Data Analytics Enhanced Data Visualization and Interrogation with Parallel Coordinates Plots

Muhammad Sajjad Akbar
Advanced Analytics Institute
University of Technology Sydney
Muhammad.Akbar@uts.edu.au

Bogdan Gabrys
Advanced Analytics Institute
University of Technology Sydney
Bogdan.Gabrys@uts.edu.au

Abstract: Parallel coordinates plots (PCPs) suffer from curse of dimensionality when used with larger multidimensional datasets. Curse of dimentionality results in clutter which hides important visual data trends among coordinates. A number of solutions to address this problem have been proposed including filtering, aggregation, and dimension reordering. These solutions, however, have their own limitations with regard to exploring relationships and trends among the coordinates in PCPs. Correlation based coordinates reordering techniques are among the most popular and have been widely used in PCPs to reduce clutter, though based on the conducted experiments, this research has identified some of their limitations. To achieve better visualization with reduced clutter, we have proposed and evaluated dimensions reordering approach based on minimization of the number of crossing pairs. In the last step, k-means clustering is combined with reordered coordinates to highlight key trends and patterns. The conducted comparative analysis have shown that minimum crossings pairs approach performed much better than other applied techniques for coordinates reordering, and when combined with k-means clustering, resulted in better visualization with significantly reduced clutter.

I. Introduction

Parallel coordinates (PC), also known as Parallel coordinates plots (PCPs), are widely used in data visualizations [1, 2, 3]. Three important characteristics are attached to PCPs: (1) the angles of line segments for identifying positive or negative relationships; (2) the co-location of line segment crossings to measure the strength of the relationships; and (3) density of line segments. Figure 1: (a) (b) (c) (d) illustrates these characteristics.

While being very versatile visualisation and interactive data interrogation tool, there are a number of challenges when using PCPs with large datasets [1] including: visual clutter, perceptual scalability, overlapping lines between adjacent axes, finding relevant projections, selection of meaningful variables, noise and interactive semi-automated data analysis [3, 4, 5]. Clutter is produced by a large number of lines which potentially hide the data trends and relationships between variables. Clutter reduction techniques are categorized into two types: data driven and screen tested. Several methods have been proposed to solve the clutter issue, where the most often used ones are filtering, aggregation, spatial distortion and variable reordering. Moreover, identification of relationships among coordinates is also a challenging task. The types of relations usually searched for are linear, non-linear, and monotonic. Identification of these relations help to view meaningful data trends. Researchers have been trying to ad-

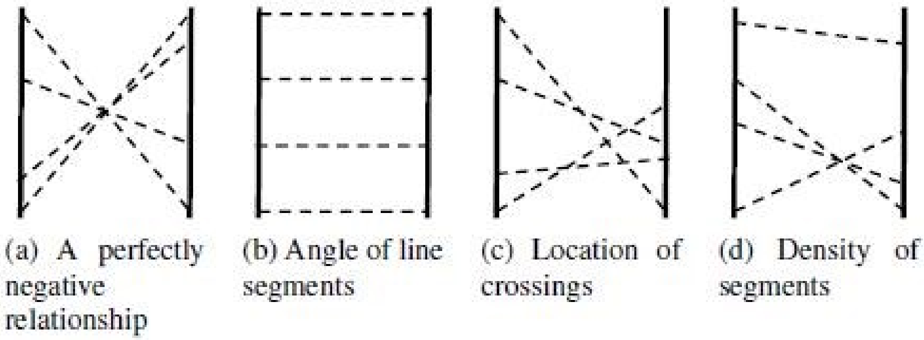
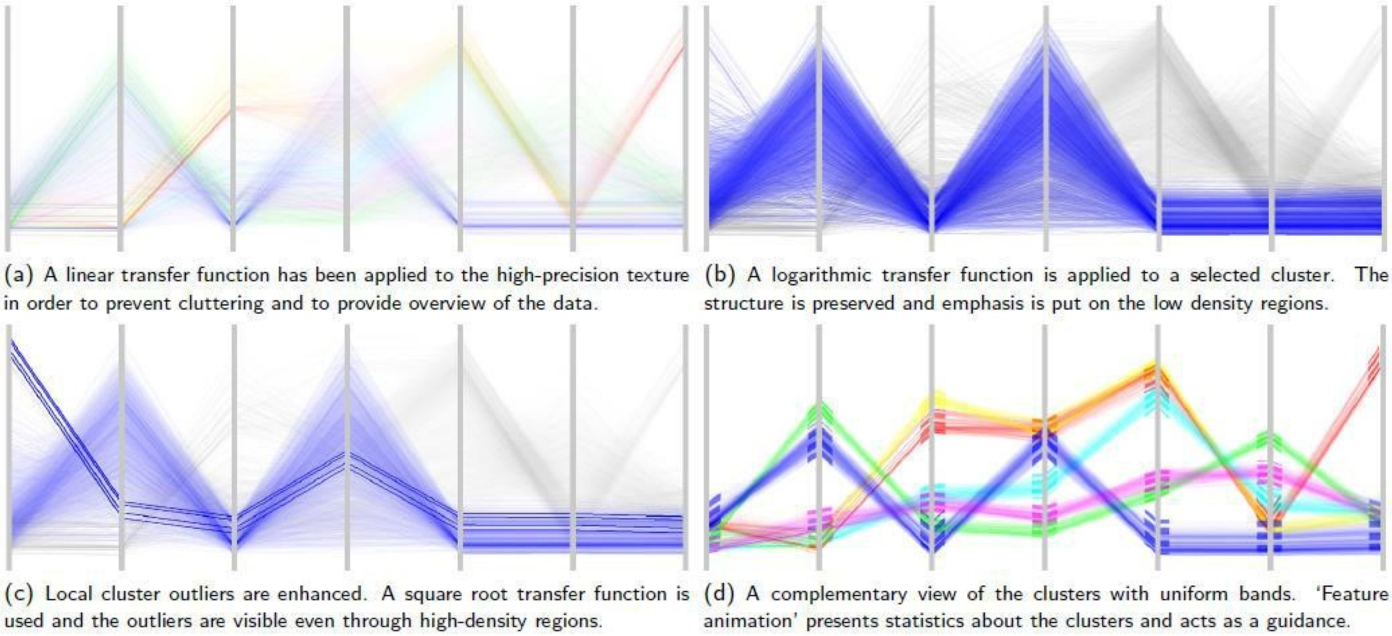

(a) A perfectly negative relationship
(b) Angle of line segments
(c) Location of crossings
(d) Density of segments

Figure 1: Characteristics of parallel coordinates

dress these problems through a number of automatic data analysis and visualization approaches that cover the whole spectrum of possibilities from fully automatic to fully interactive. In this paper, we have evaluated some prominent clutter reduction techniques to check the extent to which their claims of clutter reduction can be validated. The purpose is to identify the most suitable technique which can reduce clutter and provides a better visualization. Lastly, we proposed a combination of coordinates reordering technique with k-means clustering to achieve a comparatively better visualizations of large, high dimensional datasets.

