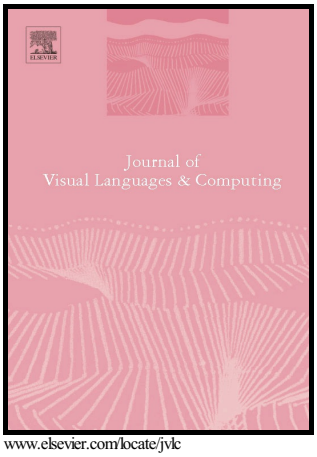
Author's Accepted Manuscript

Two Axes Re-ordering Methods in Parallel Coordinates Plots

Liang Fu Lu, Mao Lin Huang, Jinson Zhang



www.elsevier.com/locate/jvlc

PII: S1045-926X(15)30037-9
DOI: http://dx.doi.org/10.1016/j.jvlc.2015.12.001
Reference: YJVLC742

To appear in: *Journal of Visual Language and Computing*

Received date: 19 November 2015
Accepted date: 1 December 2015

Cite this article as: Liang Fu Lu, Mao Lin Huang and Jinson Zhang, Two Axes Re-ordering Methods in Parallel Coordinates Plots, *Journal of Visual Language and Computing,* http://dx.doi.org/10.1016/j.jvlc.2015.12.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain

ACCEPTED MANUSCRIPT

Two Axes Re-ordering Methods in Parallel Coordinates Plots

Liang Fu Lu [1,2] **, Mao Lin Huang** [3,2,*] **, Jinson Zhang** [2]

[1.] Mathematics Department, Tianjin University, Tianjin, P. R. China.
[2.] Faculty of Engineering and IT, University of Technology, Sydney, Australia.
[3.] School of Computer Software, Tianjin University, Tianjin, P. R. China.
*Corresponding Author E-mail: Mao.Huang@uts.edu.au

Abstract— Visualization and interaction of multidimensional data are challenges in visual data analytics, which requires optimized solutions to integrate the display, exploration and analytical reasionning of data into one visual pipeline for human-centered data analysis and interpretation. Though it is considered to be one of the most popular techniques for visualization and analysis of multidimensional data, parallel coordinate visualization is also suffered from the visual clutter problem as well as the computational complexity problem, same as other visualization methods which visual clutter occurs with the volume of data needs to be visualized increasing. One straightforward way to address these problems is to change the ordering of axis to reach the minimal number of visual clutters. However, the optimization of the ordering of axes is actually a NP-complete problem. In this paper, two axes re-ordering methods are proposed in parallel coordinates visualization: 1) a contribution-based method and 2) a similarity-based method.

The contribution-based re-ordering method is mainly based on the singular value decomposition (SVD) algorithm. It can not only provide users with the mathmetical theory for the selection of the first remarkable axis, but also help with visualizing detailed structure of the data according to the contribution of each data dimension. This approach reduces the computational complexity greatly in comparison with other re-ordering methods. While a similarity-based re-ordering method is based on the combination of nonlinear correlation coefficient (NCC) and SVD algorithms. By using this approach, axes are re-ordered in line with the degree of similarities among them. It is much more rational, exact and systemic than other re-ordering methods, including those based on Pearson's correlation coefficient (PCC). Meanwhile, the paper also proposes a measurement of contribution rate of each dimension to reveal the property hidden in the dataset. At last, the rationale and effectiveness of these approaches are demonstrated through case studies. For example, the patterns of Smurf and Neptune attacks hidden in KDD 1999 dataset are visualized in parallel coordinates using contribution-based re-ordering method; NCC re-ordering method can enlarge the mean crossing angles and reduce the amount of polylines between the neighboring axes.

Index Terms—Multidimensional data visualization, visual analytics, parallel coordinates, axes re-ordering, singular value decomposition, nonlinear correlation coefficient.

Accepted manuscript

INTRODUCTION

The rapid growth of data communication through the Internet and World Wide Web has led to vast amounts of information available online. In addition, business and government organizations create large amounts of data contain both structured and unstructured information which needs to be processed, analyzed, and visualized. Therefore, multi-dimensional data analysis is becoming commonplace as the number of applications increases, such as statistical and demographic computation, digital libraries and so on. However, traditional visualization techniques for these data sets usually require dimensionality reduction or selection to generate the meaningful visual representations. Dimensionality reduction, as Sara Johansson et. al [1] pointed out, is always employed prior to visualization for dealing with the data with a large number of attributes. Currently, many dimensionality reduction methods are available to preserve the information inside the data as much as they can remove some less relevant data items or attributes from the original dataset. While dimension selection is mainly referred to dimension re-ordering which means the corresponding axes of the dimension in a parallel coordinate visualization can be positioned in accordance to some effective rules such as similarity of dimensions to achieve good visual structures and patterns. This paper focuses on the dimension re-ordering instead of dimension reduction to address the problems of visual clutter and computational complexity.

In 1998, Mihael Ankerst et al. [2] presented the method of using the similarity of dimensions to improve the quality of visualization of multidimensional data, that is using global and partial similarities for one or two-dimensional visualization methods. Pearson's Correlation Coefficient (PCC) is one of the most common methods used for measuring similarity between two dimensions. PCC can be used for dimension reduction, clutter reduction and clustering in data visualization. At the same time, it has also been proved that the PCC based re-ordering problem is a NP-complete problem. Therefore, many researchers applied heuristic algorithms to figure out an optimal order of axes (or dimensions) in the multi-dimensional data visualization.

This paper proposes two rational dimension re-ordering methods to support the visual analytics in parallel coordinates. And those two methods can also be easily applied to other visualization techniques.

Firstly, method to find out the contribution of each dimension in the dataset is developed on the basis of the Singular Value Decomposition (SVD). After the calculation of contribution rates, axes (or dimensions) are re-ordered and visualized as parallel coordinates from left to right according to the degree of their significances. Though the traditional heuristic algorithms can

ACCEPTED MANUSCRIPT

optimize the order of axes for one- or two-dimensional visualizations, most studies have not been done to a deeper investigation on how to determine the first dimension (the most significant dimension) in multi-dimensional data visualizations. The first dimension always attracts much more user's attention than the others. Therefore, the one with the highest contribution rate can be considered as the first dimension to simplify the traditional similarity-based re-ordering methods and the one to find out the optimal order of parallel axes in a short time period.

