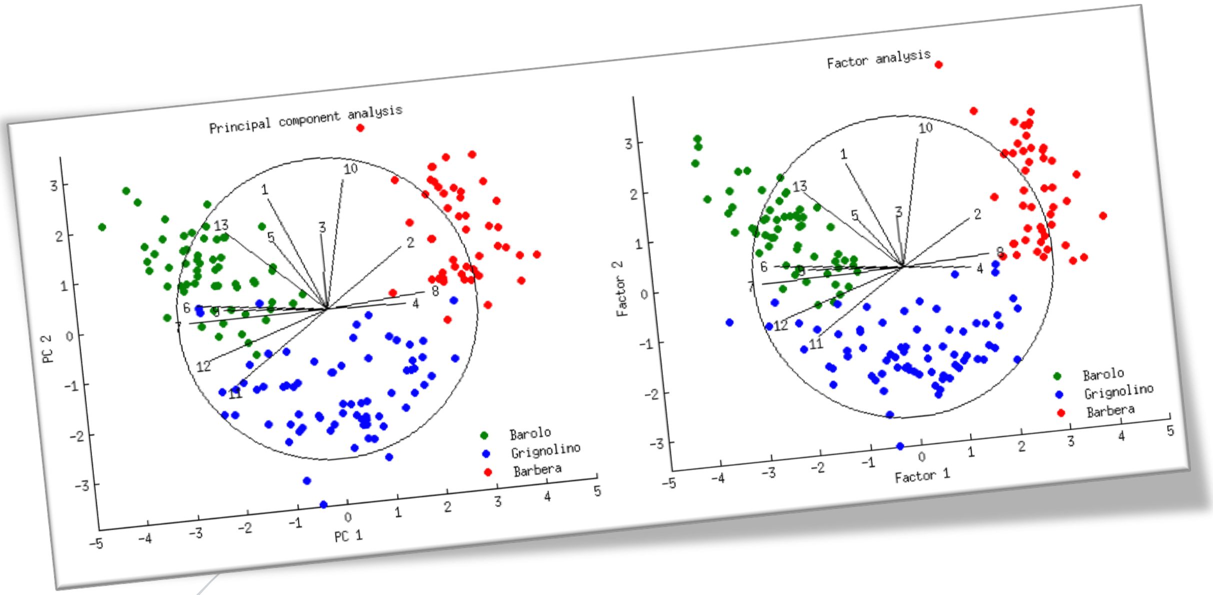
10/2/2018

Principal Component Analysis Report



Submitted By:

Kautuk Desai	50247648
Sujay Vijay Purandare	50205931
Palwinder Singh	50247454

Algorithms and their Implementation

1. **PCA**
   - We implemented the PCA algorithm to obtain new 2-Dimensional co-ordinates of the original attributes.
   - Each point is colored based on the disease it represents in the provided data (we have different colors for different diseases).
   - The scatter plot signifies the two principal components with the maximum variation (component 1 and component 2 which are labeled on the axis).

2. **SVD**
   - In SVD the components show more variations than the PCA.
   - Again, each point is colored based on the disease it represents in the provided data (we have different colors for different diseases).

3. **t-SNE**
   - Now, for t-SNE we implement a function using TSNE package (reference*: https://medium.com/@luckylwk/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b)*
   - Next, we set number of iterations to 1000 to optimize the resulting clusters.
   - Perplexity with values like 30, 40, 50 and 60 were tried but well-defined clusters were obtained at the default value of 30.
   - To get well defined cluster learning rate was set to 100.
   - Each point is colored based on the disease it represents in the provided data (we have different colors for different diseases).

