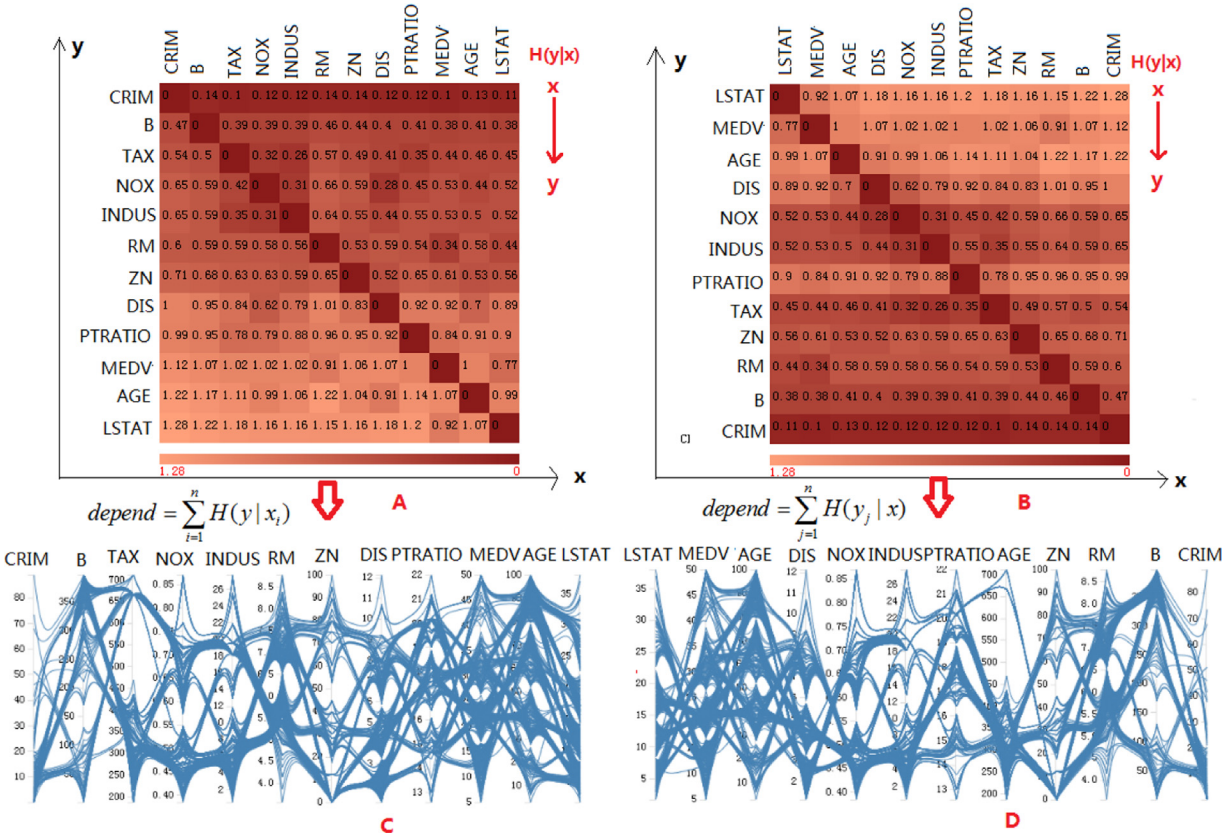

Fig. 13. Reordering by dependency. All dimensions are reordered according to the values of *depend* from small to large. Conditional information entropy are shown in the map graph(A,B) and the results of reordering are displayed by the parallel coordinates(C,D).

they are weakly dependent on each other. So, when two variables are weakly dependent on each other, they have a weak correlation.

In Fig. 12, we can see that all dimensions are reordered according to the linear correlations by selecting "NOX" which is close to a normal distribution (Fig. 12B and C) as the primary dimension. The reordering results are shown in the map graph (Fig. 12A) and parallel coordinates (Fig. 12D). However, for the House dataset have no apparent correlations between dimensions, we cannot easily find the data relationships between dimensions from the parallel coordinates.

In Fig. 13, we can see that all dimensions are reordered by the dependencies of variables in the House dataset. In view A and view C of Fig. 13, all dimensions are reordered by $depend = \sum_{i=1}^{n} H(y|x_i)$. The dimensions in front of Fig. 13(C) represent that they are strongly dependent on the other dimensions. In view B and view D of Fig. 13, all dimensions are reordered by $depend = \sum_{j=1}^{n} H(y_j|x)$. The dimensions in front of Fig. 13(D) represent that they are strongly dependente on the other dimensions. From the results of reordering we can know that all dimensions have some one-way dependent relationship. The dimensions of "CRIM" and "B" are strongly dependent on the other dimensions, but are weakly dependented by the other dimensions. The dimensions of "LSTAT", "AGE" and "MEDV" are weakly dependent on the other dimensions, but are strongly dependented by the other dimensions.

4.3. Discussion

From the analysis results for the datasets of Car and House, we can know that our linear analysis method of improved Pearson coefficient and non-linear analysis method based on the mutual information correlation analysis are better. When the variables are extremely skewed distribution, our linear and non-linear methods can both obtain the precise correlations between the variables. In addition, our linear analysis method can judge the correlation is negative or positive. And our non-linear correlation analysis method can also find the dependencies between dimensions. For the dataset with apparent data relationships, these two method can both get a good reordering view according to the correlations between dimensions. However, when the dataset has no apparent data relationships between dimensions, we cannot find the valuable information from the reordering view based on correlation. At this time, our linear analysis method is invalid. However, our non-linear method can help to analyze the dependencies of all dimensions. And several important dimensions can be found from the reordering view based on dependency. Besides, we also find that the correlation and dependency have a certain relationship. When two variables are strongly dependent on each other, they may have a strong correlation. If one of them is weakly dependent on another one, they may have a weak correlation.