II. Related work

Wegman [4] encourages parallel coordinate's visualization with respect to geometry, statistics and graphics, which has been extensively applied [5]. For visualizing clustered data sets, several methods have been proposed using parallel coordinates [6]. Instead of visualizing each data item as a polyline, each cluster pattern is visualized as a fuzzy stripe [7]. Fua et al. [8] envision clusters by variable-width opacity bands which focuses on the global pattern of clusters. The general shape of a cluster, however, might be affected by a small number of outliers inside a cluster. Figures 2 a-d show examples of applied clustering and different enhanced data visualizations.



(a) A linear transfer function has been applied to the high-precision texture in order to prevent cluttering and to provide overview of the data.

(b) A logarithmic transfer function is applied to a selected cluster. The structure is preserved and emphasis is put on the low density regions.

(c) Local cluster outliers are enhanced. A square root transfer function is used and the outliers are visible even through high-density regions.

(d) A complementary view of the clusters with uniform bands. 'Feature animation' presents statistics about the clusters and acts as a guidance.

Figure 2: An example data set containing 7,800 7-dimensional data items classified into 6 clusters [6]

In [8], authors proposed a variation on parallel coordinates by using proximity-based colouring to assure that clusters from similar parts of the hierarchical structure are shown using the same colours. Novotny [16] characterized each cluster as a polygonal area and used multiple opacity values and textures to differentiate clusters. In [17], striped envelope and ellipse plots have been used to discover the properties of subsets. Wong and Bergeron [18] designed a wavelet brushing technique for large datasets. [19] used angular brushing for data subsets having particular trends.

Various visual enhancement schemes have also been proposed for parallel coordinates. In [20], authors used high-precision textures and transfer functions to disclose clusters. In [21] a useful outlier-preserving focused visualization tool was developed. The authors in [22, 23], proposed extensions for parallel coordinates which interactively explore categorical data.

The dimension reordering is associated with PCP's visualization [9, 10]. In [4], Wegman highlighted the problem and proposed a solution on how to enumerate the minimum number of permutations such that every pair of coordinates can be visualized in at least one of the permutations. The grand tour animated a static display in order to examine the data from several views [11, 12, 13]. The authors in paper [1] proposed a method to rearrange dimensions such that dimensions showing a similar behaviour are positioned next to each other. [14] used a number of crossings as a sign of clutter between two adjacent coordinates, and then applied Branch-and-Bound optimization for dimension reordering. The research in [15] used crossings to understand correlation between two dimensions of a dataset. Figure 3 shows an example for clustering and filtering approach. Large numbers of dimensions not only cause clutter in multidimensional visualizations, but also make it difficult for users to navigate the data space. Effective dimension management, such as dimension ordering, spacing and filtering, is critical for visual exploration of such datasets. Automatic dimension reordering techniques, such as those explored in this paper, make use of similarity based solutions.

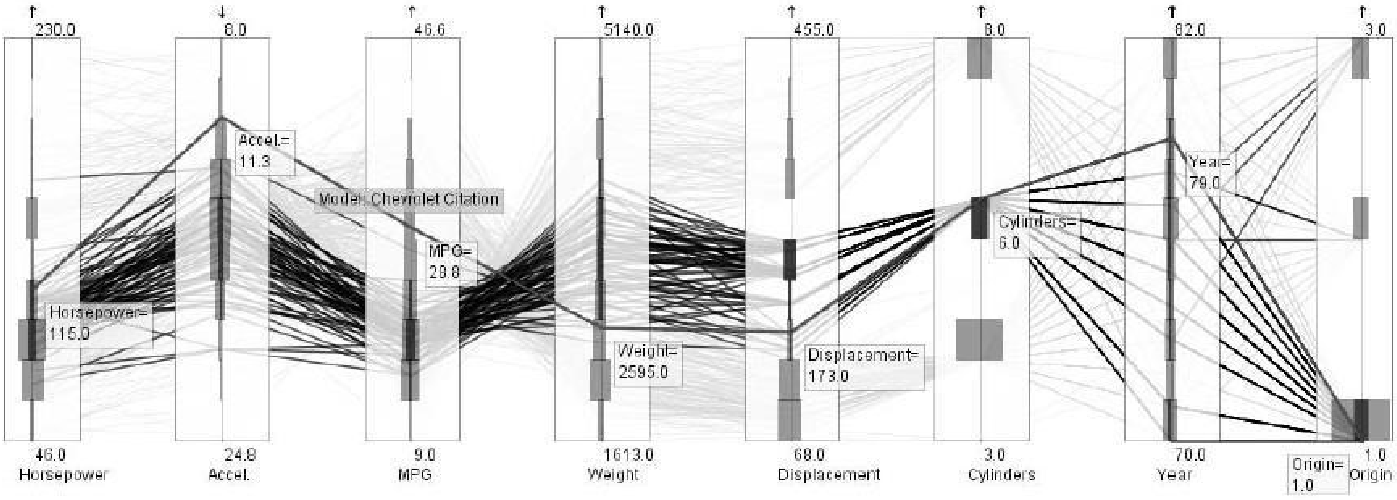


Figure 3: A sample view of the cars data-set: cars with six cylinders were emphasized through brushing [19]

III. Dimensions reordering approaches: Evaluation and discussion

We started work on improving PCPs visualization with two goals: reducing the clutter and exploration of hidden data patterns in dataset. Various approaches are presented in literature based on dimension reordering. Most of them are similarity based and other do reordering by identification of optimal dimension reordering [39, 40, 41]. Similarity based techniques mostly use the Pearson correlation while some of them also used Kendall and Spearman correlations [42, 43]. Pearson's correlation coefficient is a linear correlation measurement between two variables. In contrast to Pearson correlation, Spearman correlation determines the strength and direction of the monotonic relationship between two variables irrespective if the relationship is linear or not. Similarly to Spearman correlation, Kendall correlation is also a non-parametric rank correlation test that measures the strength of dependence between two variables but with often quoted smaller gross error sensitivity (i.e. it is more robust) and a smaller asymptotic variance (i.e. it is more efficient) [44, 45]. The mentioned literature compute the correlations among dimensions and then reorder them accordingly. Another promising dimension reordering approach is based on counting and minimizing the number of crossing of edges among dimensions. Both of mentioned approaches claimed that clutter can be significantly reduced. In this section, our aim is to identify the suitable technique to reduce the clutter in a given dataset by reordering the coordinates and then use such reordered dataset as an input to a clustering algorithm for better data summarization. In order to illustrate the effectiveness of the investigated approaches and substantiate our discussions we have used Male athlete strength dataset (http://www.artofstat.com/datasets.html) as a test-case for experiments. This dataset shows a performance comparison of 62 athletes based on various parameters including physical exercises, body conditions and age. Figure 5 shows PCPs for the dataset. In the red oval, a high clutter area can be seen, moreover, Figure 5 does not show any clear data trend among dimensions.