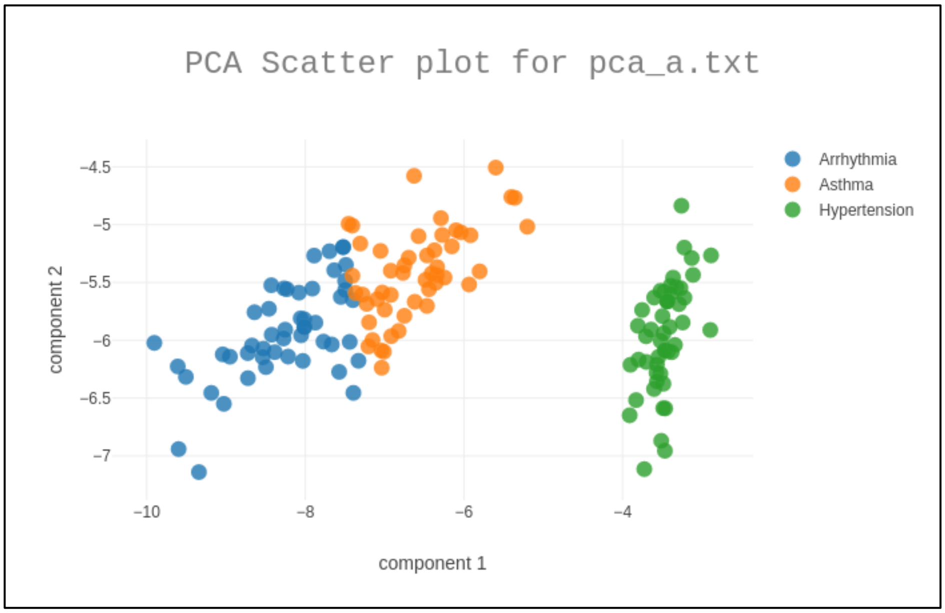
Inference

1. **Principle Component Analyis (PCA) compactly** represents the ways original data deviates from the mean, therefore corresponds to centering the dataset and then rotating it to obtain points with maximum variance as the top principal components.
2. **Singular-value decomposition (SVD**) in simple words corresponds to compactly summarizing the data in the way it deviates from zero. The results and the plots will be similar to the PCA algorithm.
3. The t-SNE algorithm focusses more on the nearest neighbor accuracy.
4. We program a function to get the t-SNE scatter plots by setting parameters.

