Interactive Multivariate Data Analysis and Visualization

Group Information:

Di Wang, u1072369@utah.edu, u1072369;

Shenruoyang Na, u1142914@utah.edu, u1142914

Project Respository: https://github.com/orpheus92/RegulusHD

Background and Motivation:

One important application of information visualization is that it helps domain experts understand multivariate data, which is hard to visualize in conventional ways. Dimension reduction method such as PCA can help understand the data, but some features of the high dimension space are lost through the dimension reduction process. The main inspiration of this project is the previous work *Visual exploration of high dimensional scalar functions* by Samuel Gerber. In this paper, a method that involves computation of Morse Smale Complex is used to cluster high dimensional data by calculating the significance level of features shown in the data. Then an inverse regression method is used to show the behavior of the point clouds in different clusters. The existing tool that uses this algorithm is written in C++ and python, which does not provide powerful interaction as what javascript does in frontend. Designing such interface will help users do better analysis on multivariate data and understand the high dimension space.

The dataset used for this visualization is not limited to any single dataset. The first dataset that is going to be analyzed is related to nuclear simulation are obtained from Nuclear Energy University Program. Other multivariate datasets of users' interests from different fields can also be used for this visualization.

Project Objectives:

There are many existing visualizations that people have implemented to show multivariate data. However, many of the existing visualizations are domain specific, which means these visualizations only help the users with limited datasets (some visualizations are only created for one dataset and cannot be used for any other purposes). Thus, we are trying to create an interface that is capable of doing multivariate analysis and can be reused. The visualization itself is also capable of conveying important information of the dataset. The main goal is to help the users understand more about the high dimensional space and the relationship between the dependent variable and independent variables in the dataset. One thing we want to achieve at the same time is to use promise for the interaction, which would make our project more flexible.

Data:

The first dataset we want to work with is a nuclear dataset Di gets from his NEUP project. Other datasets related to price of the stock market will be analyzed later. Tushare is a python library that can be used to generate historical and Real-time Quotes data of China stocks. The corresponding repository is https://github.com/waditu/tushare. We are able to generate multivariate dataset stored as csv files.

Data Processing.

The main requirements for the dataset we use are that each attribute will need to be a numerical value. There should also be a numerical dependent variable. For the dataset we need to analyze, there is no preprocessing process that is required for this. However, since we do not want to implement the whole algorithm that involves the computation of Morse Smale Complex, we would run the algorithm in python locally to generate a json file that specifies how are the points clustered in the high dimensional space. If the have enough time before the end of the semester, we will try to figure out whether there is a way to make the javascript communicate with the python kernel so that the computation of Morse Smale Complex does not need to be done locally.

Visualization Design.

Layout:

There are many ways to visualize multivariate data. Our idea is to include four major blocks in the design. With the block A contains some button that lets users load the data. After the data is loaded, the information of this data such as number of dimensions, sample size as well as the name of each attribute will shown in this block. Block B and C will contain two type of visualizations of the input data. Block D will include some buttons for interaction commands that the users can play with. The basic framework is shown in. Figure 1.

Block A **Load Data** ● Data Inforamation: ● Selection Information:	Block B **Visualization A** Drop-down box ● Click/Hover/Brush for update
Block C **Visualization B** Drop-down box ● Click/Hover/Brush for update	Block D **Interaction Commands:** ● Brush for selection ● Feature 1... ● Feature 2...

Figure 1. Draft Version for the Final Layout of our Design

Focus

As mentioned before, the main focus of this project is to create an interactive interface that users can use to analyze multivariate data. This interface can be reused by different people to work on different datasets. The novel part of this design is to include the algorithm mentioned in the paper *Visual exploration of high dimensional scalar functions* to help the users understand the dataset better. By computing the corresponding Morse Smale Complex for the dataset, a persistence value will be calculated for each cluster that represents the significance level of the feature it represents. Then the users are able to select the specific cluster they are interested in to analyze with the interactions that d3 provides.