In our implementation, correlation values are computed between any given variable and all remaining variables, then we perform reordering of the variables to display using an algorithm that searches for highly correlated variables in order to display them next to each other. Such variable reordering based on correlation gives the flexibility to the user to analyze the same data in different logical visualizations.

It is also important to validate the implemented approaches with regard to their ability to reduce the clutter. The main source of clutter is data density and crossing of edges in parallel coordinates. Therefore crosses of edges (CE) can be computed as:

$$TC = CE * d \tag{1}$$

$$CE = \frac{n * (n-1)}{2} \tag{2}$$

Here, n is the number of data items, TC represents total number of crosses, d is the number of dimensions and CE is the number of crosses between a pair of axes. Suppose that the number of variables is 5 and n=10 then total number of possible crosses will be 450. There are various ways of calculating the actual total number of crossings in PCPs given an order of displayed variables and an interested reader is directed to [14] for details. In our algorithm we have used an optimization procedure that minimizes the total number of crossings by suitable reordering of the coordinates.

To illustrate the impact of the reordered coordinates on quality of visualization, we used Male athlete strength dataset as a test-case (http://www.artofstat.com/datasets.html). This dataset shows a performance comparison of 62 athlete based on various parameters including physical exercises, body conditions and age. Figure 5 shows PCPs for the dataset. In the red oval, a high clutter area can be seen, moreover, Figure 4 does not show any clear data trend among dimensions.