5. Conclusion

In this paper, we have introduced two correlation analysis methods to detect the linear and non-linear relationships between variables in the multidimensional data. Our linear method based on dimension grouping and Pearson correlation coefficient can detect the linear correlations between two variables with any distributions. Our non-linear method cannot only detect the correlations between variables with any distributions, but also can find their dependencies which show the relationships between dependent and independent variables. It is not influenced by the noise points. We also presented two interactive frameworks that enable correlation analysis and visual association mining for the multidimen-

sional data. Our relationship and feature graphs can guide users to reorder all dimensions in a good view. After reordering, the parallel coordinates can show the data relationships more clearly.

However, a present limitation is that correlation can show only pairwise relationship of two single variables, but strong relationships may exist between two sets of variables. In the future the method of canonical correlation coefficient should be considered in the multidimensional visualization.

In this paper, we only analyze the relationships between the whole variables. The strong relationships do not exist between the two whole variables. But it may exist between the subspace of the two variables. In the future, we should focus on the correlations between variables within the subsets of the multidimensional data.

References

[1] A. Gisbrecht, B. Hammer, Data visualization by nonlinear dimensionality reduction, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 5 (2) (2015) 51–73.
[2] F. Kemp, Applied multiple regression/correlation analysis for the behavioral sciences, J. R. Stat. Soc. 52 (4) (2003) 227–229.
[3] W.K. Hardle, L. Simar, Canonical Correlation Analysis, Springer Berlin Heidelberg, 2012.
[4] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, M.V. Gilean, P.J. Turnbaugh, E.S. Lander, M. Michael, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011). 1518–24.
[5] P.J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson Correlation Coefficient, Springer Berlin Heidelberg, 2009.
[6] K.T. Mcdonnell, Z. Erez, M. Klaus, Visual correlation analysis of numerical and categorical data on the correlation map, IEEE Trans. Vis. Comput. Graph. 21 (2) (2015) 289–303.
[7] J. Seo, B. Shneiderman, A rank-by-feature framework for interactive exploration of multidimensional data, Inf. Vis. 4 (2) (2005) 96–113.
[8] C. Spearman, The proof and measurement of association between two things, Int. J. Epidemiol. 39 (5) (2010). 1137–50.
[9] L.A. Goodman, W.H. Kruskal, Measures of association for cross classifications. ii: Further discussion and references, J. Am. Stat. Assoc. 54 (285) (1959) 123–163.
[10] K. Baba, R. Shibata, M. Sibuya, Partial correlation and conditional correlation as measures of conditional independence, Aust. N. Z. J. Stat. 46 (4) (2004) 657C664.
[11] Y. Zhou, L.U. Xiao-Wei, C.T. Cheng, Parallel computing method of canonical correlation analysis for high-dimensional data streams in irregular streams, J. Softw. 23 (5) (2012) 1053–1072.
[12] J.Y. Liang, C.J. Feng, P. Song, A survey on correlation analysis of big data, Chin. J. Comput. (2016).
[13] G.J. Szekely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Stat. 35 (6) (2008) 2769–2794.
[14] G.J. Szekely, M.L. Rizzo, The distance correlation *t*-test of independence in high dimension, J. Multivar. Anal. 117 (3) (2013) 193–213.
[15] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Appl. Phys. Lett. 3 (6) (2003) 1157–1182.
[16] S.J. Fernstad, J. Shaw, J. Johansson, Quality-based guidance for exploratory dimensionality reduction, Inf. Vis. 12 (1) (2013) 44–64.
[17] T. Cagatay, F. Peter, H. Helwig, Brushing dimensions–a dual visual analysis model for high-dimensional data, Vis. Comput. Graph. IEEE Trans. 17 (12) (2011). 2591–9.
[18] C. Turkay, A. Lundervold, A.J. Lundervold, H. Hauser, Representative factor generation for the interactive visual analysis of high-dimensional data, IEEE Trans. Vis. Comput. Graph. 18 (12) (2012) 2621–2630.
[19] W. Peng, M.O. Ward, E. Rundensteiner, et al., Clutter reduction in multi-dimensional data visualization using dimension reordering, in: Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on, IEEE, 2004, pp. 89–96.
[20] A.O. Artero, M.C.F. De Oliveira, H. Levkowitz, Enhanced high dimensional data visualization through dimension reduction and attribute arrangement, in: Proceedings of the Conference on Information Visualization, 2006, pp. 707–712.
[21] B.J. Ferdosi, J.B.T.M. Roerdink, Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis, in: Eurographics / IEEE - Vgtc Conference on Visualization, 2011, pp. 1121–1130.
[22] X. Zhao, A. Kaufman, Structure revealing techniques based on parallel coordinates plot, Visual Comput. 28 (6–8) (2012) 541–551.
[23] T. Weinkauf, H.P. Seidel, A. Oulasvirta, M. Bachynskyi, G. Palmas, An edge-bundling layout for interactive parallel coordinates, in: IEEE Pacific Visualization Symposium, 2014, pp. 57–64.
[24] Y. Chen, H.E. Zhong-Shi, M. Huang, The study on improved k-means algorithm, Manuf. Autom. (2012).
[25] D. Harrison, D.L. Rubinfeld, Hedonic housing prices and the demand for clean air, J. Environ. Econ. Manage. 5 (1) (1978) 81–102.