Secondly, Pearson's Correlation Coefficient method is applied into the further investigation on axes re-ordering. As a correlation metric between each pair of dimensions in the dataset, PCC is available for characterizing linear systems statistically. Inspired by PCC, a similarity-based re-ordering method is presented in parallel coordinates which is based on the combination of a Nonlinear Correlation Coefficient (NCC) and the SVD algorithms. NCC is sensitive to any relationship, not just the linear dependence [26]. It is more rational than the current PCC method in theory and it can improve the quality of multi-dimensional visualizations significantly in terms of effectiveness and exactness. In our experiments, the effectiveness of the new method can be proved by visualizing the patterns and enlarging the mean crossing angles for better visual representation.

The paper is organized as follows. The current situation and background of researching on similarity measure and dimension reordering in high-dimensional data visualization is introduced in Section 1. And next two dimension reordering approaches are stated in detail in section 2. While the experimental evaluation for our new ideas as well as the effectiveness of our methods in parallel coordinates visualization are further elaborated and proved in section 3. Section 4, the last part of the paper, is designed to make the conclusions and look forward to future work.

1 Related work

An effective way to improve the quality of multi-dimensional visualizations is to re-order the dimension axes in parallel coordinates based on similarity of data attributes. In this section, paper begins to summarise the previous researches finished in the area of high-dimensional visualization.

Parallel coordinates[3, 4], scatter plot matrix[5], table lens[6] and pixel-oriented display[7] et al. are well-known and accepted as visualization techniques for high-dimensional data sets.

Similarity measurement as one aspect of quality metrics in high-dimensional data visualization has been addressed in the past few years [1, 8-11]. It is worth noting that Enrico Bertini et al [8] systematically presented an overview of quality metrics in many visualization techniques through a literature review of nearly 20 papers and considered correlation between two or more dimensions to be the main characteristic of similarity measurement. Sara Johansson[1] introduced a weighted quality metrics to their task-dependent and user-controlled dimensionality reduction system, where small correlation values are ignored to reduce the dataset that preserves the important structures within the original dataset. Andrada Tatu et al. proposed similarity-based function for classified and unclassified data based on Hough Space transform on the resulting image of parallel coordinates [10, 11]. Aritra Dasgupta et al. [9] introduced binned data model and branch-and-bound algorithm as the screen-space metrics for parallel coordinates to reduce the computations and find the optimal order of axes.

To enhance the high-dimensional data visualization, some studies on dimension reordering have been done to find good axes layouts in visualization techniques both in one- or two-dimensional arrangement[1, 8-10, 12] [13, 14] [2, 15-18, 32]. Mihael Ankerst et al. [2] defined similarity measures which determined the partial or global similarity of dimensions and argued that the reordering based on similarity could reduce visual clutter and do some help in visual clustering. Wei Peng et al.[15] introduced the definition of the visual clutter in parallel coordinates as the proportion of outliers against the total number of data points and they tried to use the exhaustive algorithm to find the optimal axes order for minimizing the member of edge crossings (or visual clutter). As mentioned in [16], the computational cost $o(n \cdot n!)$ hampers applications of this technique to large high dimensional data sets. Almir Olivette Artero et al. [16] introduced the dimension configuration arrangement based on similarity to alleviate clutter in visualizations of high-dimensional data. They proposed a method called SBAA (Similarity-Based Attribute Arrangement), which is a straightforward variation of the Nearest Neighbor Heuristic method, to deal with both dimension ordering and dimensionality reduction. Other studies have been done on the dimension reordering based on the similarity[11, 17, 18] [19] [20]. Michael Friendly et al.[17] designed a framework for ordering information, including arrangement of variables. However, the arrangement of variables is decided mainly according to the users' desired visual effects. J. Yang et al. [18] established a hierarchical tree structure over the attributes, where the similar attributes were positioned near each other. Diansheng Guo [19] developed a hierarchical clustering method based on comparing and sorting dimensions by using the maximum conditional entropy. Georgia Albuquerque et al. [20] introduced the quality measures to define the placement of the dimensions for Radviz and also to appraise the information content of pixel and Table Lens visualizations.

The most of recent dimension reordering methods are established on the basis of Pearson's Correlation Coefficient. From the statistics point of view, PCC is taken as a method for measuring the linear correlation between the two random variables. Therefore, it is irrational to reorder the dimensions in similarity only depending on the calculation of PCC. Though Pargnostics, proposed by Aritra Dasgupta et al. in[1], is the most similar with our approach, the probability and joint probability during the computational process are both denoted as their special axis histograms, which lack the support by mathematical theories.

Accepted manuscript

ACCEPTED MANUSCRIPT

Moreover, it can be seen from the definition of the mutual information that does not range in a definite closed interval as the correlation coefficient does, which ranges in $[-1,1]$.

Hence, it is of great importance that a rational and useful method should be proposed for correlation analysis among the dimensions for conveying better visual structures and patterns. In this paper, we propose two methods for dimensions reordering in parallel coordinates: contribution-based and similarity-based reordering methods. The contribution-based reordering, based on the SVD algorithm, which not only can provide theoretical support for the selection of the first dimension but also can visualize the clear and detailed structure of the dataset with the contribution of each dimension. Operated with it, much less computational complexity can be reduced and much more time can be saved than did with any other traditional reordering methods. The similarity-based reordering method is based on the combination of NCC and SVD algorithms, where dimensions are reordered in line with the degree of correlations among the dimensions. It is more rational, exact and systemic than other traditional methods.

2 METHODOLOGY

The ordering of dimension has large impact on how easily we can perceive different structures in the data[2]. Completely different displays and conclusions may be obtained if we interactively switch between different dimension reordering. How to reordering the dimensions in high-dimensional data sets meaningfully is one of the most significant problems of the researches on quality metrics in data visualization due to its influences on the quality of visualization in terms of readability and understandability. In this section, we visualize them in a rational way rather than arrange them only according to the empiricism or good visual effects.

Therefore, we propose a contribution-based reordering method based on SVD algorithm to visualize the data sets in parallel coordinates according to the contribution of each dimension to the whole dataset in section 2.1. Moreover, this method provides some theoretical support on how to determine the first dimension to visualize as well. In section 2.2, we present a similarity-based reordering method based on the combination of SVD and NCC algorithms.