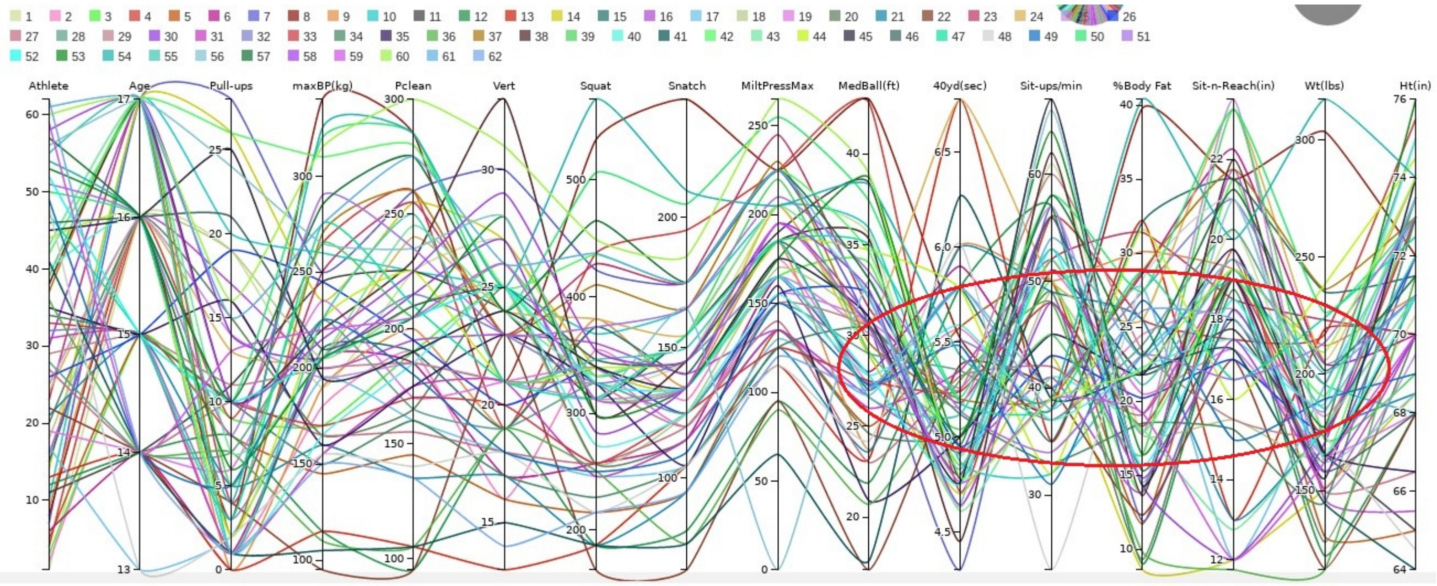


Figure 4: Man Strength Athlete database with the original order of variables

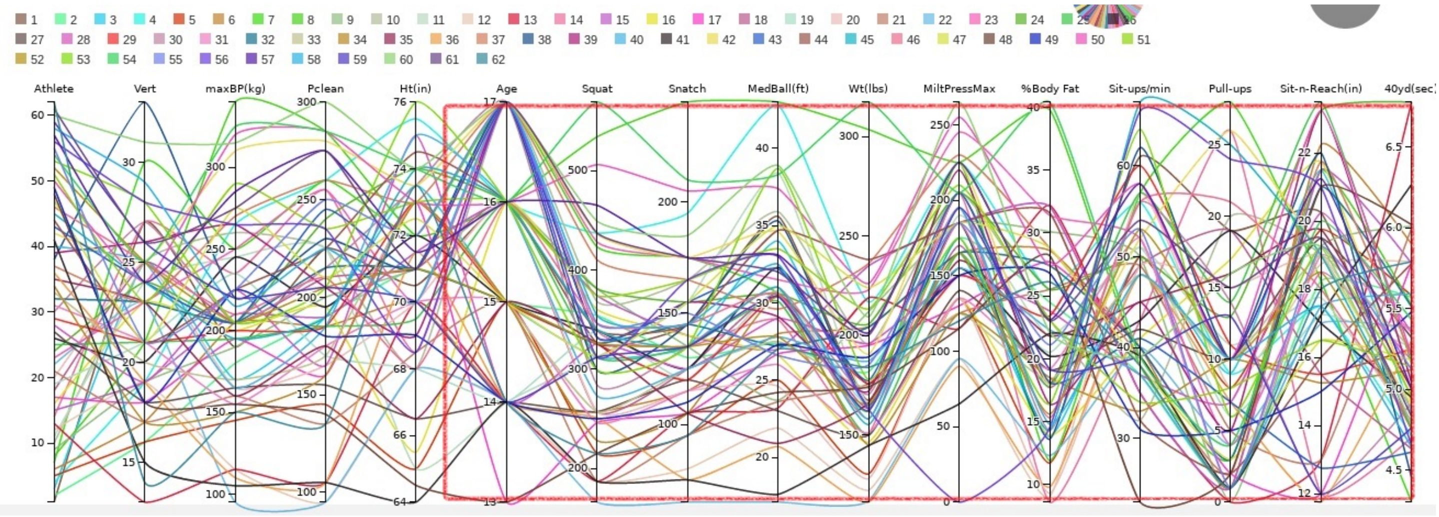


Figure 5: Man Strength Athlete database reordered based on the Pearson correlation

Figure 5 shows the Pearson based reordering of the Male athlete dataset. It can be seen from Figure 5 that the coordinates "Vert", "maxBP", and "pclean" etc. with highest linear correlation coefficient values have been displayed next to each other. In the highlighted red rectangular area, clearer trends of dimensions with respect to each other can be more easily noticed. For instance, the athletes with the age around 17 performing squats between 300 and 500. Although in Figure 5 one can see more data patterns than in the Figure 4, there is still substantial clutter present.

Figure 6 shows PCP visualization of the analyzed dataset with Kendall's correlation based reordering. As expected the order of the dimensions is different due to the nonlinear correlation measure and coefficients used in their ordering. The highlighted red rectangle shows few more easily identifiable data trends with less visual clutter. This is particularly noticeable for the first eight displayed dimensions. In some sense the visualization of Figure 6 is better than Figure 4, yet there are still a number of clearly seen clutter areas.

Figure 7 shows PCP visualization of the analyzed dataset with Spearman's correlation based reordering. Once again we can see a somewhat different view of the same data allowing one to potentially identify different relationships among variables. The blue rectangular highlights few more easily identifiable data trends with less cluttered area, particularly for the dimensions from "Age" to "Sit-ups(min)". While the visualization of Figure 7 can be considered to be better than Figure 4, yet there are still various clutter areas present.

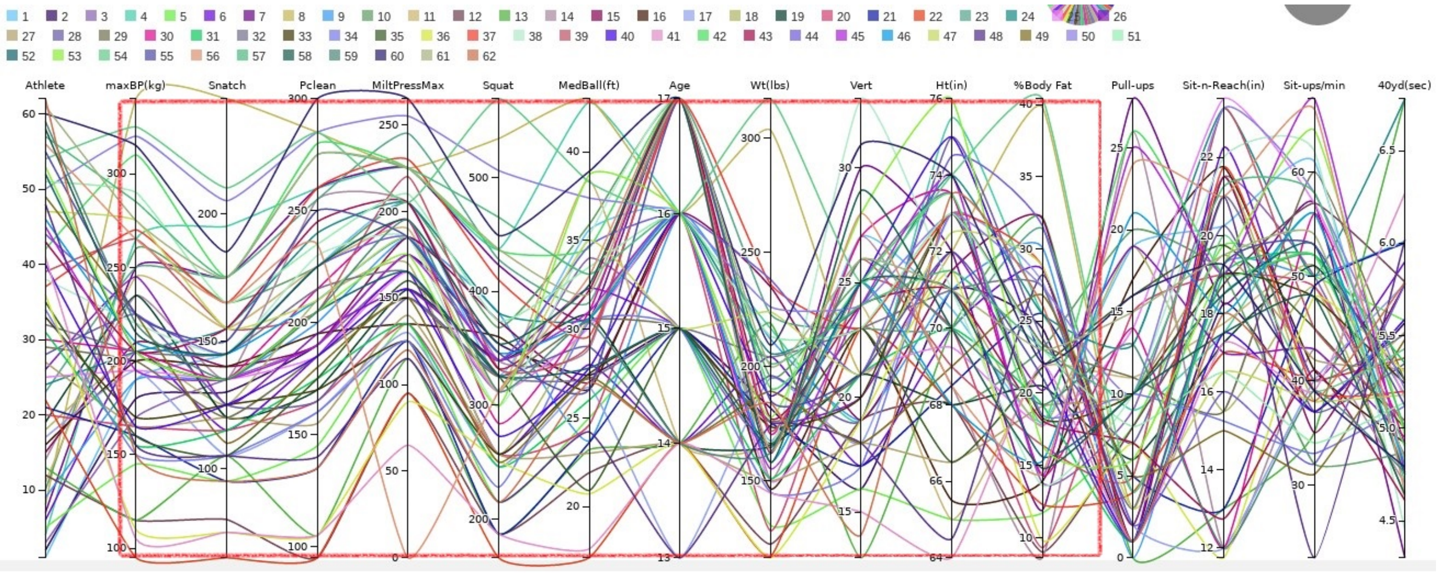


Figure 6: Man Strength Athlete database reordered based on the Kendall's correlation values