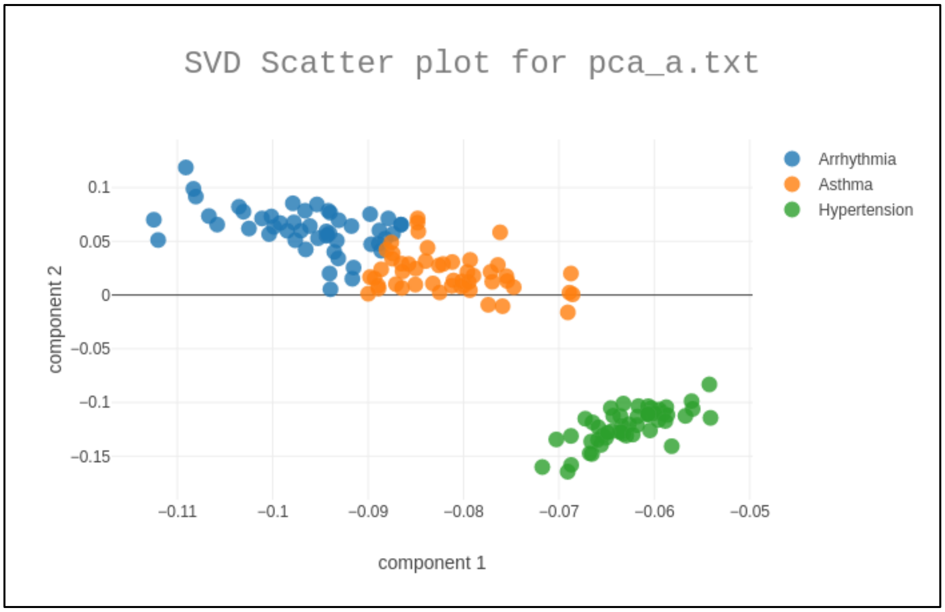
Scatter Plots

1. Pca_a.txt
   Diseases: Arrhythmia, Asthma, Hypertension

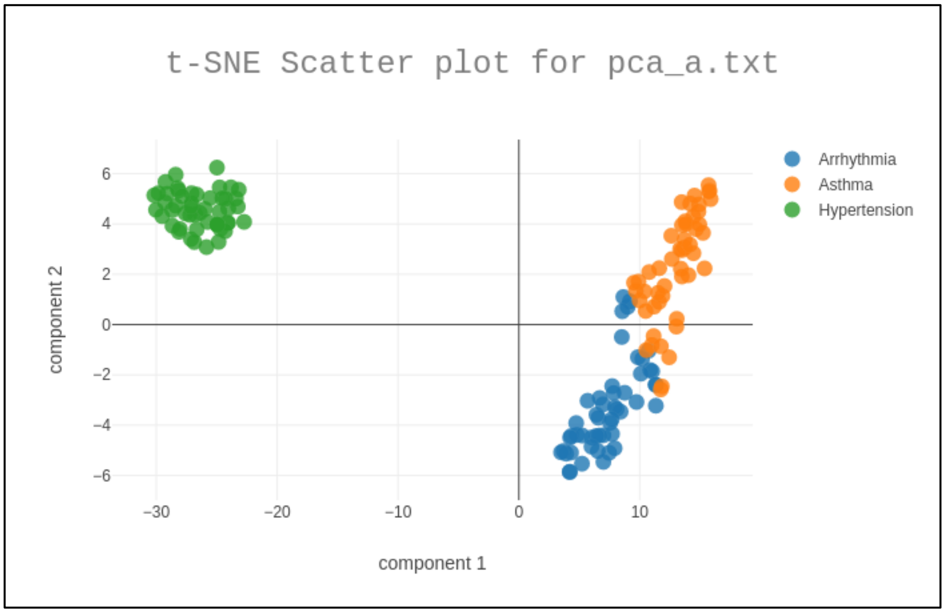
   Algorithm: PCA



   Algorithm: SVD

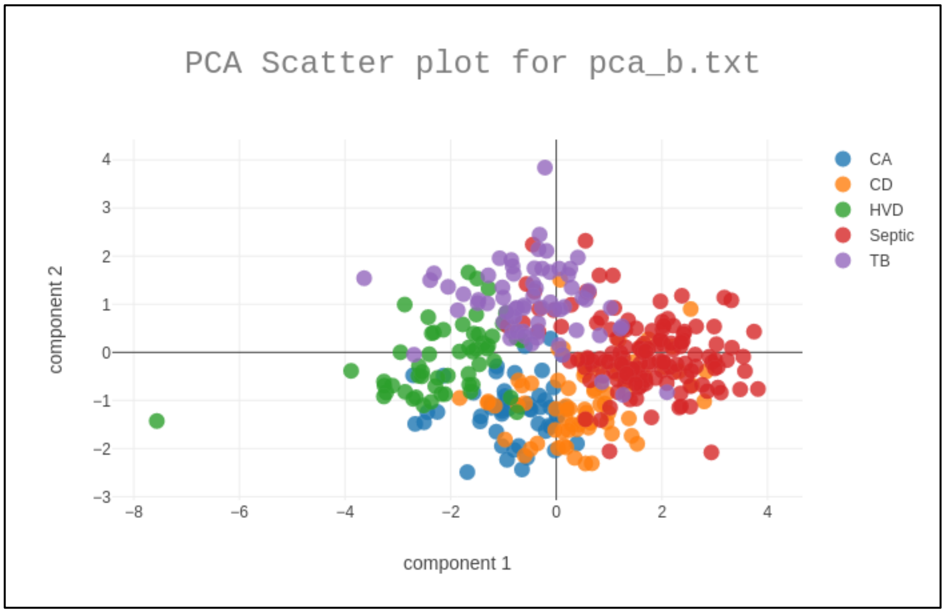


Algorithm : t-SNE