Throughout this section the following notation is used: a dataset D is composed of n dimensions (variables) with m data items for each one. In some cases we need to measure the statistical characters between the two dimensions X and Y, where $X=(x_1,x_2,\cdots,x_n)^T$, $Y=(y_1,y_2,\cdots,y_n)^T$.

2.1 Contribution-based Re-ordering

In the research field of matrix computation, singular value decomposition plays an important role in revealing interesting and attractive algebraic properties, and conveying important geometrical and theoretical insights about transformations. The entries of each matrix obtained by the SVD algorithm have their special physical significances. Here we apply these significances of matrix to our method to measure the contribution of each dimension to the dataset.

For an $m\times n$ matrix D, the singular value decomposition of it is defined as the following form[21]:

$$D=U\Sigma V^* \qquad (1)$$

Where, U and V^* (V^* is the conjugate transpose of V) are $m\times m$ and $n\times n$ unitary matrices respectively. Σ is an $m\times n$ rectangular diagonal matrix with nonnegative real numbers (singular values of D) in order of decreasing magnitude on the diagonal.

There are many properties of SVD for the matrices. For example, the singular values of the matrix D are the square roots of eigenvalues of matrix D^TD; the Euclidean norm of D is equal to the largest singular value and so on. Among these properties, what impressed us most are that the columns of the matrices U and V form the orthonormal basis for the space spanned by the columns and rows of D. For example, in the literature [30], characteristic modes are defined to reconstruct the gene expression patterns based on this property. By combining and analyzing these properties, we can conclude the following property in perspective of the numerical properties for matrix:

Property: The entries of the first column of V in the singular value decomposition, which are denoted as v_{1j}, $j=1,2,\cdots,n$, show the contributions of columns of D to the space spanned by them, i.e. $span\{d_1,d_2,\cdots,d_n\}$, d_i is the ith column of D.

Based on the above property, we can design a contribution-based reordering method according to these entries of the column and visualize the dimensions of the dataset from left to right in parallel coordinates (See Fig. 1 and Fig. 2). From the perspective of data values, this reordering method provides us effective and clear visualization structure of the data. It can help us take deeper insight into the dataset.

It is natural that we can compute the contribution rate of each dimension to the whole dataset using the following possible measure:

$$C_i=\frac{v_{1j}}{\sum_{j=1}^{n}v_{1j}}\times 100\% \qquad (2)$$

Accepted manuscript

ACCEPTED MANUSCRIPT

This approach not only provides us a new reordering method helping us take much more insights into the dataset but also gives rise to the following re-ordering method which can help in determination of the first dimension with the most contribution.

2.2 Similarity-based Re-ordering

The correlation of two variables (dimensions/attributes) is a statistical technique that can indicate the magnitude relationship between the two variables. It also shows how the two variables interact with each other. In this section, we present the reordering methods based on the two correlation measures: Pearson's correlation and nonlinear correlation information measures.

2.2.1 Linear/Nonlinear Correlation Analysis

Pearson's Correlation Coefficient [22], as one of the most popular similarity measures in visualization of multidimensional data, is a linear correlation measurement for each pair of random variables:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$,

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, $\bar{x}$ and $\bar{y}$ behave the mean of variables X and Y respectively. The value of PCC ranges in the closed interval [-1, 1], which indicates the linear correlation degree of the two variables. When it is close to 1 or -1, the PCC value denotes a strong relationship; and if close to 0, it means a weak relationship between the two variables. A positive and negative correlation coefficient denotes that both variables are in the same way or in the opposite way.

Although linear correlation can detect the relationship between two dependable variables, the correlations can also be nonlinear in the real world. Mutual Information can be thought of as a generalized correlation analogous to the linear correlation coefficient, but sensitive to any relationship, not just linear correlation. Moreover, NCC is a method that can measure nonlinear relationship based on mutual information[23, 24] and redundancy[25], which is sensitive to any relationship, not just the linear dependence[26]. Zhiyuan Shen et al.[26, 27] did further researches on the effects of statistical distribution to it and made it range in a closed interval $[0,1]$.

Corresponding to the literature[2], we mainly apply NCC to compute the partial similarity measures of dimensions in multidimensional data visualization, while SVD is used for measuring the global one. We introduce the detailed NCC in the following paragraphs. Mutual information plays an important role in the computation of NCC, which is defined as

$$I(X;Y) = H(X) + H(Y) - H(X;Y) \quad (3)$$

where $H(X)$ is the information entropy of variable X:

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

$H(X;Y)$ is the joint entropy of the variables X and Y:

$$H(X;Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \ln p_{ij}$$

p_i denotes the probability distribution that random variable X takes the value x_i, and p_{ij} denotes the joint probability distribution $p(X = x_i, Y = y_i)$ of the discrete random variables X and Y.

After revising joint entropy of the two variables X and Y,

$$H^r(X;Y) = -\sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{n} \log_b \frac{n_{ij}}{n} \quad (4)$$

in which $b \times b$ rank grids are used to place the sample pairs $\{(x_1, y_i)\}_{1 \le i \le n}$. n_{ij} is the number of samples distributed in the $ijth$ rank grid, Wang et al. [27] proposed the calculation method for nonlinear correlation coefficient as follows:

$$\begin{aligned} NCC(X;Y) &= H^r(X) + H^r(Y) - H^r(X;Y) \\ &= 2 + \sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{n} \log_b \frac{n_{ij}}{n} \end{aligned} \quad (5)$$

In the following section 2.2.2, the above formula is applied to measure the linear or nonlinear relationship between the two dimensions in multidimensional data sets because of its sensitivity to any relationship.

Accepted manuscript

ACCEPTED MANUSCRIPT

2.2.2 Similarity-based Reordering

Since the problem of dimension reordering is similar to the Traveling Salesman problem, many researchers applied heuristic algorithms, such as genetic algorithms, colony optimization and nearest neighbor heuristic method etc. [2, 16], to overcome exhaustive time. In the method SBAA proposed by Almir Olivette Artero et al.[16], the largest value $s_{i,j}$ in their similarity matrix S (lower diagonal) is considered to be the initial dimension "ij" in the new order. And then, they try to search for the dimensions which will be positioned in the right of it. It seems rational that we just reorder all the dimensions in line with this similarity. However, some dimensions always attract much more concentrations from the whole visual structure. For example, in parallel coordinates, the first and the last dimensions can draw much more attention than the other axes do. Therefore, different from the existed methods, we propose a new dimensions reordering algorithm based on the NCC and SVD algorithms. And these methods help users reduce the computation complexity and improve the visual readability greatly.