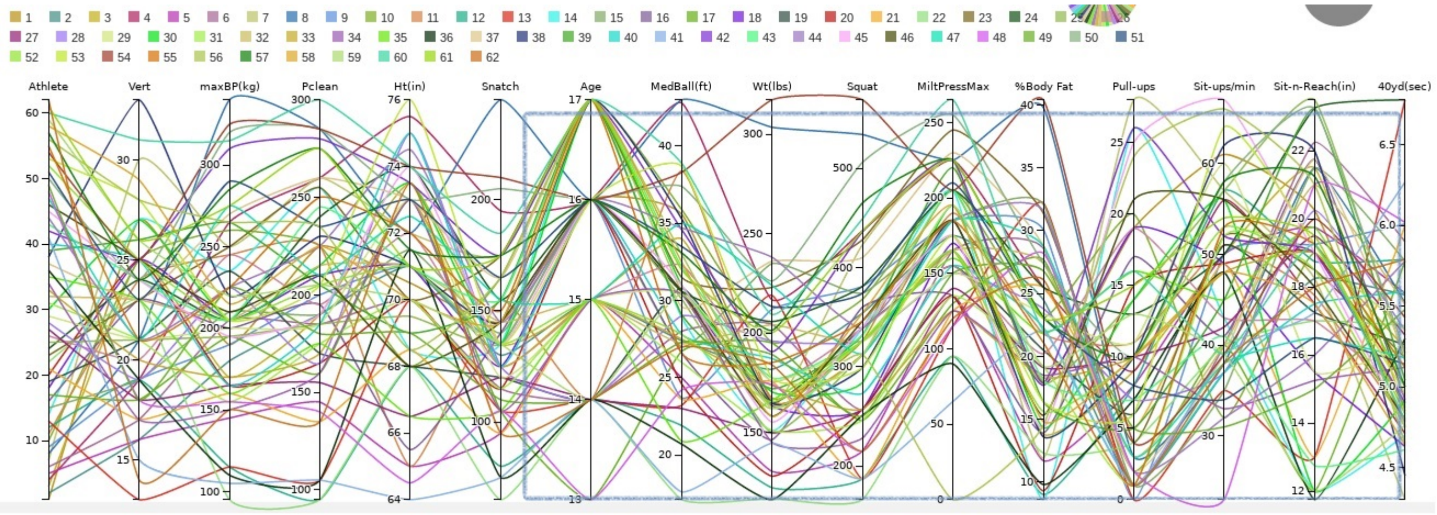


Figure 7: Man Strength Athlete database reordered based on the Spearman's correlation values

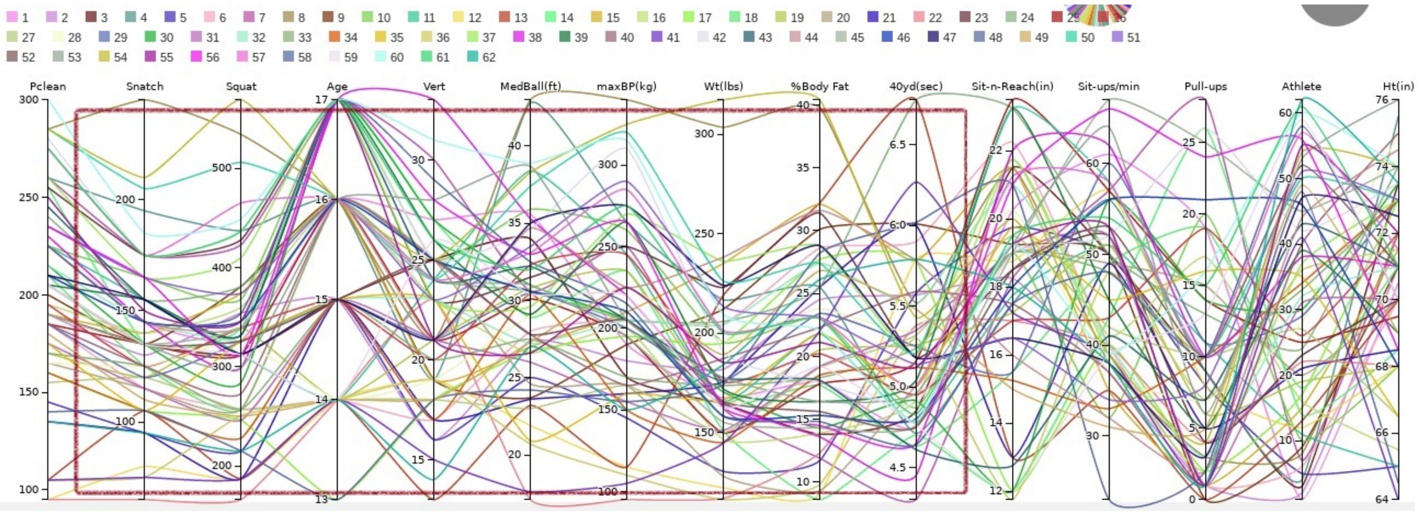


Figure 8: Man Strength Athlete database reordered based on the minimum crossing pairs

While the previous reorderings have been based on very often used for this purpose correlation measures, Figure 8 shows minimum crossing pairs based reordering of Male athlete dataset. In this reordering, we find the pairs of dimensions with the smallest number of crossings.

The red rectangle highlights an area with few more easily identifiable data trends and less clutter. Table 1 and Figure 9 provide the comparison among applied reordering techniques in terms of reduced clutter. The relative clutter is computed based on the total number of crossings. For instance, there are 18457 crossings present in the dataset before applying any reordering techniques. It can be clearly seen from Figure 9 that the reordered coordinates employing the minimum crossing pairs method was able to reduce the total number of crossings (and related to it visual clutter) by 29%, whereas the Pearson correlation based reordering was able to reduce the total number of crossings by 10.5%. From the above analysis we can conclude that if one is interested in reducing overall clutter in all displayed coordinates, the method based on directly minimizing the total number of crossings should be used, while correlation based reordering techniques can have a value for identifying interesting subsets of coordinates (those showing particularly high linear or non-linear correlation) while not paying attention to overall clutter reduction in the remaining coordinates.

Having said all of the above, even with the reduction of the total number of crossings by 29%, visual clutter (due to the number of items displayed and the significant number of crossings remaining) is still present and can cause an issue with clear identification of characteristic patterns across all displayed variables. In order to enhance the identified patterns, clustering techniques on reordered datasets based on correlations and minimum crossings pair can be applied.

Table 1: Number of crossings comparison

Applied Techniques	**No. of crossings**
Original dataset	18457
Pearson correlation	16524
Kendall correlation	16851
Spearman correlation	16997
Minimum crossing pairs	13087