Sketches

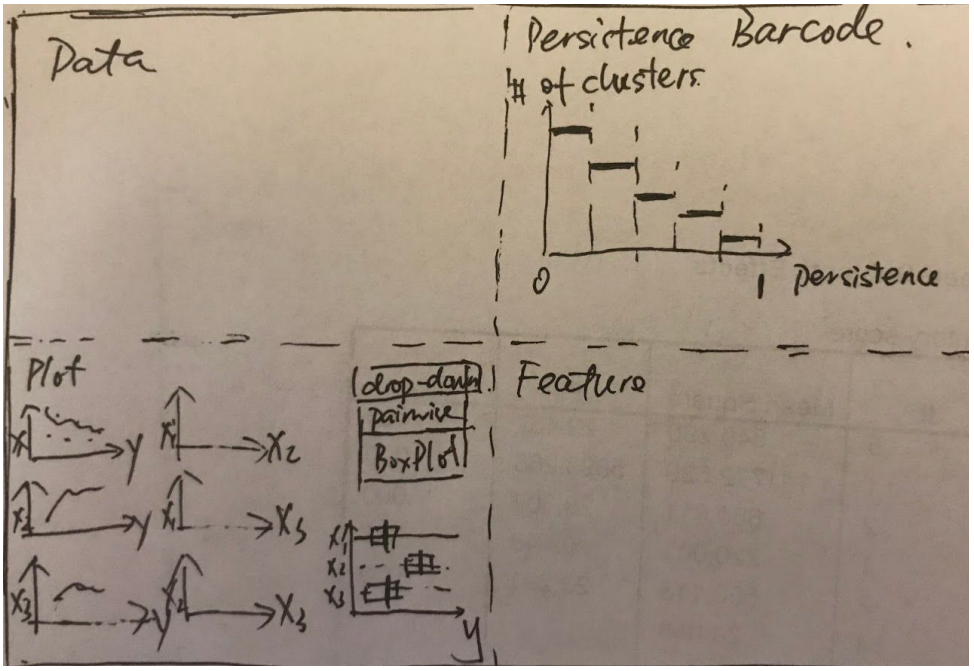

Figure 2. Sketch 1 for the Design

The first sketch is just comprised of certain visualizations from the original paper. Block B includes a chart for persistence barcode is used to show the merging/splitting process of different clusters (in the paper referred to as crystals). However, this visualization lost the geometric property of the data and it was not clear which cluster splits or merges. The plots are capable of giving some information of the data, but other visualizations may also be tried before deciding the final design.

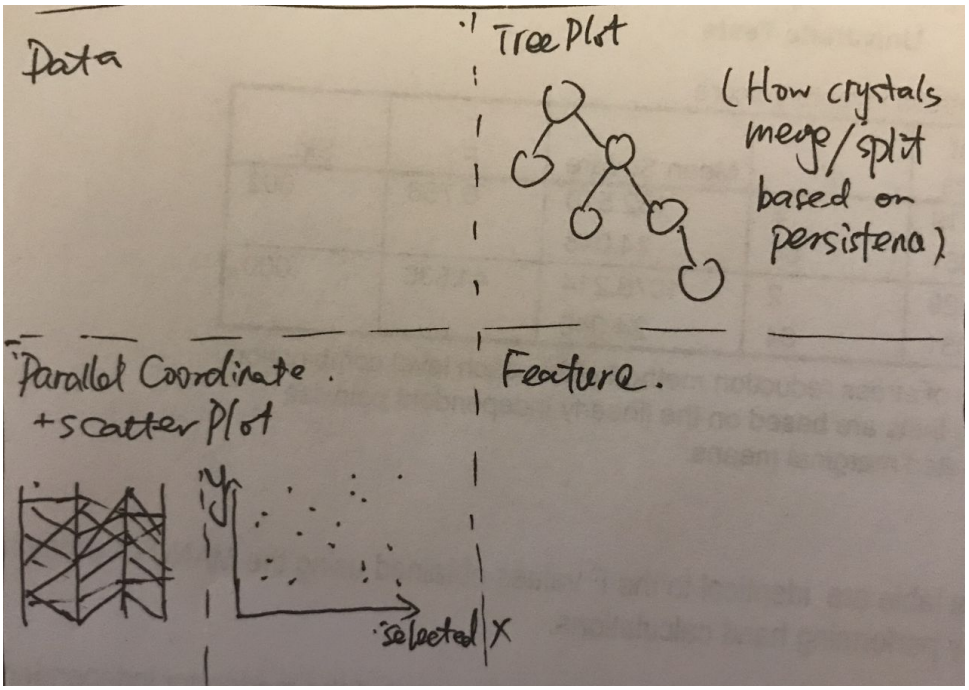

Figure 3. Sketch 2 for the Design

Since the persistence barcode is not something easy for the users to understand and it does not show how crystals are merged/splitted based on persistence, we want to think of some other way to visualize this. One option we come up with next is the tree plot, which will fit our design and show the information we want. For Block C, we are trying parallel coordinates witg scatter plot. There is no specific reason for this. We would just like to try different sketches before we decide which one we want to use.