We define the similarity matrix S, which is symmetric, as follows:

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$$

where

$s_{ij}=s_{ji}\ (i\neq j)$, which are calculated by use of nonlinear correlation coefficient.

s_{ii} $(i=1,2,\cdots,n)$ (We also can denote them as v_{1i}) behaves the contribution value of the ith dimension to the whole data values, which is calculated by SVD algorithm.

Similarity-based Reordering Algorithm

Step1. Form the matrix D of the data sets.

Step2. Calculate the singular value decomposition of matrix D, and get the contribution factors s_{ii}, $i=1,2,\cdots,n$.

Step3.Compute the other elements s_{ij} of similarity matrix S, using our nonlinear correlation coefficient method, besides s_{ii}, $i=1,2,\cdots,n$ which have calculated in step2.

Step4. Choose the largest value of s_{ii}, $i=1,2,\cdots,n$ as the extreme left attribute to start display the data sets. We denote this attribute as S_{ll}, $l\in\{1,2,\cdots,n\}$.

Step5. Get the largest value S_{lr} from $\{S_{li}, l<i\}$. Therefore, the r_1th attribute is appended to the lth attribute. We get the first two elements of neighbouring sequence $NS=\{l, r_1\}$.

Step6. Repeat step5 using the r_1th attribute as the left neighbouring attribute from $\{S_{r_1i}, r_1<i\}$ until inserting all attributes into the NS.

It is worth noting that this visualization method can not only provide us the similarities between each pair of dimensions, but also express some ideas of the self-property of each dimension. During the computation process of the nonlinear correlation coefficient, we chose the $b\times b$ rank grids according to the empirical formula, which is mentioned in [28]:

$$b=1.87\times(n-1)^{2/5} \qquad (6)$$

3 CASE STUDIES

With the application examples to demonstrate the effectiveness of our rational dimension reordering methods, we analyzed many data sets in this section, such as the one describing KDD Cup 1999 and Cars models for contribution-based reordering visualization and another one about Cars and Liver Disorders data set for our similarity-based reordering method. All of these data sets we tested come from the literature [29].

3.1 Contribution-based visualization

In this section, we utilize two data sets to show the effectiveness of our contribution-based method. Firstly, part datum from *KDD 1999* consisting of 1113 data items with 42 attributes (including "normal" and "abnormal" labels) are analyzed in Fig. 1.

Accepted manuscript

ACCEPTED MANUSCRIPT

To the whole 42 attributes, we use contribution-based reordering method as a dimension reduction step for visualizing data set. By setting the contribution rate as one of the simplest techniques to retain as much characteristics of the whole data set as we can do, we get the six attributes which retain the 99.98% of the overall information. It can be easily found that there are two different attacks in these 1113 data items: purple lines and red lines represent *"Smurf"* and *"Neptune"* attacks respectively. From the polylines among the attributes *"dst_host_count"* and *"count"*, we can find there is a big fluctuation between the normal and abnormal lines, which just presents us the pattern of attacks.

On the other hand, Cars dataset with seven dimensions, which consists of 392 values, is considered to be the second example to test the contribution-based reordering method. We calculated the contribution of each dimension to the whole dataset using the property in section 2.1 and the fourth attribute named *"Weight"* enjoys the largest contribution factor 0.9991 (The other contributions are listed as the diagonal elements in the following matrix S of next section). Fig. 2 visualizes the dataset in line with their contribution to the values in parallel coordinates. From the characteristic of values within the dataset, the visualization illustrates the contribution of each dimension from the highest rate to the lowest one. It is worth to be noticed that the polylines among the *"Cylinders"* and *"Origin"* attributes simplify the visual patterns due to their last two lower contributions to the overall data and neighboring each other.

3.2 Similarity-based reordering visualization

In this section, we visualize and compare the PCC with the NCC re-orderding method using Cars and Liver Disorders data sets. According to the literature [31], the larger the crossing angle between the polylines is, the less the cognitive load becomes, and the better the visualization efficiency turns to be. Therefore, in order to show the advantages in the readability and understandability of our method, we calculated the mean angles occurred among the polylines between two neighboring attributes using the following formula:

$$mean\_angle = \frac{total\_angle}{total\_edge\ cross\sin g}$$

3.2.1 Cars Dataset

Based on the theory in section 2.2, the similarity matrix of Cars data set was calculated as the following S .

$$S = \begin{bmatrix} 0.0067 & 0.5950 & 0.3236 & 0.0561 & 0.9078 & 0.8104 & 0.0302 \\ 0.5950 & 0.0018 & 0.5806 & 0.5028 & 0.8944 & 0.0261 & 0.6288 \\ 0.3236 & 0.5806 & 0.0354 & 0.1313 & 0.5223 & 0.9544 & 0.0104 \\ 0.0561 & 0.5028 & 0.1313 & 0.9991 & 0.3389 & 0.6968 & 0.0302 \\ 0.9078 & 0.8944 & 0.5223 & 0.3389 & 0.0047 & 0.9598 & 0.0197 \\ 0.8104 & 0.0261 & 0.9544 & 0.6968 & 0.9598 & 0.0235 & 0.0117 \\ 0.0302 & 0.6288 & 0.0104 & 0.0302 & 0.0197 & 0.0117 & 0.0004 \end{bmatrix}$$

After positioning the first dimension *"Weight"*, which makes a significant contribution to the whole data set, we try to find out the one from the unordered dimensions with the largest similarity value to this dimension: $S_{46} = 0.6968$. Therefore, the $6-th$ dimension is considered to be the strongest correlation with the $4-th$ one. And then, we make the $6-th$ attribute to be appended to the $4-th$ one. Similar to this process, we can get the final rational dimension order, which is

$4 \to 6 \to 5 \to 1 \to 2 \to 7 \to 3$

Corresponding to the initial Cars dataset, the reordering dimensions calculated using our algorithm is

Weight → Year → Acceleration → MPG
→ Cylinders → Origin → Horsepower

The reordering results after our analysis are visualized in parallel coordinates in Fig. 3(a).

