DATA501 - Design Proposal

Name: Chau Thi Mai Tram

Student ID: 300643163

1. Introduction

Data visualization is an effective way to explore data for data scientist. However, it is difficult for humans to interpret visuals in more than three dimensions space. We usually plot data with box plot, scatter plot, and others, which are most feasible for two dimensions space. Although it is possible to encode more dimensions into visual by colour, shape, size; there is still limit on how many dimensions can be visualized, also this jamming practice is not a good approach. The real-world data rarely has just two or three dimensional. With the exploration of computing capacity and storage, more and more data are captured in high dimensional space and may lead up to hundreds of dimensions.

Transformation or decomposition is a useful technique to deal with multi dimension data. These algorithms transform high dimensional data to lower dimensional data while retaining as much information as possible. Transformation is benefit in representing data for interpretation, in compressing data, and in removing noise from original space.

Currently, the three transformation algorithms (PCA, NMF, t-SNE) were implemented on CRAN in three separated packages. This new package would reimplement these three algorithms to standardize them in the same interface to the end users. Besides that, this package will provide functionality on summarizing and visualizing the three algorithms results.

2. References

The ideas of these three techniques are briefly described following:

- PCA and NMF are similar in terms of mapping the features into a set of components. While PCA orders the components based on the direction of variance, NMF’s components represent additive components without intrinsic order.
- t-SNE tries to maintain the distance among data points i.e. maintain the neighbourhood.

Existing R packages:

- stat:: prcomp
- NMFN::nmfn
- tsne::tsne

Possible addition feature (maybe low priority or not feasible to implement):

- Handling the dataset with categorical columns (the existing packages return error).
- Whitening: scaling components after transform (to understand more about the use cases).
- Inversing transformation (undo rotation and add mean back to the original space) based on selected components to understand which information the component is capturing.

The developing package ideas reference to the following sources:

- Existing R packages mentioned above.
- The version of these three techniques implemented in scikit-learn python library.
- Book: Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido (Chapter 3 section 3.4).

3. Functionality

User transforms the high-dimensional dataset to a lower dimension dataset by PCA, NMF, or t-SNE algorithm with configuration:

- Number of components: the target number of lower dimension dataset.
- Scaling: to scale data before transforming.
- Ignore categorical columns.
- Whiten: to scale data after transforming (scaling components).

User views summary and visualization for exploratory data analysis:

- Using plot function for transformer object.
- Using summary function for transformer object to view the transformation attributes.

4. Appendix:

Figure 1 shows result of three techniques from three existing CRAN R packages on the *iris* dataset; PCA, NMF, and tSNE from left to right.

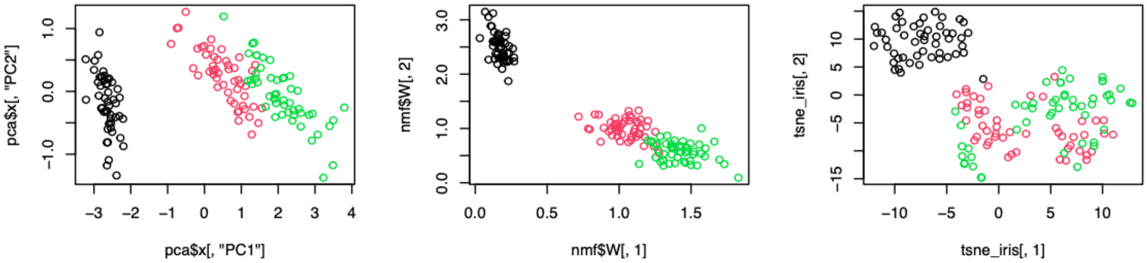
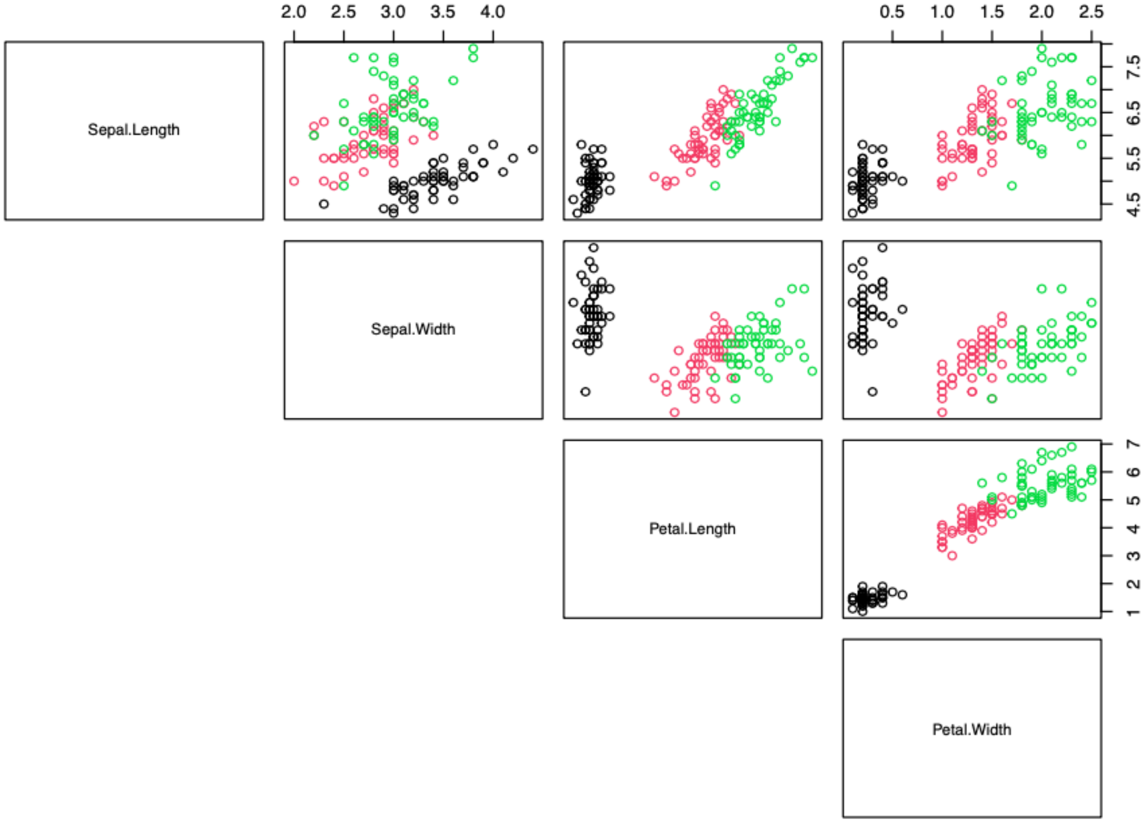


Figure 2 shows scatter plots by original features of Iris dataset.


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width