Process book

* Overview 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. We got some good feedbacks from through the peer review session about the size of the data to visualize. When the size of the dataset is too large, we would have a function that only selects certain samples from the whole dataset to visualize.

* Questions:

Could we get some meaningful relationships between different attributes of the data? how is each dimension related to each other?

How is the data distributed in the space, how are the points clustered in the high dimensional space? Are the tree view we try to create consistent with the geometric location of data in high dimensional space?

Can our visualization help to answer the above questions clearly? Which part can be addressed well and which part is not?

The original method from the paper calculates the clusters and store the data into local directory. Can we update the tree structure during runtime?

* Data: Source, scraping method, cleanup, etc.

We start with nuclear simulation dataset of 10000 points. Each attribute is numeric. Preprocessing is not required. However, the current available algorithm that calculates the clusters is not in javascript. Since we do not want to spend a long time rewriting the algorithm in javascript or dealing with python server, we decide to treat this part as preprocessing and run the algorithm locally, which is also the advice from the TA. The result is stored in a json file that gives specification of how the clusters would merge or split as persistence changes. The front-end requires both raw data in the form of csv as well as the json file.

Although there is not tree structure mentioned in the original paper, we try to create the treeview first since it should give some notion of the algorithm. Since what we want to show through other views are clusters existing in the current tree structure, the tree is the main view that all other views need to be linked to.

There are also other views such as scatter plot, histogram and boxplot in our proposal, which we will also put in our initial prototype as the first step.

Together with the treeview, we also think of adding a view that shows the clusters in high dimension space.

Currently we have the first dataset we would like to analyze together with its processed version that contains the information of how clusters merge or split. We believe the existing algorithm should be robust for other multivariate data.

We have several views that show different aspects of the dataset. However, the tree view does not fully function yet. After the preprocessing, we have the specification json file. We have written the algorithm that updates the clusters in python, which is It will read a user’s input of persistence level and update the number of clusters, with which a tree will be updated. We will convert it to javascript soon. After the tree is implemented, all other views will connected to the tree for interaction.

* Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?
* Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?
* Implementation: Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.
* Evaluation: What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?