We visualize Car dataset using the traditional reordering method-Pearson's Correlation Coefficient in Fig. 3(b). The corresponding order of dimensions is as follows:

Weight → Cylinders → Horsepower → MPG
→ Year → Acceleration → Origin

Comparing with these two images in Fig. 2, we can find that visualization structures between the *"Cylinders"* and *"Origin"* dimensions become much clearer and simpler with our method. In the visualization graph of NCC, the mean angle between the attributes *"Acceleration"* and *"MPG"* gets to 22.359°. Moreover, the mean angle between *"Cylinders"* and "Origin" attributes is 28.162°.Compared to the mean angle of the overall polylines produced in the PCC reordering method, 0.422°, the

Accepted manuscript

ACCEPTED MANUSCRIPT

angle in NCC reordering one is 21.2 times larger than it. Therefore, we can find the visual effect of our reordering method is much better than the traditional ones.

Table 1 presents the detailed comparisons between the similarity values of attributes, which are calculated using PCC and NCC. The numbers from 1 to 7 denote the dimensions: *"MPG, Cylinders, Horsepower, Weight, Acceleration, Year and Origin"* separately. Note that no matter which method we use, the similarities between the two dimensions are the same, that is $s_{ij} = s_{ji}\ (i \neq j)$. It is obvious that there are big differences between the similarity values with two methods. For example, to our knowledge, the similarity between the $3-th$ *("Horsepower")* and $7-th$ (*"Origin"*) dimensions of the dataset is not strong enough at all. For example, the computation result of similarity by using PCC is 0.4552, while ours NCC is 0.0104.

3.2.2 Liver Disorders Dataset

Liver Disorders dataset consists of 345 instances with 7 dimensions. Fig. 4 illustrates us the final visualization result of the whole dataset according to their similarities calculated by NCC and PCC methods respectively, where the dimension *"MCV"* makes the most significant contribution to the whole dataset and occupies the first place in the two reordering visualization.

It is easy to find that polylines among the "SF" and "DN" attributes are much less than those among any other neighboring attributes in Fig. 4 (a) and (b). The mean crossing angle of these two dimensions, 43.515°, as the largest one in the dimensions reordering visualizations, simplifies the visual representation greatly. The mean crossing angle of our NCC reordering method to this dataset is 12.322°, which is 3.722° larger than the result calculated using PCC method.

We also tested the other data sets such as Nursery, Iris et al. large scale ones to illustrate the advantages of our methods. All of them showed our methods can enlarge the mean crossing angles for better visualization.

4 Conclusions and Future Work

In this paper, two new methods are proposed to improve the readability and understandability of parallel coordinates theoretically. At the first stage, singular value decomposition algorithm provides a new way of looking into the dimensions within data sets. We propose a contribution-based reordering method and a formula for contribution rate of each dimension. At the second stage, we present a method, named as similarity-based reordering method, for calculating the similarity between the two dimensions based on the nonlinear correlation coefficient and singular value decomposition algorithms rather than the traditional Pearson's correlation coefficient, and then visualize the optimal dimension order according to the similarity in parallel coordinates. At last, the experimental evaluations demonstrate the effectiveness and rationale of our approaches: the patterns of Smurf and Neptune attacks hidden in KDD 1999 dataset are visualized in parallel coordinates using the contribution-based reordering method; NCC reordering method enlarges the mean crossing angles of the whole data set and reduces the amount of polylines between some neighbouring dimensions.

During the process of calculation for nonlinear correlation coefficient, the more exact choice of rank grids will do much more help in the speed up of calculation. Therefore, we consider this issue to be our first work in priority. Secondly, we will apply our methods with interactive techniques to more real-world data sets and help users analyze the data sets using visualization.

Acknowledgments

The authors wish to thank anonymous reviewers for spending valuable time in reviewing our paper and providing us with important comments to improve the paper. This work was supported in part by Science Foundation of Tianjin grant 15JCQNJC00200.

References

[1] Johansson, S. and J. Johansson, Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics. *IEEE Transactions on Visualization and Computer Graphics,* 2009. 15(6): p. 993-1000.

[2] Ankerst, M., S. Berchtold, and D.A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *Proceedings of IEEE Symposium on Information Visualization*, 1998:p. 52-60.

[3] Inselberg, A., The plane with parallel coordinates. *The Visual Computer*, 1985. 1(2): p. 69-91.

[4] Wegman, E.J., Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 1990. 85(411): p. 664-675.

[5] Becker, R.A. and W.S. Cleveland, Brushing Scatterplots. *Technometrics*, 1987. 29(2): p. 127-142.

[6] Rao, R. and S. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. *Proceedings of the SIGCHI conference on Human factors in computing systems.* 1994. Boston, MA, USA: ACM.p.318-322.

[7] Keim, D.A., Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and Computer Graphics,* 2000. 6(1): p. 59-78.

Accepted manuscript

ACCEPTED MANUSCRIPT

[8] Bertini, E., A. Tatu, and D. Keim, Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Transactions on Visualization and Computer Graphics,* 2011. 17(12): p. 2203-2212.

[9] Dasgupta, A. and R. Kosara, Pargnostics: Screen-Space Metrics for Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 2010. 16(6): p. 1017-1026.

[10] Tatu, A., et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. *Proceedings of IEEE Symposium on Visual Analytics Science and Technology, 2009,* VAST 2009. p. 59-66.

[11] Tatu, A., et al., Automated Analytical Methods to Support Visual Exploration of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics,* 2011. 17(5): p. 584-597.

[12] J.Bertin, Semiology of graphics. 1983: University of Wisconsin Press.

[13] Hahsler, M., K. Hornik, and C. Buchta, Getting things in order : an introduction to the {R} package seriation. *Journal of Statistical Software*, 2008. 25(3):p. 1-34.

[14] Hurley, C.B. and R.W. Oldford, Pairwise Display of High-Dimensional Information via Eulerian Tours and Hamiltonian Decompositions. *Journal of Computational and Graphical Statistics,* 2010. 19(4): p. 861-886.

[15] Peng, W., M.O. Ward, and E.A. Rundensteiner, Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering, *Proceedings of the IEEE Symposium on Information Visualization. 2004,* IEEE Computer Society. p. 89-96.

[16] Artero, A.O., M.C.F.d. Oliveira, and H. Levkowitz, Enhanced High Dimensional Data Visualization through Dimension Reduction and Attribute Arrangement, *Proceedings of the conference on Information Visualization. 2006*, IEEE Computer Society. p. 707-712.

