Assignment 6

Information Visualization & Visual Analytics (WS 2020/21)

Chatterjee, Nishan
nishan.chatterjee@ims.uni-stuttgart.de

Fischer, Stefan
stefan.fischer@ims.uni-stuttgart.de

December 13, 2020

1 Task 1: British Lawns

(a) The agency presents the results by visualizing the distribution of each variable. Describe why this presentation may lack important details and why it is generally challenging to visually analyze multivariate data sets.

Dataset = [age, gender, occupation, garden size, city, current planting, household income, household size]

Visualizing the distribution of each variable often fails to provide the patterns that might exist in visualizing all the variables together. On the other hand, if all variables are plotted in a scatterplot, it may be infeasible to interpret it, as it's difficult even in 3D to visualize without rotation.

Finally, there might be dependant variables in this dataset which may be hard to determine. For instance, household size may be dependant on household income, and household income may depend on occupation, age, and gender.

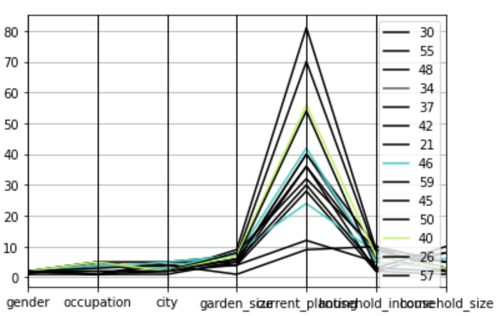
(b) Sketch how you would present the results to better visualize the relationships within the data sets (you can come up with fictitious data). What are the advantages and disadvantages of your approach?

With the use of brushing, we can compare between entries. An example of this has been mentioned using a dummy dataset by highlighting age and occupation. In the Age diagram, we can also see brushing for specific entries. We can also easily observe at a glance the coorelation between axes which are not neighbours.

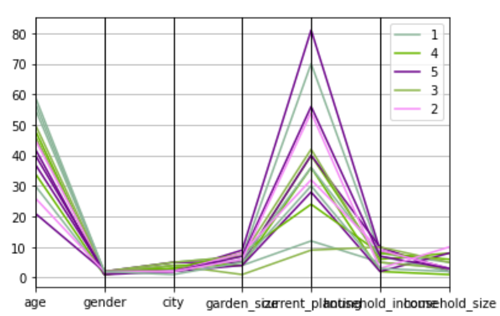
The problem with this approach is that for dataset containing thousand of entries, it becomes overplotted and hard to understand.

	age	gender	occupation	city	garden_size	current_planting	household_income	household_size
0	24	M	Engineer	Stuttgart	5	15	9973	8
1	59	M	Teacher	Stuttgart	1	1	5642	6
2	21	M	Doctor	Berlin	8	24	2695	10
3	22	F	Doctor	Munich	3	3	5543	9
4	29	M	Scientist	Stuttgart	2	16	9602	9
5	25	M	Doctor	Hamburg	3	9	7955	5
6	59	F	Engineer	Berlin	1	2	4260	7
7	53	M	Scientist	Cologne	5	40	6941	10
8	27	F	Business	Munich	7	42	3719	6
9	32	F	Teacher	Cologne	1	9	9475	3
10	28	M	Teacher	Berlin	3	6	9111	10
11	23	M	Engineer	Hamburg	7	21	5045	3
12	23	M	Business	Cologne	5	20	6647	6
13	23	F	Scientist	Stuttgart	7	21	6261	5
14	47	M	Engineer	Cologne	2	8	4152	10

Figure 1: Dataset used for generating Parallel Coordinate Plot



(a) Age with Brushing of Certain Values



(b) Occupation

Figure 2: Parallel Coordinate Plot generated by using the above dataset

2 Task 2: Scatter Plot Matrices vs Parallel Coordinates

(a) Describe which advantages and disadvantages Scatter Plot Matrices (SPLOMs) and Parallel Coordinates (PCPs) offer compared to each other.

Parallel coordinates are designed for visualizing higher dimensional data, but the increase in polylines makes it very difficult to understand as it becomes overplotted. Also, in PCP display, it is difficult to determine if a pair of data dimensions are strongly correlated unless the corresponding axes are adjacent.