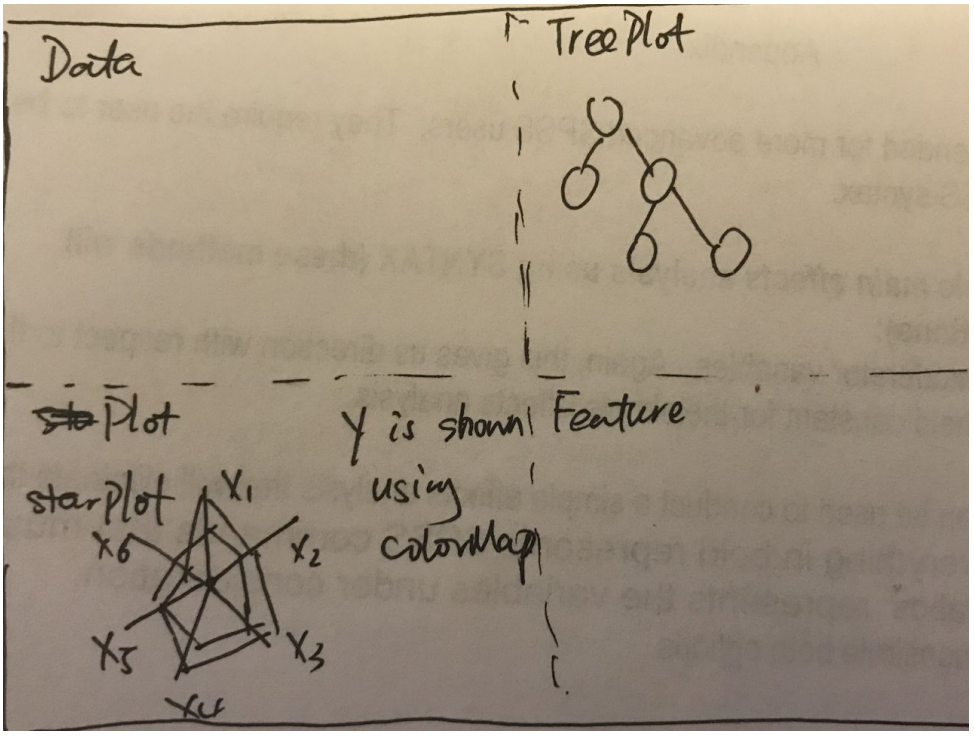

Figure 4. Sketch 3 for the Design

We did not come up with better idea for Block B and kept the tree plot. However, we may include star plot for Block C, which might look neater than parallel coordinates in some sense. We also use the colormap to represent the value of our dependent variable of interest. This might show some interesting feature of the dataset.

Discussion

The treeplot seems a good visualization technique for our interface. We compare the treeplot with dendrogram and find that the traditional treeplot seems more intuitive for this case. However, when the tree is really long, there might be some problem with it. That is something we will keep in mind.

As we show with the three sketches, there are many ways to visualize the points. When we talk with the people that are working on high dimensional data analysis, many of them prefer the traditional ways (as the one in Figure 2). One reason behind this is that it is straightforward and easy to understand. It is still robust when the size of the data increases. Also, if we are going to fit some mathematical model to certain clusters, the best way would be showing it directly in the 2D plot. However, we will have a drop-down menu for different kinds of traditional plots that users might be interested in.

Final Design

Our final design is shown by Figure 5 with some instructions in each of the block. We switch the location of Block C and Block D because we think the users always prefer have the interactions, buttons or commands on the left side and visualization on the right side.