[17] Friendly, M. and E. Kwan, Effect ordering for data displays. *Computational Statistics & Data Analysis,* 2003. 43(4): p. 509-539.

[18] J. Yang, M.O.W., E.A. Rundensteiner and S. Huang, Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Data sets, *in Joint EUROGRAPHICS - IEEE TCVG Symposium on Visualization (2003)*, S.H. G.-P. Bonneau, C. D. Hansen, Editor. 2003. p. 19-28.

[19] Guo, D., Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization,* 2003. 2(4): p. 232-246.

[20] Albuquerque, G., et al. Improving the visual analysis of high-dimensional data sets using quality measures. *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST). 2010:*p.19-26.

[21] Golub, G.H. and C.F.V. Loan, Matrix Computations. 1983, Baltimore: John Hopkins University Press.

[22] Rodgers, J. and A. Nicewander, Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 1988. 42(1): p. 59-66.

[23] Zheng Rong, Y. and M. Zwolinski, Mutual information theory for adaptive mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2001. 23(4): p. 396-403.

[24] Matsuda, H., Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 2000. 62(3): p. 3096-3102.

[25] Drmota, M. and W. Szpankowski, Precise minimax redundancy and regret. *IEEE Transactions on Information Theory*, 2004. 50(11): p. 2686-2707.

[26] Zhiyuan, S., W. Qiang, and S. Yi. Effects of statistical distribution on nonlinear correlation coefficient. *Proceedings of IEEE Instrumentation and Measurement Technology Conference (I2MTC)*, 2011.

[27] Wang, Q., Y. Shen, and J.Q. Zhang, A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena*, 2005. 200(3–4): p. 287-295.

[28] Zhuang Chu Qiang, W.Y.S., Mathematical Statistics with Applications. 1992, Guangzhou: South China science and technology university press.

[29] University of California, Irvine. Center for Machine Learning and Intelligent Systems: http://archive.ics.uci.edu/ml/data sets.html.

[30] K. Simek, "Properties of a singular value decomposition based dynamical model of gene expression data," *International Journal of Applied Mathematics and Computer Science*, 2003. 13(3), p. 337-345.

[31] W. H. M. L. Huang, "Exploring the Relative Importance of Number of Edge Crossings and Size of Crossing Angle: A Quantitative Perspective," *International Journal of Advanced Intelligence*, 2011.3(1), p. 25-42.

[32] Ayan Biswas, Soumya Dutta, Han-Wei Shen, Jonathan Woodring. "An Information-Aware Framework for Exploring Multivariate Data Sets". *IEEE Transactions on Visualization and Computer Graphics,* 2013, 19(12), p.2683-2692.

Accepted manuscript

ACCEPTED MANUSCRIPT

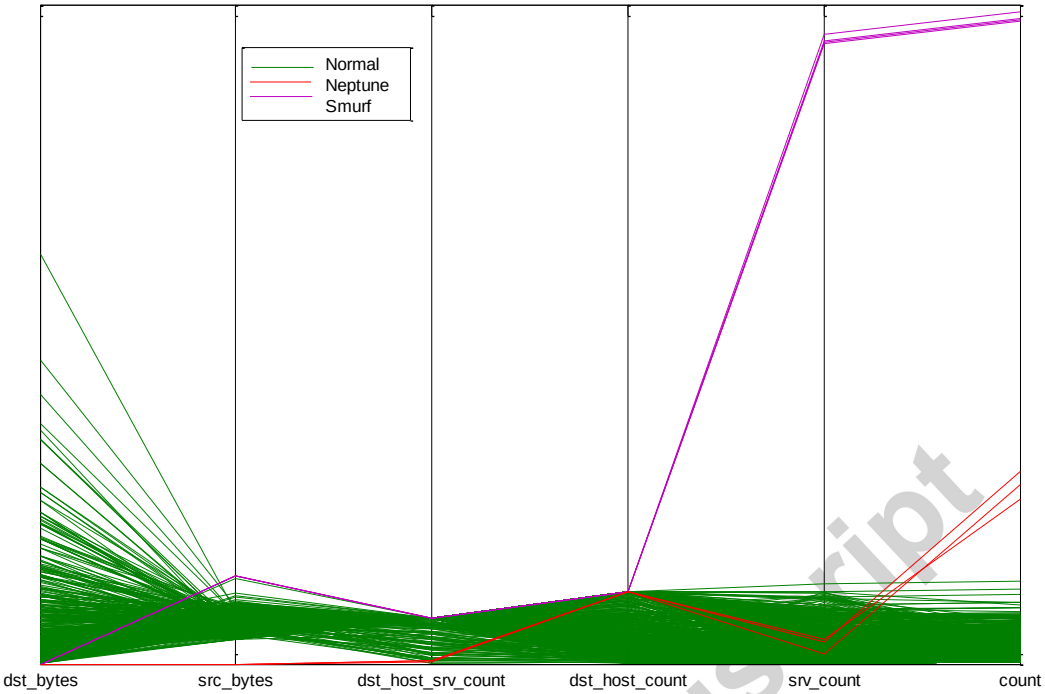


Fig. 1. Contribution-based reordering of KDD 1999 dataset in parallel coordinates.

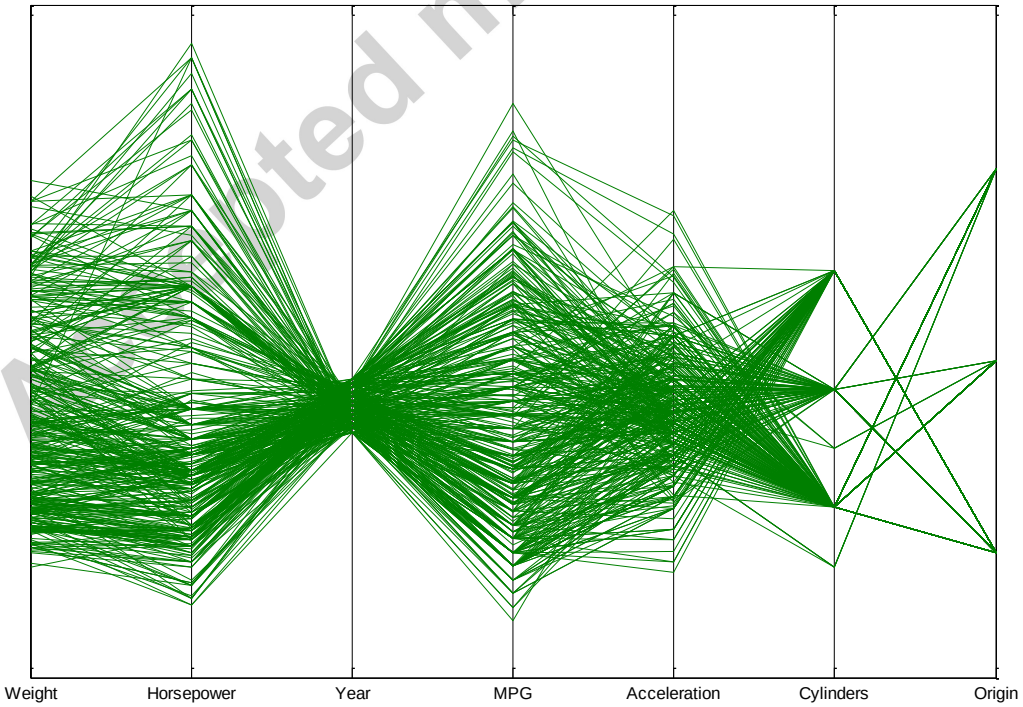
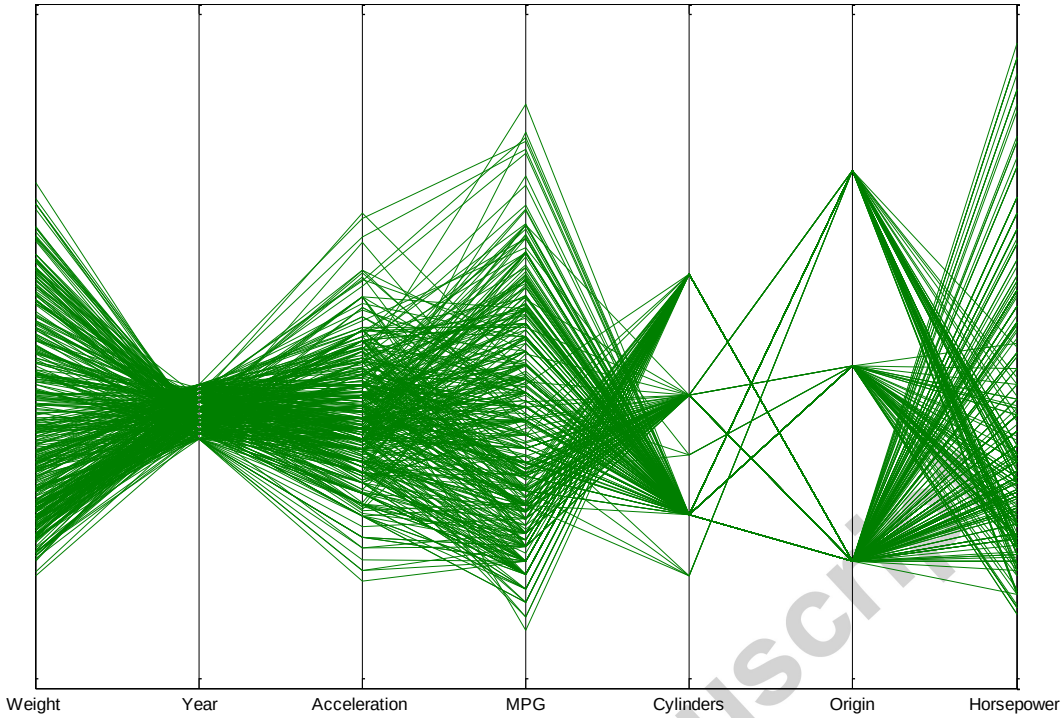


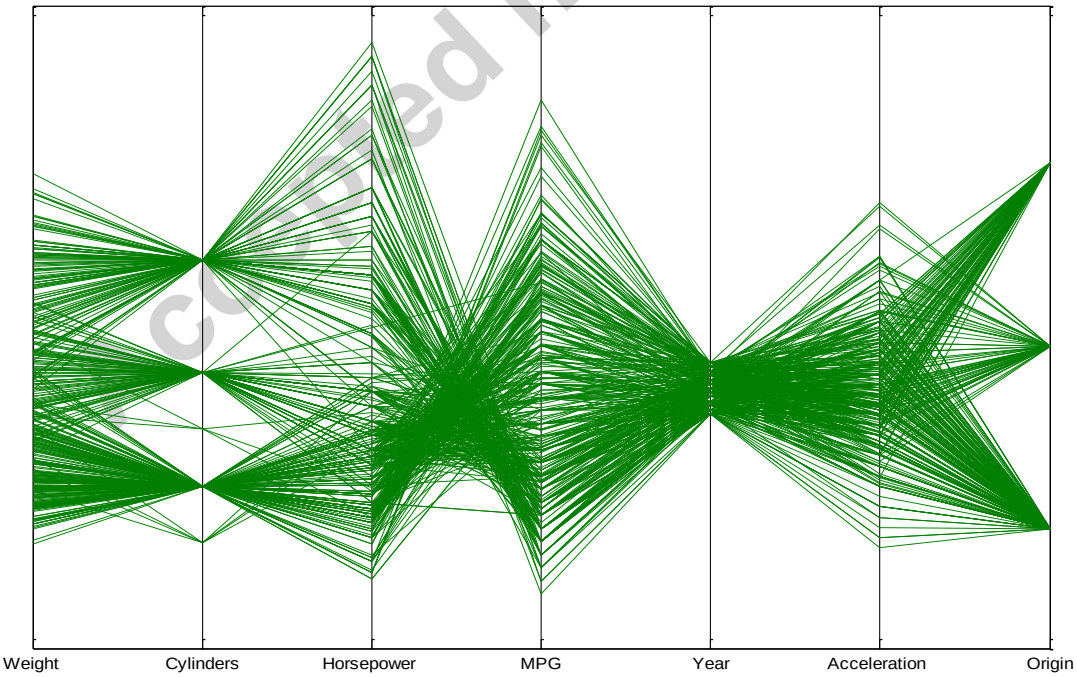
Fig. 2. Contribution-based reordering of Cars dataset in parallel coordinates.