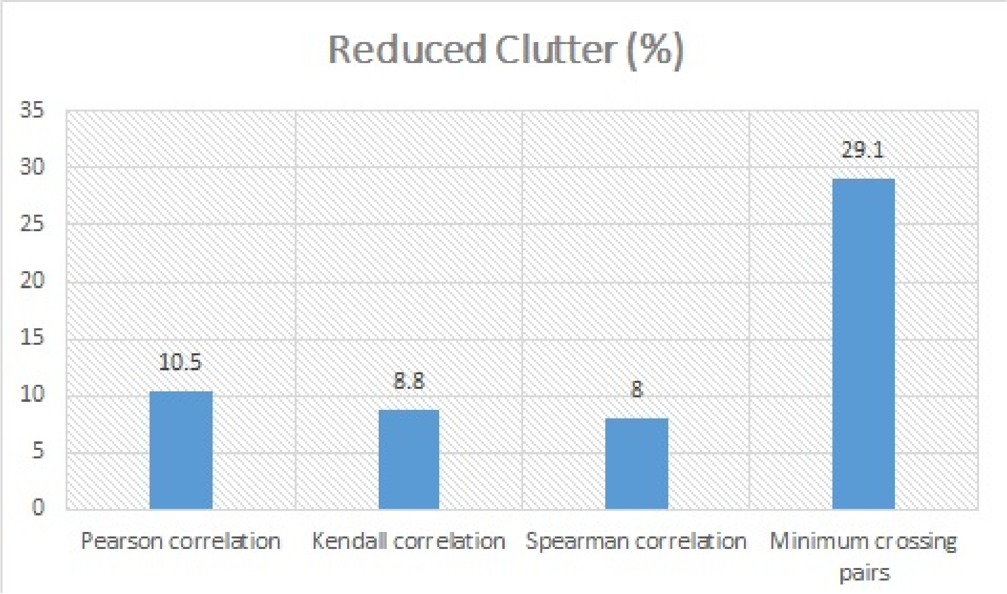


Figure 9: Performance comparison of applied techniques

1 IV. Representing multiple relationships

In the next stage, we apply K-means clustering to the results of the first step (reordered datasets) which is obtained by applying correlation techniques and minimum crossing pairs method. This step will further improve the complex visualization of PCPs. It is important to find optimal number of clusters for a given dataset before applying clustering. For this, we used Elbow method on Male athlete dataset. Figure 10 shows the elbow is at k=4 indicating the optimal k for this dataset is 4.

Figures 11, 12, 13 and 14 show PCPs visualization of k-means clusters for Male athlete reordered datasets in previous section for Minimum crossing pairs, Pearson correlation, Kendall correlation and Spearman correlation.

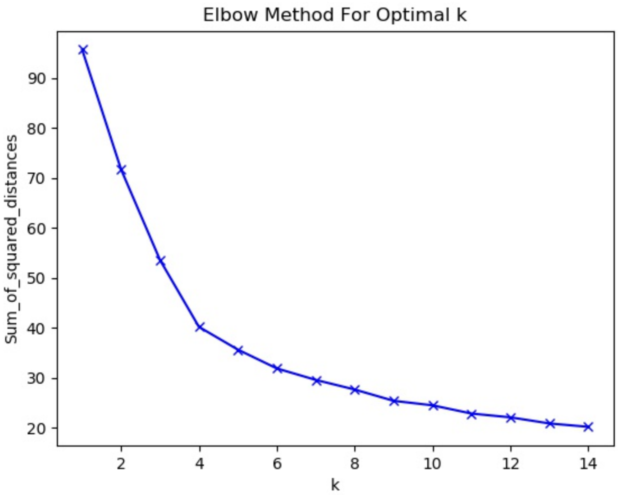


Figure 10: Optimal clusters for Male athlete dataset using Elbow method

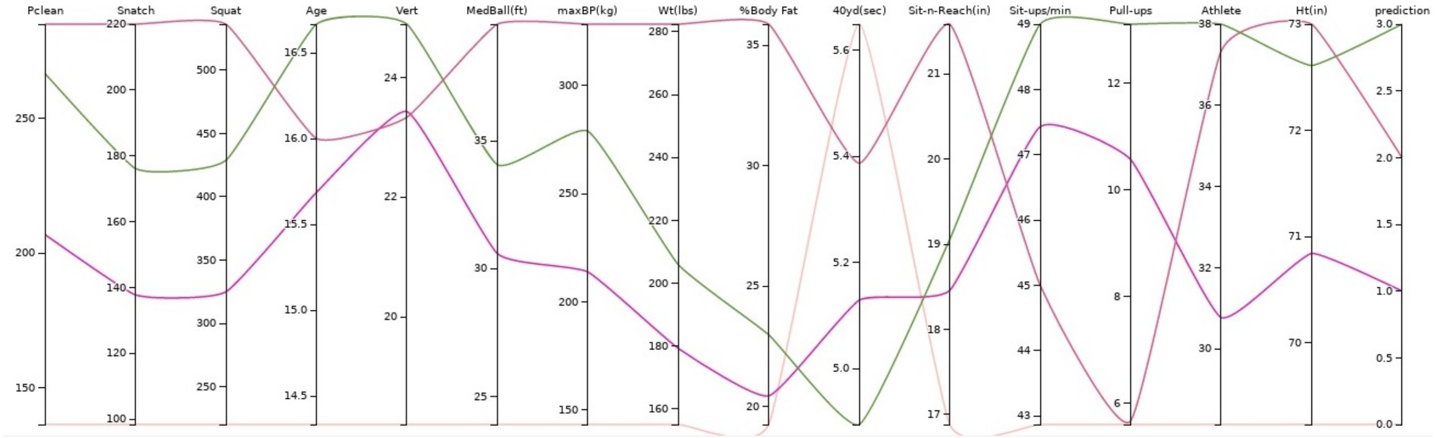


Figure 11: K-means clustering using Minimum crossing pairs reordered dataset

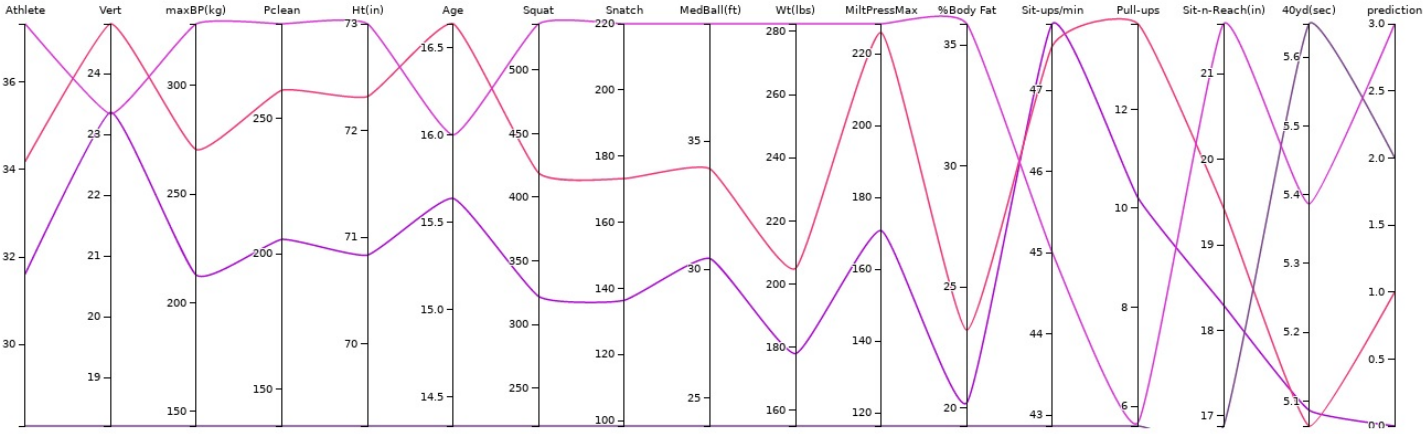


Figure 12: K-means clustering using Pearson correlation reordered dataset

From the above figures, one can see how data reordering in combination with clustering can help with identifying different trends in data and provide visualization which is much easier to interpret and interrogate. To provide an intuitive measure of reduced clutter, we compute the number of crossings seen in Figures 11, 12, 13 and 14. Table 2 shows the comparison for these datasets.