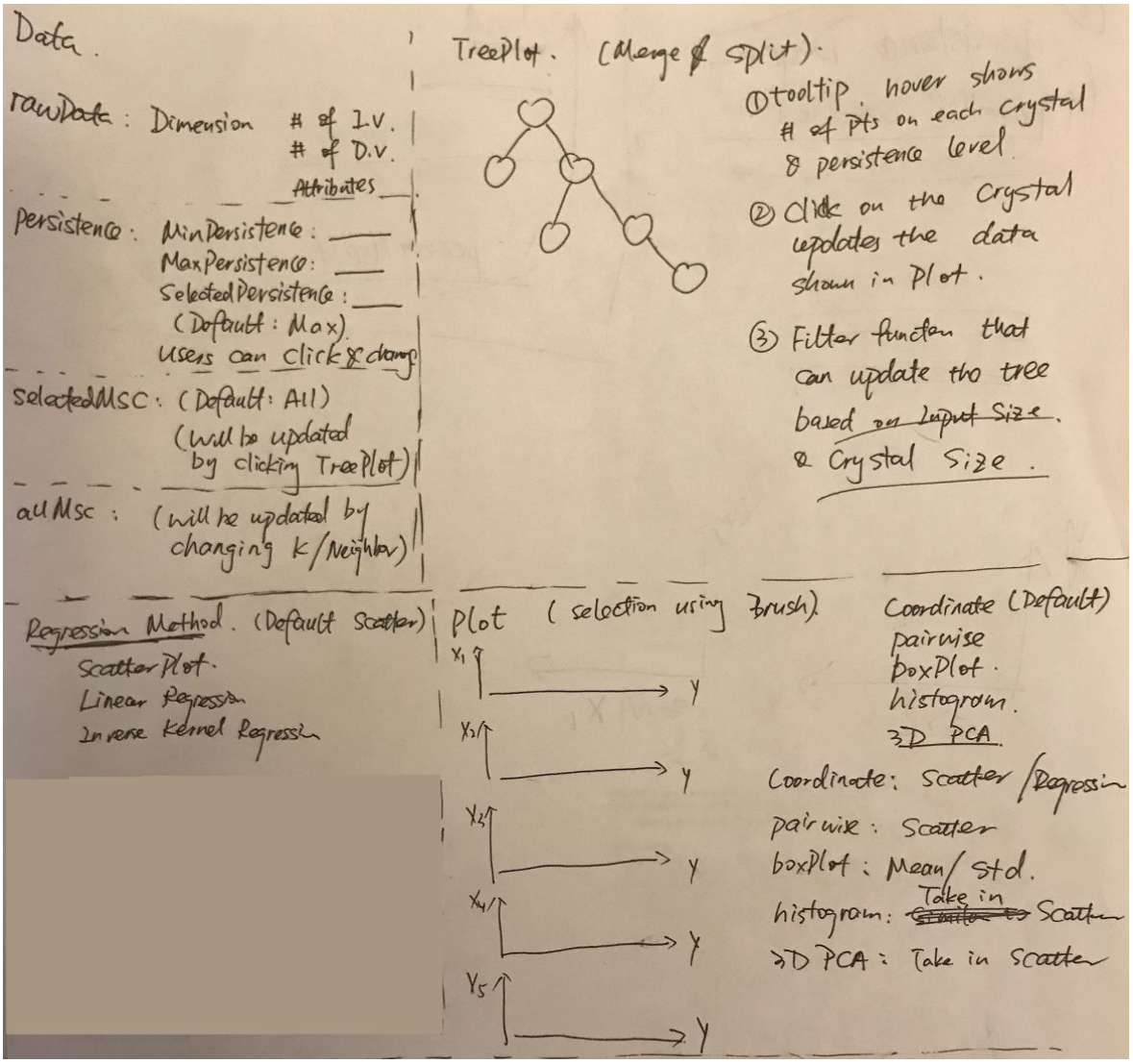

Figure 5. Draft for Final Design

Must-Have Features.

- Since the major application for this design is to analyze multivariate data, the first must-have feature is the interaction between different visualizations. For example, selecting specific cluster in the treeplot will update the data that is shown in other visualizations. Similarly, selecting specific points on the plot will also show which cluster are those selected points located. By selecting different attributes from the data, it will also change the number of independent variables and update the visualizations accordingly.

- The design should support multiple kinds of plots so that the users can select their prefered methods to visualize the data. For example, the users could choose to use different plots such as scatter plot, histogram or box plot with the drop-down box.
- According to the algorithm mentioned in the paper, there is a unique persistence value for each cluster. The clusters could merge or split depending on the persistence level that the users are interested. Specifying such value will update the tree structure.
- For some cases, there will be some outliers in the dataset. As a result, certain clusters will end up only having a few points. However, sometimes people prefer not to see those outliers in their analysis. Thus we want to have a filter function in our design such that users can use to get rid of the part of the data that they consider as noise.

Optional Features.

Besides the scatter plot, sometimes people would like to see the relationship between dependent variable and the independent variables. Fitting specific regression curve may help users understand the data better. The first optional feature would be fitting the regression curve of user's interests to the scatter plot. The kind of regression methods we have in mind so far are linear regression and kernel regression. We would like to make a drop-down box and incorporate functions to do regression.

The second optional feature would be adding PCA to the visualization. Although some features may be lost during dimension reduction, PCA still gives the users some sense of how the dataset looks like in high dimensional space. The concept of PCA is not that hard, but creating a good visualization of it that allows user's interaction is not simple. Together with the PCA, the design should also allow interaction or communication between the PCA projection and other visualizations

Project Schedule.

There are six weeks in total for our project with the milestone due in three weeks. To avoid rushing at the end of the semester, we decide to meet twice per week and keep progress report for this project. Sometime early in the week, the group members will meet and discuss about the achievable goals for each week. Sometime later in each week, the group members will meet to update about their progress and challenges.

For the first three weeks, we will focus on the main framework of the design and get each part of the design and interactions between different visualizations working. We will try to have a working prototype that we can play with before the project milestone. We will also include our must-have features during this time.

For the remaining three weeks, we will focus on the optimization of the project to make the interface more user-friendly. We will also work on optional features as we mentioned before

during this time. If possible, we will explore more datasets and get feedbacks from other people for our final design.