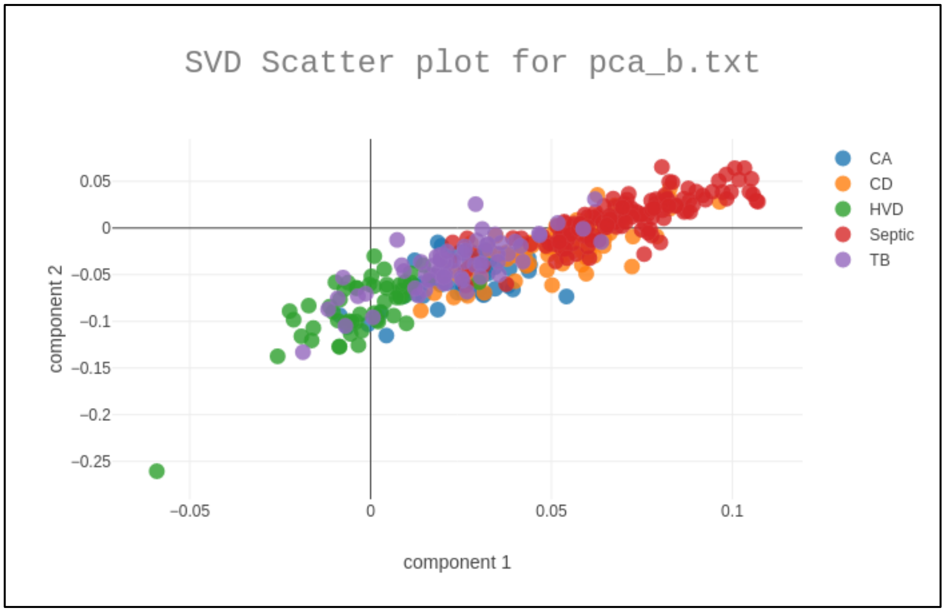


2. Pca_b.txt
   Diseases: CA, CD, HVD, Septic, TB

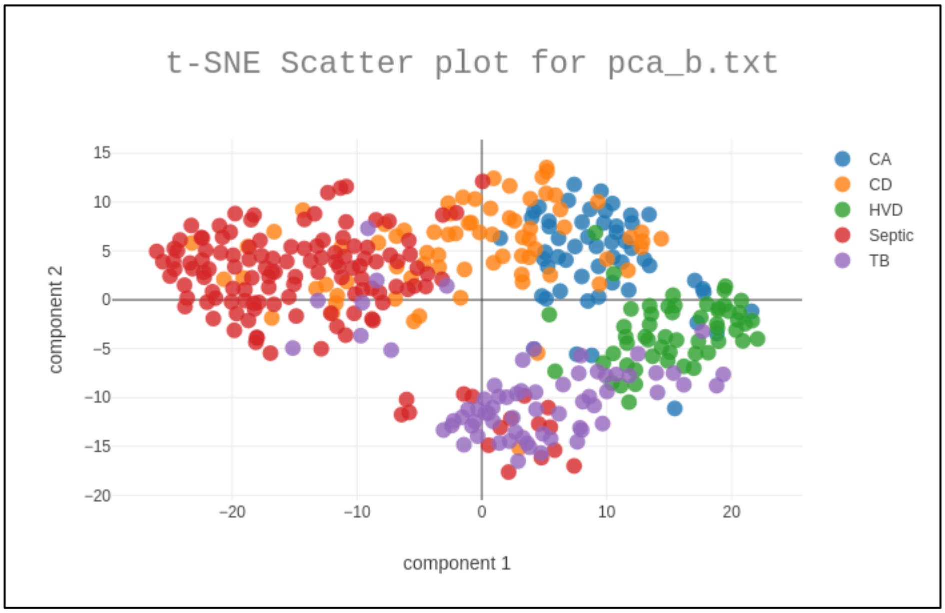
   Algorithm: PCA



Algorithm: SVD



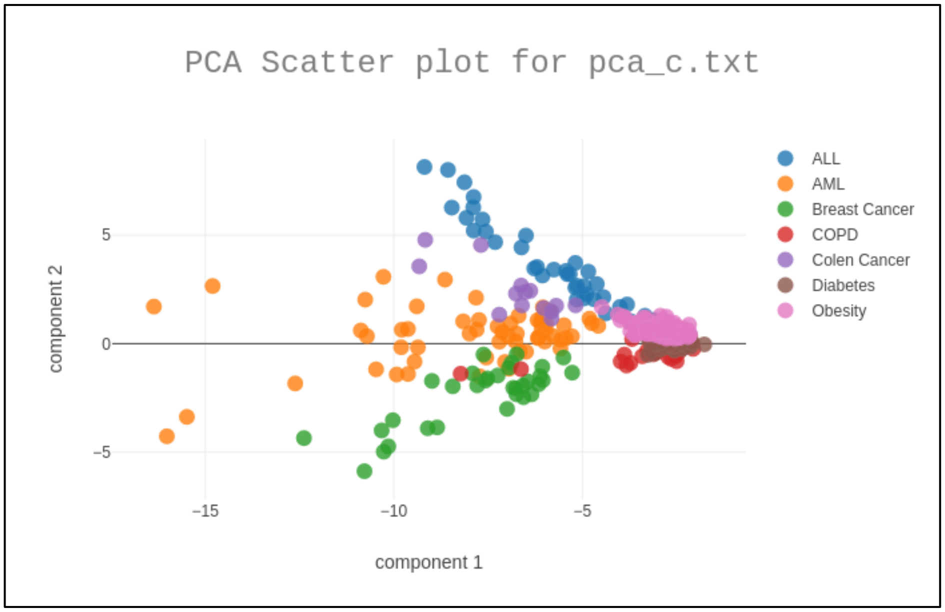
Algorithm : t-SNE



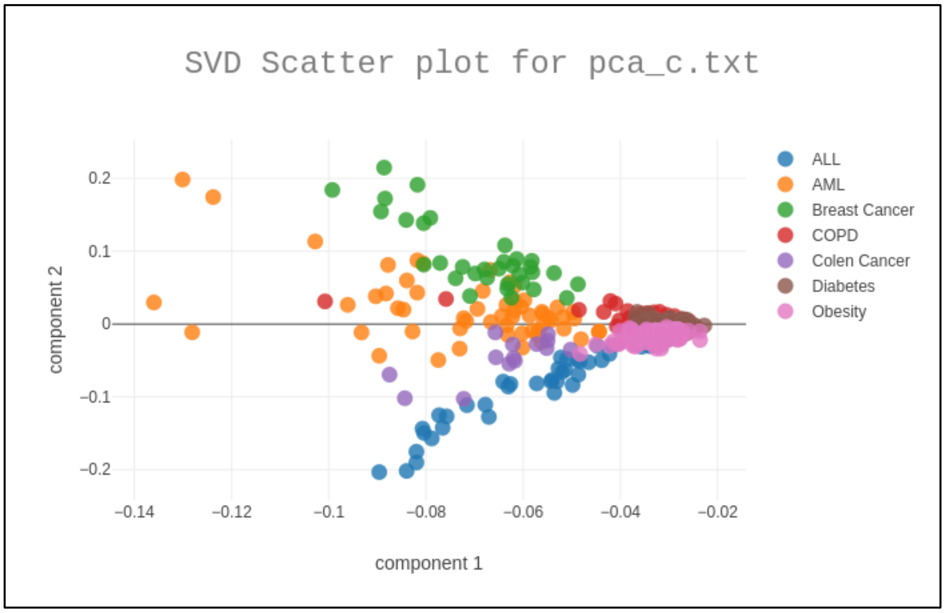
3. Pca_c.txt