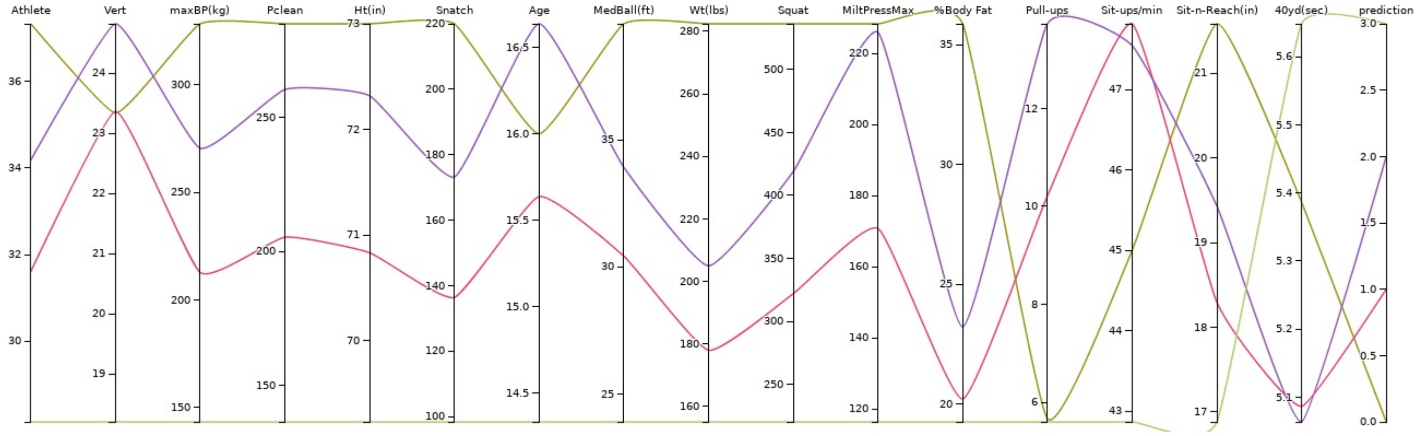


Figure 13: K-means clustering using Kendall correlation reordered dataset

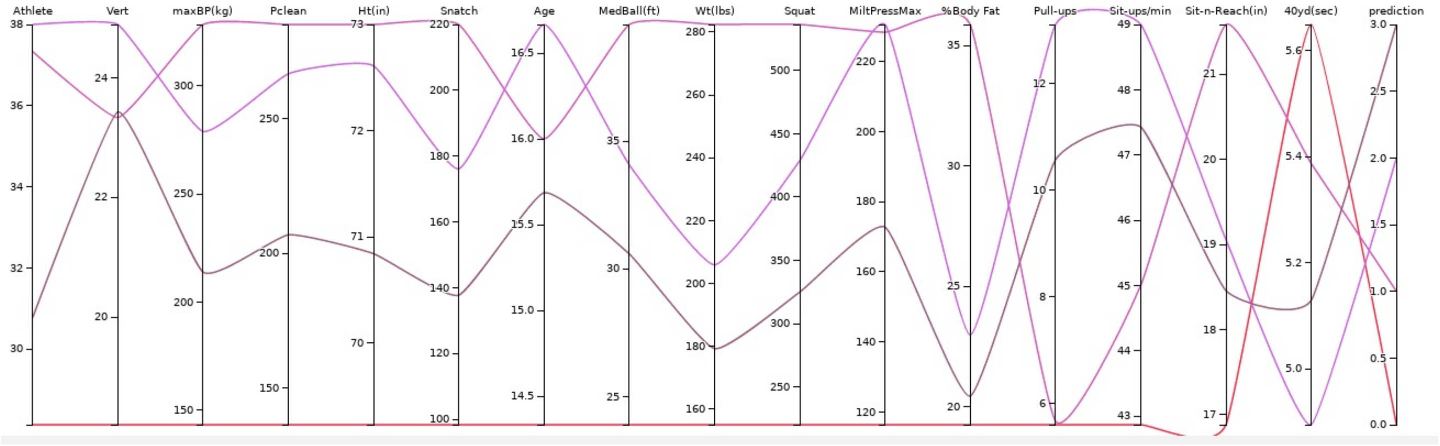


Figure 14: K-means clustering using Spearman correlation reordered dataset

Table 2: Number of crossings comparison

Applied Techniques	**No. of crossings**
Pearson correlation	16
Kendall correlation	16
Spearman correlation	19
Minimum crossing pairs	11

V. Conclusion

The purpose of the conducted exploratory research reported in this paper was an investigation of the complementarity of different visualizations using parallel coordinates plots (PCPs) and the power of combining various data analytics techniques (in our case variable reordering and clustering) in identifying interesting patterns in high volume, multi-dimensional data. While overall dataset clutter reduction techniques like directly minimizing the number of crossings can reduce the overall perceived clutter in the whole visualized dataset, the correlation based reordering techniques can have a positive effect in identifying subsets of features which when displayed next to each other could reveal interesting patterns otherwise not easily visible in PCPs. In our admittedly very preliminary, exploratory investigations, we have evaluated dimension reordering techniques including Pearson correlation, Kendall correlation, Spearman correlation and minimum crossing pairs. We have performed the experiments on Male athlete strength dataset as an illustrative example. The investigated dimension reordering approaches have been combined with k-means clustering in order to extract clear data patterns and visualize the data with sufficiently reduced clutter otherwise obscuring interesting data trends and dependencies. Our very limited results show that the minimum crossings pair approach performed better in terms of the clutter reduction on the whole dataset than the correlation based techniques. This preliminary analysis and results have been intended to illustrate how a combination of (semi-)automated data analytics approaches combined with simple though extremely versatile visualization technique like PCPs can be used for very powerful interactive interrogation of voluminous and high dimensional datasets. A more thorough investigation and development of visual analytics tools and approaches to enhance PCPs is intended as out future work.