Scatterplot Matrices are intuitive to understand. However, they lack scalability. Scatterplots are more resistant to overplotting, but all one-to-one maps are succeptible to overplotting eventually.

In a survey paper by Li et al.[2] comparing single scatterplots with PCP displays of 2 axes (only), they found that participants were better at accurately assessing degree of correlation between pairs of data dimensions with the scatterplot than with the PCP. The displayed stimuli were not interactive and the scatterplots and PCP were shown in isolation (a combined view was not tested).

Another study by Kanjanabose et al.[3] compares user performance, in terms of accuracy and response time, in the context off our different visualization tasks, using either PCP or SPLOM (but,again, not both together). Their results suggest that PCP is better than SPLOM for cluster, outlier and change detection.

A radial version of PCP called Stardinates was empirically tested against regular PCPs by Lanzenberger et al. [4] - including a combination. The two views are very similar to one another (more similar than PCP and SPLOM), and hence, not particularly complementary.

(b) Given a data set with many variables, how can you improve the visualization to support the analysis? Name and describe at least two techniques that can be applied to both SPLOMs and PCPs.

Using dimentionality reduction, visualization of both SPLOMs and PCPs can be improved. One method of dimentionality reduction is by doing the Principle Component Analysis used for finding the principle components of a dataset. For instance, a dense PCA plot can be reduced in the following way.

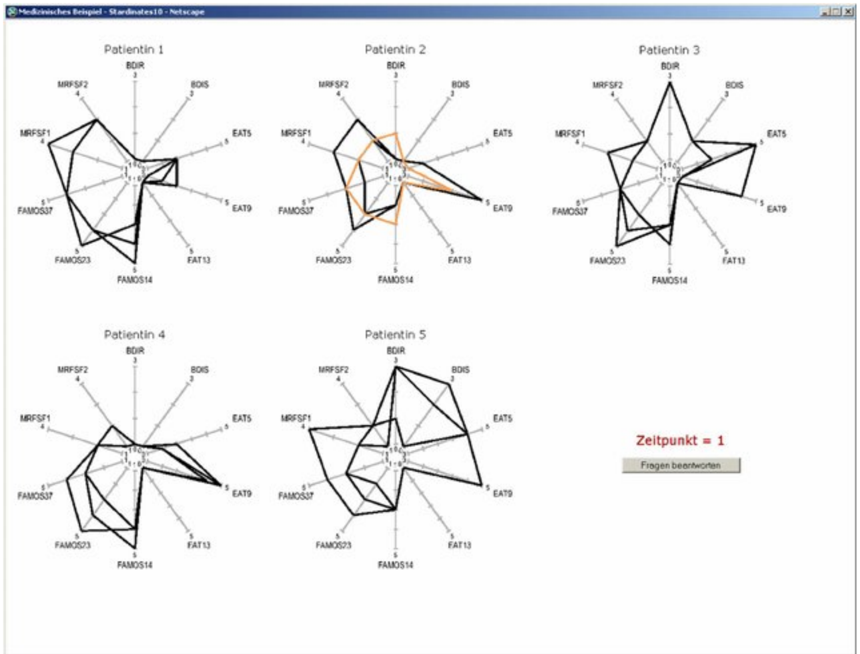
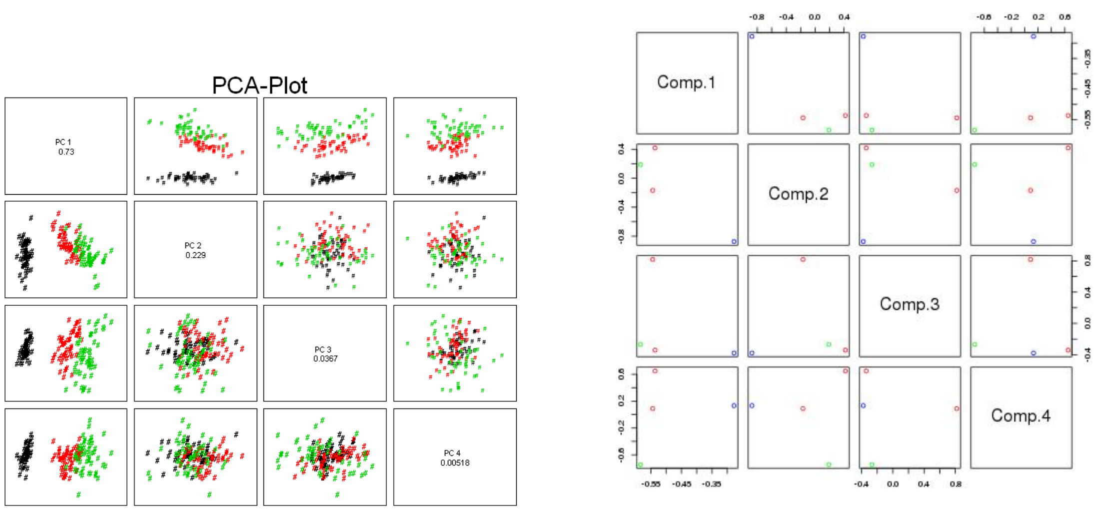


Figure 3: Psychotherapeutic data of five patients at three different times visualized by the Stardinates, the first measurement of patient ID 2 is highlighted



(a) Before (b) After

Figure 4: After applying PCA to SPLOM

The same technique can be applied for SPLOMs as well to visualize only the relevant variables with respect to each other.[5]

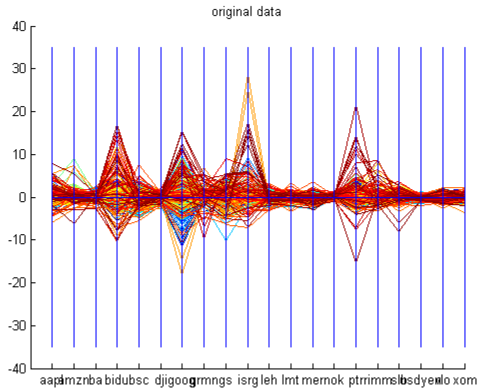
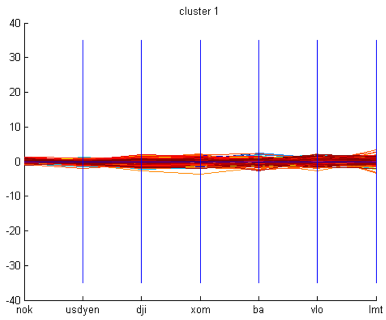
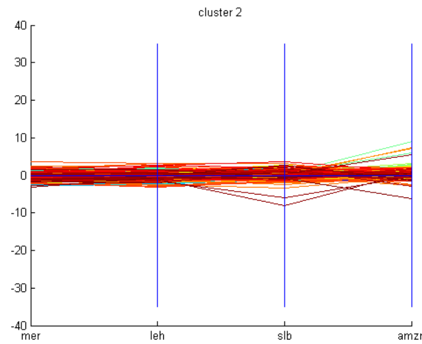


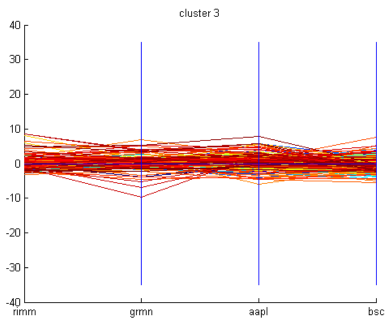
Figure 5: PCP of Financial data of the following companies: Apple, Amazon.com, Boeing, Baidu.com, Bear Sterns, Dow Jones Industrial Average Index, Google, Garmin, Goldman Sachs, Intuitive Surgical, Lehman Brothers, Lockheed Martin, Merrill Lynch, Nokia, Petro China, Research in Motion, Schlumberger, USD-Yen, Valero, and Exxon Mobil from October 29 th 2006 to October 28th 2007.