Accepted manuscript

ACCEPTED MANUSCRIPT



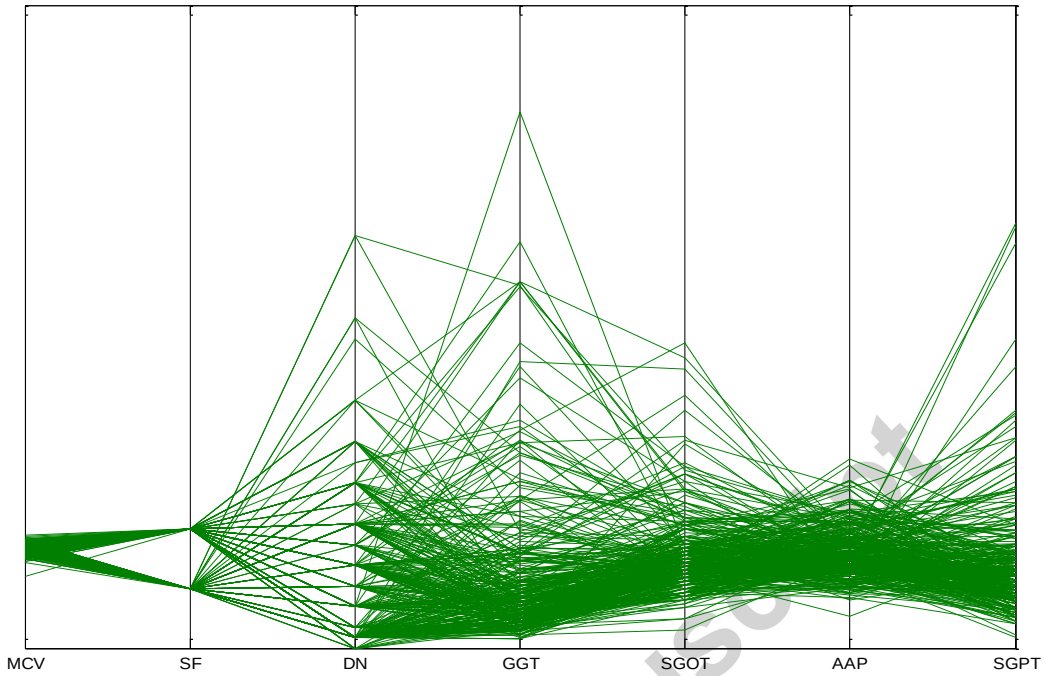
(a) Measurement with NCC



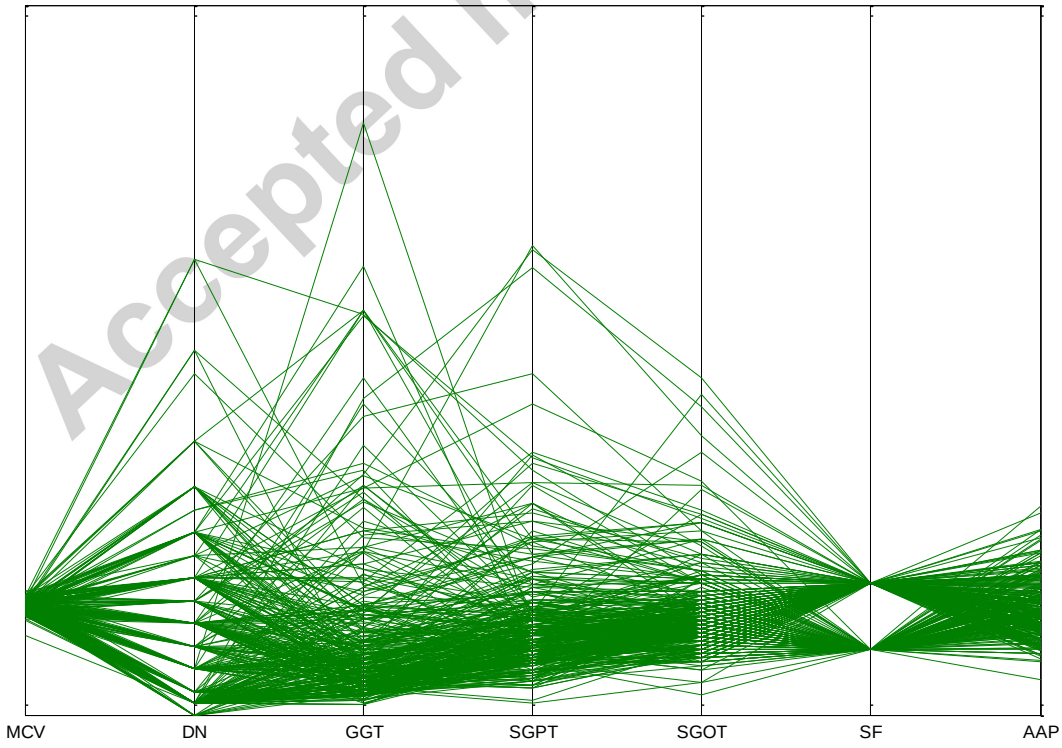
(b) Measurement with PCC.

ACCEPTED MANUSCRIPT

Fig. 3. Dimension reordering of Cars dataset in parallel coordinates.



(a) Measurement with NCC.



(b) Measurement with PCC.

Accepted manuscript

ACCEPTED MANUSCRIPT

Fig. 4. Dimension reordering of Liver Disorders dataset in parallel coordinates.

Table 1 The comparison of the similarity values using PCC and NCC to Cars dataset.

PCC / NCC	2	3	4	5	6	7
1	0.7776 / 0.5950	0.7784 / 0.3236	0.8322 / 0.0561	0.4233 / 0.9078	0.5805 / 0.8104	0.5652 / 0.0302
2		0.8429 / 0.5806	0.8975 / 0.5028	0.5046 / 0.8944	0.3456 / 0.0261	0.5689 / 0.6288
3			0.8645 / 0.1313	0.6892 / 0.5223	0.4163 / 0.9544	0.4552 / 0.0104
4				0.4168 / 0.3389	0.3091 / 0.6968	0.5850 / 0.0302
5					0.2903 / 0.9598	0.2127 / 0.0197
6						0.1815 / 0.0117

Accepted manuscript