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

[2] Siirtola, H., R¨aih¨a, K.J.: Interacting with parallel coordinates. Interacting with Computers 18(6), 1278–1309

(2006)

[3] Zhao, K., Liu, B., Tirpak, T.M., Schaller, A.: Detecting patterns of change using enhanced parallel coordinates visualization. In: ICDM, p. 747 (2003)

[4] Wegman, E.J.: Hyper dimensional data analysis using parallel coordinates. Journal of the American Statistical Association 85(411), 664–675 (1990)

[5] Moustafa, R., Wegman, E.: Multivariate Continuous Data - Parallel Coordinates. Springer, New York (2006)

[6] Johansson, J., Ljung, P., Jern, M., Cooper, M.: Revealing structure within clustered parallel coordinates displays. In: IEEE Symposium on Information Visualization (INFOVIS), p. 17 (2005)

[7] Holten, D., Van Wijk, J.J.: Evaluation of Cluster Identification Performance for Different PCP Variants. Computer Graphics Forum 29(3), 793–802 (2010)

[8] Fua, Y.-H., Ward, M.O., Rundensteiner, E.A.: Hierarchical parallel coordinates for exploration of large datasets. IEEE Visualization, 43–50 (1999)

[9] Hurley, C.B., Oldford, R.W.: Pairwise display of high-dimensional information via eulerian tours and hamiltonian decompositions. Journal of Computational and Graphical Statistics 19(4), 861–886 (2010) 516 Y. Xiang et al.

[10]. Hurley, C.B., Oldford, R.W.: Eulerian tour algorithms for data visualization and the pairviz package. Computational Statistics 26(4), 613–633 (2011)

[11] Wegman, E.J.: The grand tour in k-dimensions. In: Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface, pp. 127–136 (1991)

[12]. Wegman, E.J.: Visual data mining. Statistics in Medicine 22, 1383–1397 (2003)

[13] Wilhelm, A.F.X., Wegman, E.J., Symanzik, J.: Visual clustering and classification:

The oronsay particle size data set revisited. Computational Statistics 14, 109–146 (1999)

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

[15]Hurley, C.B.: Clustering visualizations of multidimensional data. Journal of Computational and Graphical Statistics 13(4), 788–806 (2004)

[16] NOVOTNY M.: Visually effective information visualization of large data. In Proc. of the 8th Central European Seminar on Computer Graphics (2004).

[17] ANDRIENKO G., ANDRIENKO N.: Parallel coordinates for exploring properties of subsets. In Proc. Of International Conf. on Coordinated Multiple Views in Exploratory Visualization (2004), pp. 93–104.

[18] WONG P. C., BERGERON R. D.: Multiresolution multidimensional wavelet brushing. In Proc. of IEEE Visualization (1996), pp. 141–148.

[19] HAUSER H., LEDERMANN F., DOLEISCH H.: Angular brushing of extended parallel coordinates. In Proc. of IEEE Symp. on Information Visualization (2002), pp. 127–130.

[20] JOHANSSON J., LJUNG P., JERN M., COOPER M.: Revealing structure within clustered parallel coordinates displays. In Proc. of IEEE Symp. on Information Visualization (2005), pp. 125–132.

[21] NOVOTNY M., HAUSER H.: Outlier-preserving focus+context visualization in parallel coordinates. IEEE Trans. on Vis. and Comp. Graph. 12, 5 (2006), 893–900.

[22] BENDIX F., KOSARA R., HAUSER H.: Parallel sets: visual analysis of categorical data. In Proc. of IEEE Symp. on Information Visualization (2005), pp. 133–140.

[23] ELLIS G., DIX A.: Enabling automatic clutter reduction in parallel coordinate plots. IEEE Trans. on Vis. and Comp. Graph. 12, 5 (2006), 717–724.

[24] ANKERST, M., B ERCHTOLD ,S., AND K EIM , D. A. 1998. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. Proc. of IEEE Symposium on Information Visualization, InfoVis'98, p. 52-60

[25] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In Proc. IEEE Symp. Information Visualization (InfoVis), 2003.

[26] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE Trans. On Visualization and Computer Graphics, 15:993–1000, 2009.

[27] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. IEEE Trans. on Visualization and Computer Graphics, 16:1017–1026, 2010.

[28] A. Tatu et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In Proc. IEEE Symp. Visual Analytics Science and Technology (VAST), 2009.

[29] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Trans. on Visualization and Computer Graphics, 12:1363–1372, 2006.

[30] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In Proc. IEEE Symp. Information Visualization (InfoVis), 2004.

[31] Q. Cui, M.Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. IEEE Trans. on Visualization and Computer Graphics, 12:709– 716, 2006.

[32] Tatu, A., G. Albuquerque, et al. (2009). 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 2009:59-66

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

[34] Guo, D. (2003). "Coordinating computational and visual approaches for interactive feature selection and multivariate clustering." Information Visualization 2(4):

232-246.

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

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

[37] Wang, Q., Y. Shen, et al. (2005). "A nonlinear correlation measure for multivariable data set." Physica D:

Nonlinear Phenomena 200(3–4): 287-295.

[38] Zhiyuan, S., W. Qiang, et al. (2011). Effects of statistical distribution on nonlinear correlation coefficient. IEEE Instrumentation and Measurement Technology Conference (I2MTC), 2011.

[39] Johansson, S. and Johansson, J., 2009. Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE transactions on visualization and computer graphics, 15(6), pp.993-1000.

[40] 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.

[41] 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.

[42] Zhou, L. and Weiskopf, D., 2018. Indexed-Points Parallel Coordinates Visualization of Multivariate Correlations. IEEE transactions on visualization and computer graphics, 24(6), pp.1997-2010.

[43] Richer, G., Sansen, J., Lalanne, F., Auber, D. and Bourqui, R., 2018, March. Enabling Hierarchical Exploration for Large-Scale Mutidimensional Data with Abstract Parallel Coordinates. In International Workshop on Big Data Visual Exploration and Analytics 2018.

[44] Wilkinson, L., 2018. Visualizing Big Data Outliers through Distributed Aggregation. IEEE Transactions on Visualization Computer Graphics, (1), pp.1-1.

[45] Chakuma, B. and Helbig, M., 2018, June. Visualizing the Optimization Process for Multi-objective Optimization Problems. In International Conference on Artificial Intelligence and Soft Computing (pp. 333-344). Springer, Cham.