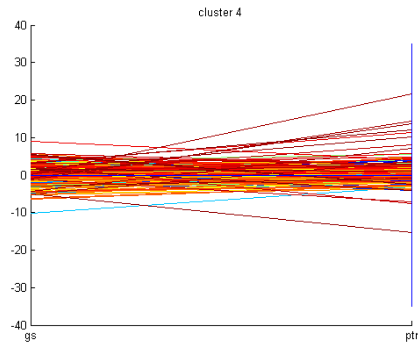
(a) Cluster 1



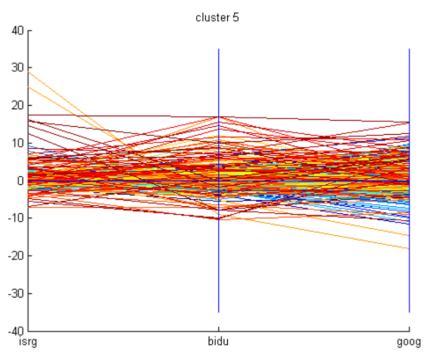
(b) Cluster 2



(c) Cluster 3



(d) Cluster 4



(e) Cluster 5

Another way to improve visualization of SPLOMs and PCPs is by combining the PCP-SPLOMview by using brushing such that when lines in the PCP view are selected or hovered-over, the points for the corresponding data items are also highlighted in the SPLOM view. Conversely, selections or hovers in the SPLOM view are also highlighted in the PCP view. [1]

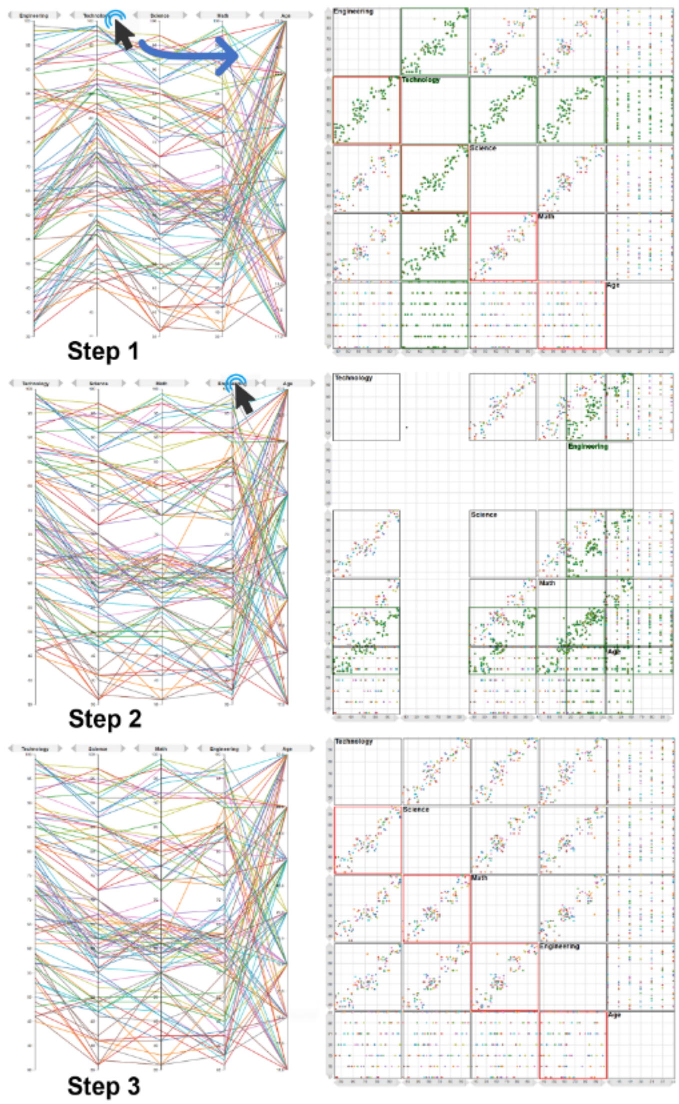


Figure 6: Reordering axes to form an animated Combined view.

References

[1] An Evaluation of Perceptually Complementary Views for Multivariate Data - Chunlei Chang et al.

[2] Judging Correlation from Scatterplots and Parallel Coordinate Plots - Jing Li et al.

[3] R. Kanjanabose, A. Abdul-Rahman, and M. Chen. A multi-task compar-ative study on scatter plots and parallel coordinates plots.

[4] M. Lanzenberger, S. Miksch, and M. Pohl. Exploring highly structureddata: a comparative study of stardinates and parallel coordinates.

[5] PARALLEL COORDINATES, ECE 491 Fall 2007, Cornell - Ankur Kumar