Diseases: ALL, AML, Breast Cancer, COPD, Colen Cancer, Diabetes, Obesity

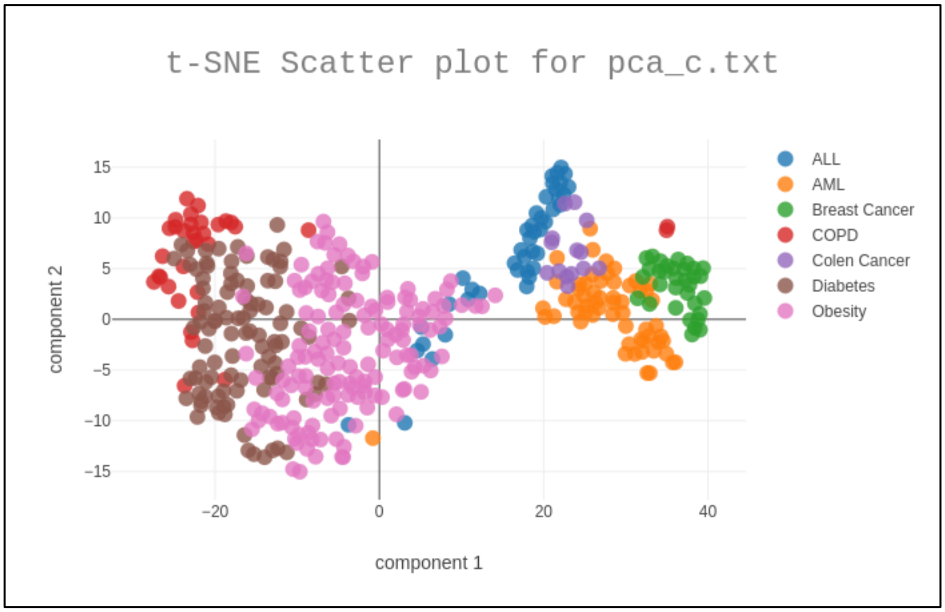
Algorithm: PCA



Algorithm: SVD



Algorithm: t-SNE


t-SNE Scatter plot for pca_c.txt
component 2
component 1
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity