Name: Thuan Tran
CSS 490
PCA

REPORT

Table of Contents

NOTE

Run the script that I included and there will be the result of the mean and covariance of each new class. Also the image that I included below (In case they are too small to see)

Analysis of calculating mean and covariance in new space

In this case, We have Ur where Ur = U*S and U is the matrix in the new space and S is the weight of each column in the U matrix.

The new mean and covariance now have the meaning of the new features in the new transformed space. Each new mean and covariance matrix in the new space is now composed of the effects of each attributes in the old space (each column has the effects of sepal length, sepal width, petal length, petal width)

Looking at the mean, I can see that they are within a small range. Maybe +-2 ? This is because the new features are now mapped to a new space. In addition, the covariance matrix of each new class indicate that there is a correlation between each new feature. Either positive correlation or negative. Since there is no 0 in the matrix, we can say none of the features are independent toward another . Note that I am referring to the new features in the new space

The reason why we do this is because we want to perform dimensional reduction. That is we want to reduce the number of features we have to use in order to classify the flowers

Result:

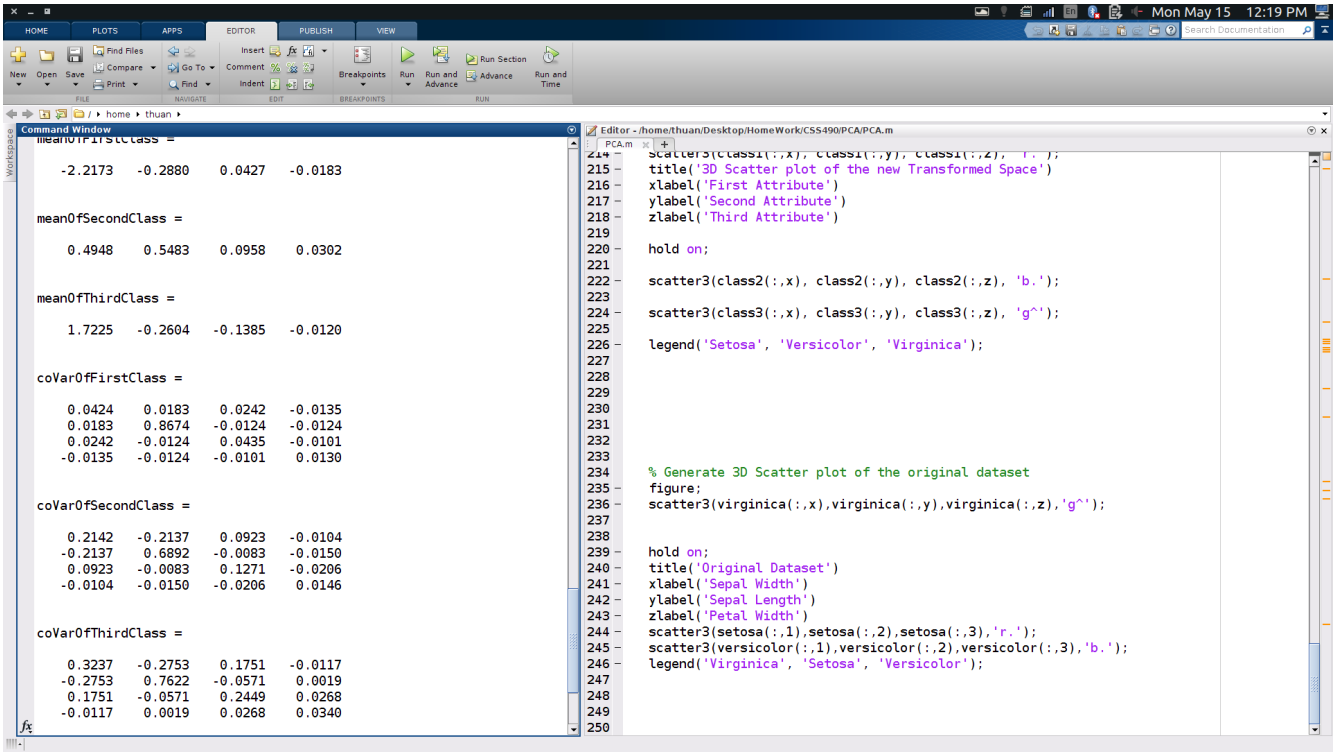

Illustration 1: Mean and covariance of new class in new space

Analysis of Scree plots

Scree plots:

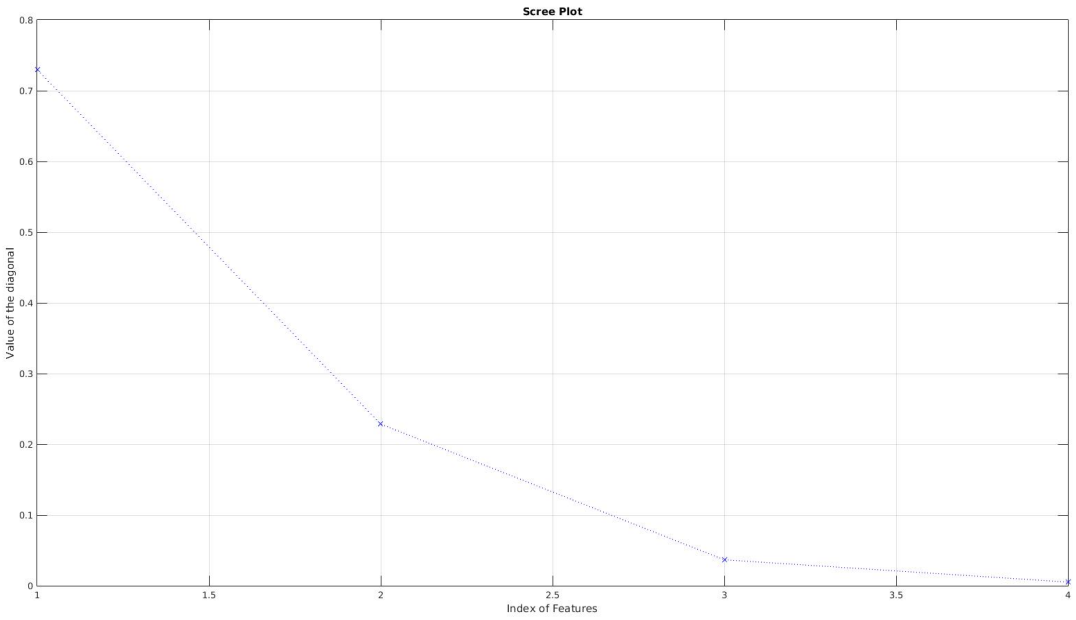

Illustration 2: Scree Plot

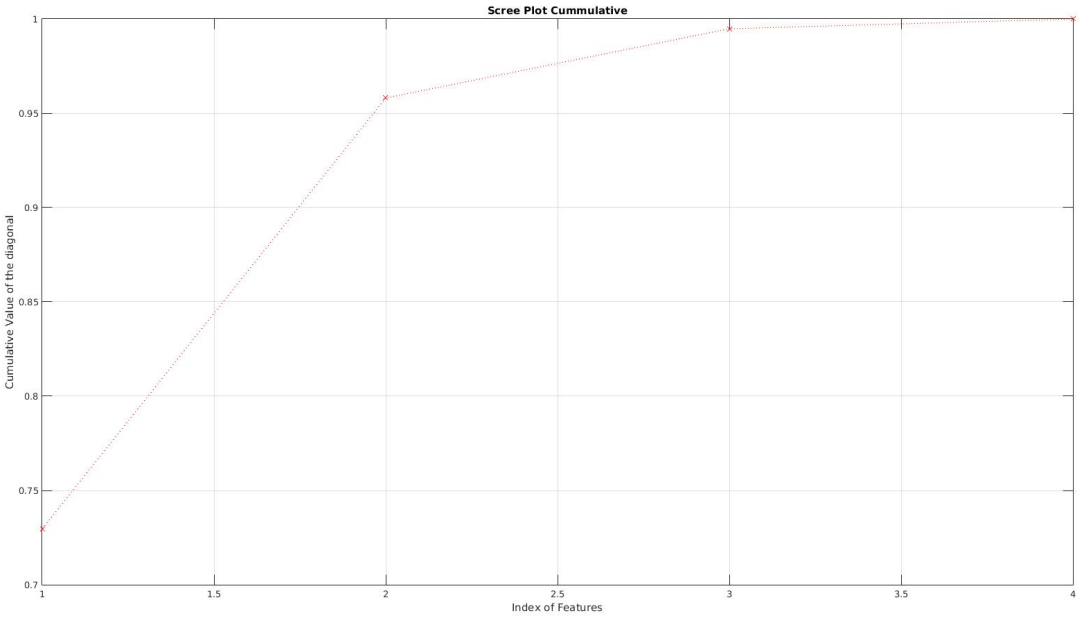

Illustration 4: Cumulative Scree Plot

The meaning of using Scree plots is that it help us to understand what features are necessary. The high elements are the one that contribute mostly to the new transformed space where the low elements are the one that do not contribute much.

From that, we can filter out the low element and only use the one that is high (relevant features). Thus reducing the dimensional

Taking at Illustraion 2 (Scree Plot), I can determine that we only need the first 3 features since they have captured most of the information. The 4^{th} feature only capture tiny bit of informaion (less than 0.05)

In fact, we might not need all 3, we only need the first 2 since they also capture a lot of information as well. The first one is around 0.72 and the second one is about 0.22 and the third one is about 0.04. Only the first one and the second one already capture 0.94 (max is 1)

Taking a look at Illustration 3(Cumulative Scree Plot), we can also say the same thing like Illustration 2 where we might only need the first 2 features. Since they capture over 0.95. We could also add feature number 3 to bring the total to 0.99 as well

Analysis of Loading Vector

The purpose of the loading vectors are to determine which features in the old space contribute to the changing new axis in the new space. Looking at their contribution (height), we can determine which one is correlated to each other. Hence we can eliminate those columns and pick 1 to reduce our dimension.

Each rows will be composed of the different effects the features in the old space to transformed the features. So for each axis in the new space, we can determine the effects of old attributes have on those new axis

Here is what the index correspond to in the old place

Index 1 is the Petal Width

Index 2 is the Petal Length

Index 3 is the Sepal Width

Index 4 is the Sepal Length

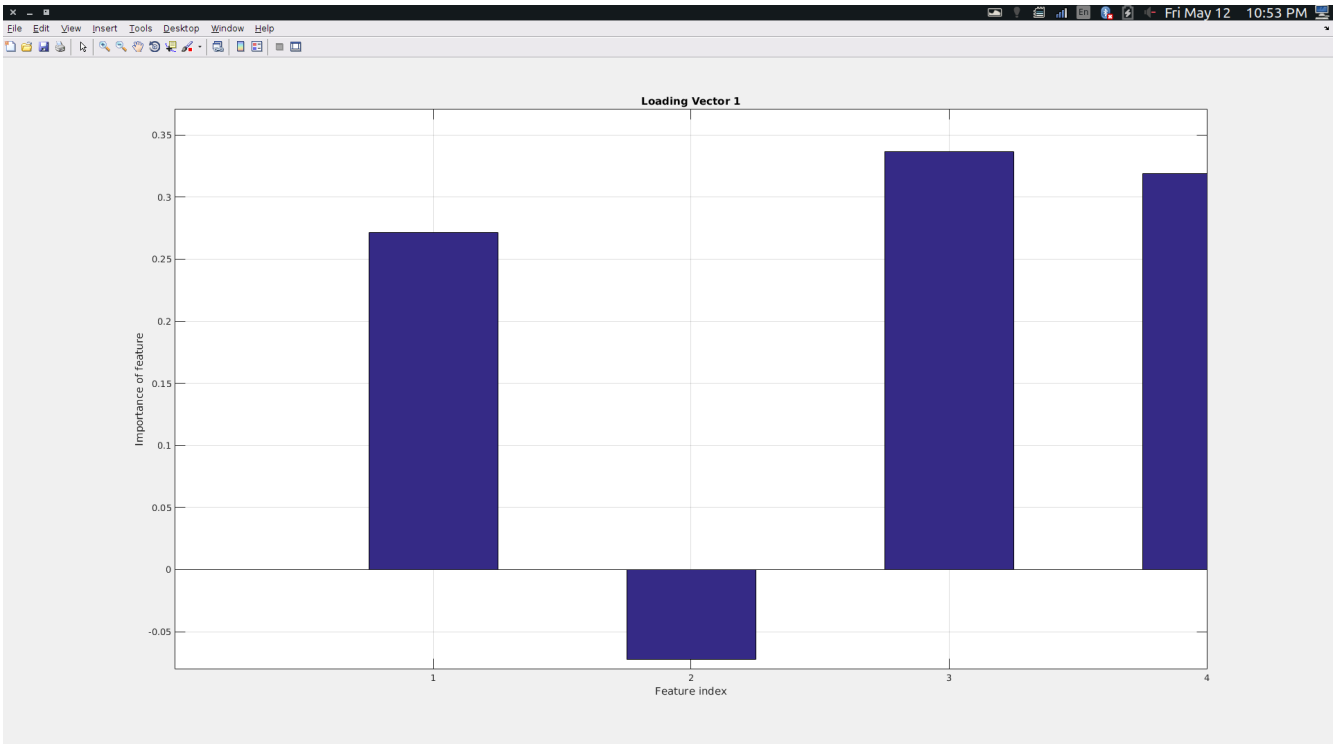

Illustration 5: First Row of V Transpose

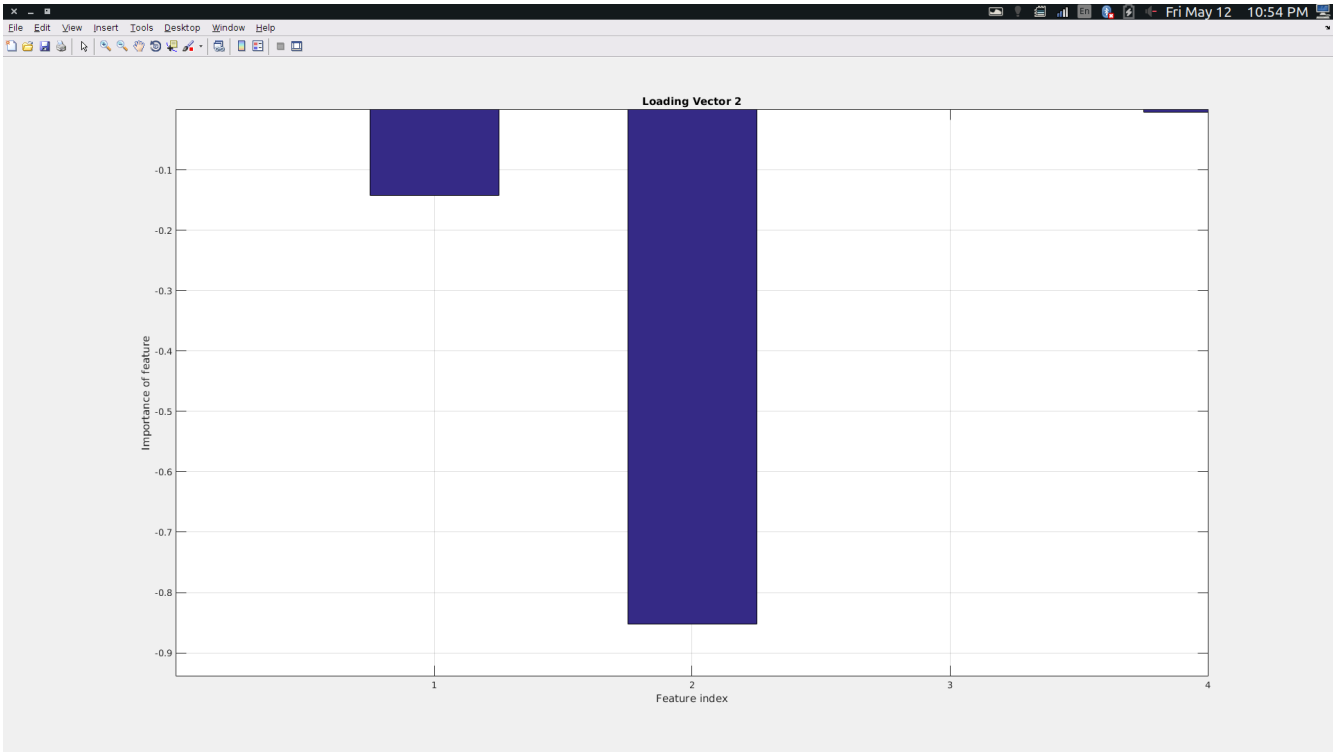

Illustration 6: Loading Vector 2

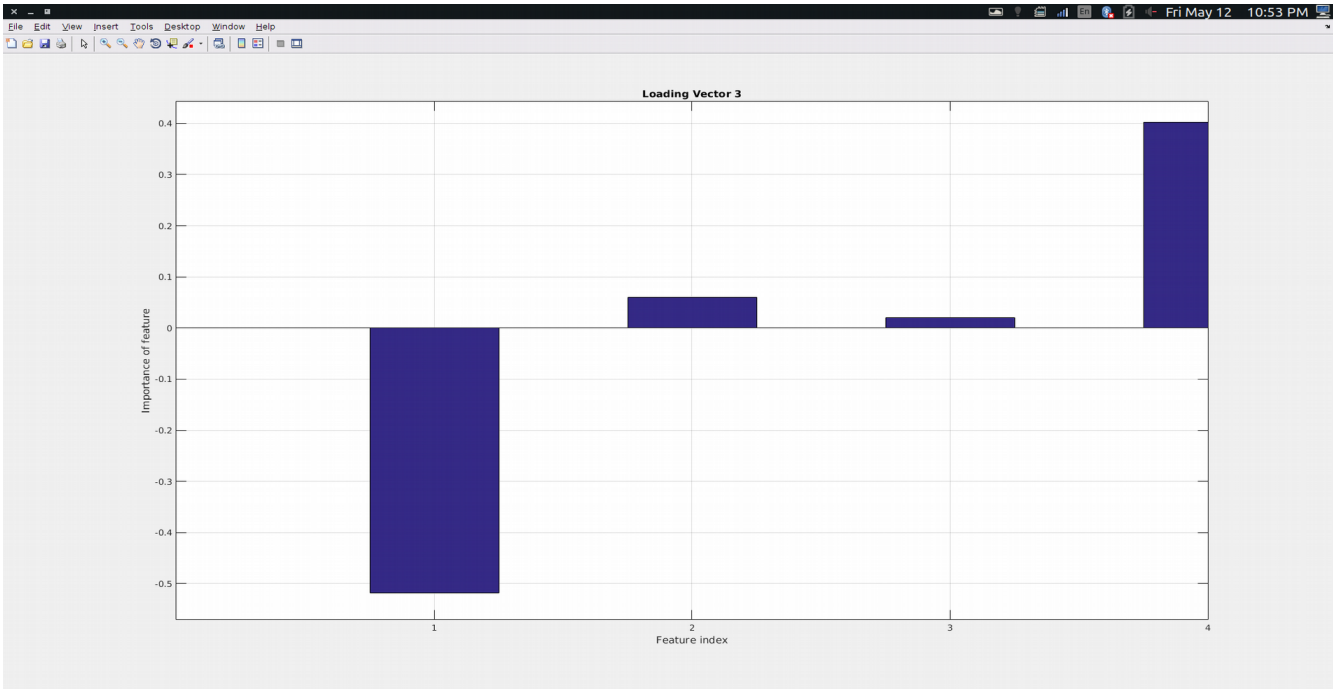

Illustration 7: Loading Vector 3

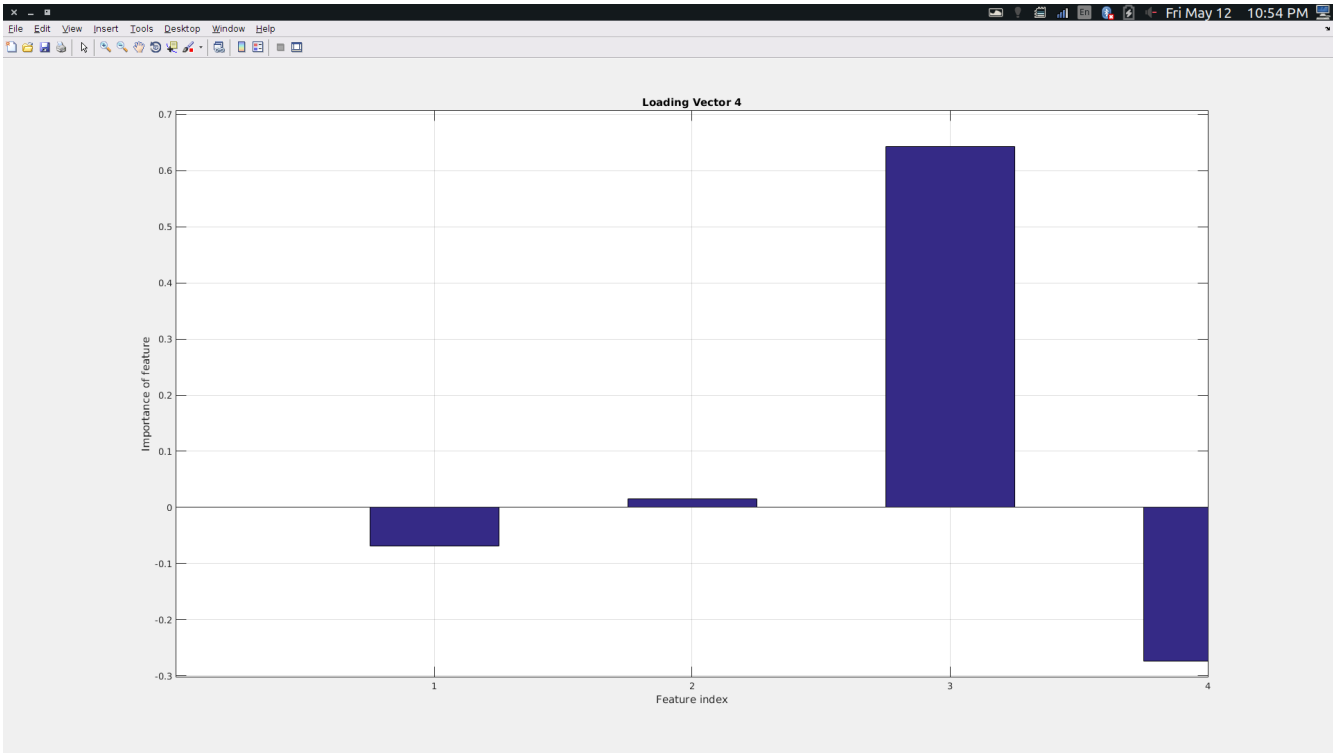

Illustration 8: Loading Vector 4

Looking at all the loading vector, all of them contribute to the new principal components in the new space. Taking a look at Loading Vector 1, at first I thought the feature 3 and feature 4 are correlated. However, taking a look at other loading vectors, we can see that feature 3 and feature 4 actually contribute differently the the new space.
Index 2 does not seem contribute much to the loading vector 1,3 and 4. But it has a huge effect on the Loading Vector 2. So we can't determine that feature at index 2 is useless and eliminate it
The same can be said about feature at index 1 where it does not seem relevant at loading vector 4 and 2. But are important in loading vector 1 and 3

All the features seem important to the new space. Note that from the scree plots, only the first 3 loading vectors are important. In which the way they captured the information in the new class

In conclusion, indexes where the bar chart is high have a high importance in the specific principal component. However, we also need to take into account of where the bars have the same height and direction. This could indicate that those features are irrelevant since they contribute the same information. In that case, we only need 1 of them. Note that we will also need to look at other loading vectors to see if there is a trend between those indexes. If they are, we can assume the above where the features are redundant.
The importance of this is that from the loading vectors, we can also identify features in the old space that is not useful in the new space or if the features in the old space are correlated. Or which one contribute the most to the new space.

2D Plots of new space and old space

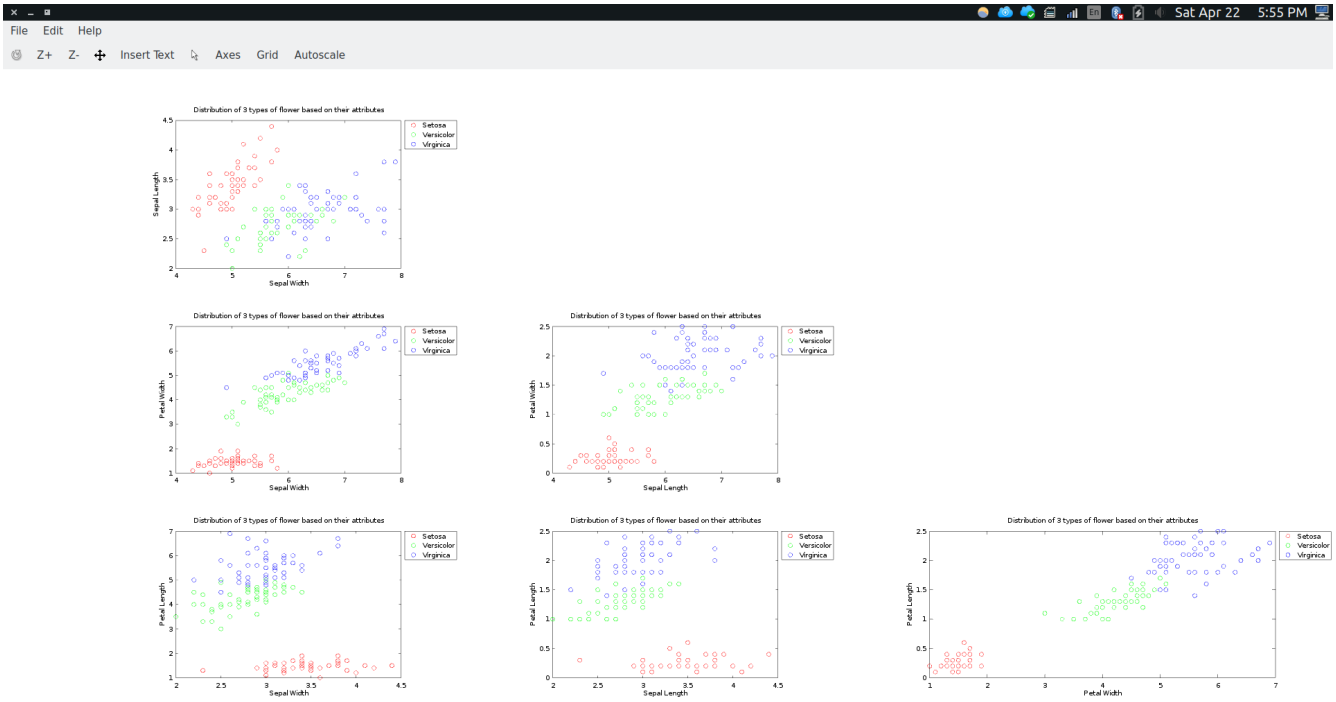

Illustration 9: Scatter Plot of the original data set

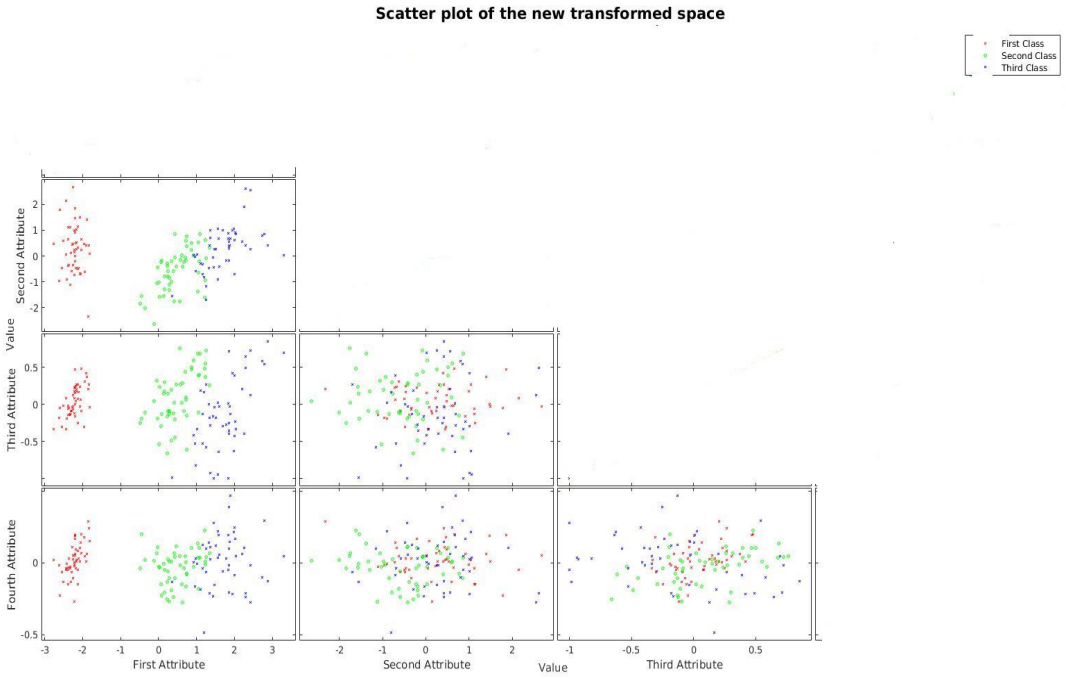

Illustration 10: Scatter Plot of New Space with Ur

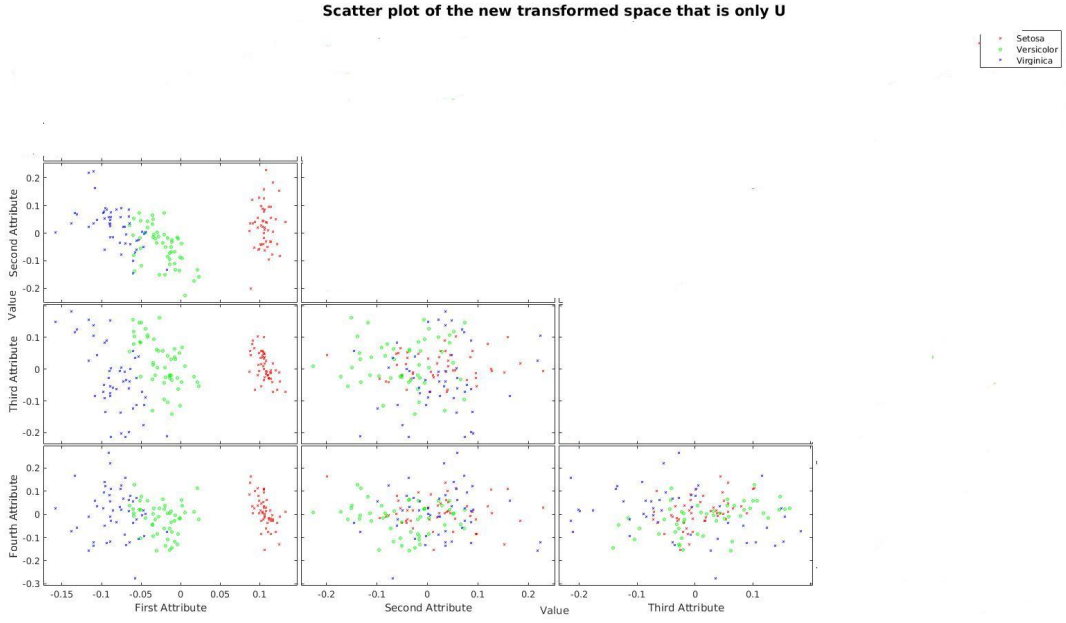

Illustration 11: Scatter Plot of the new space with only U

Looking at these scatter plots, I do not see much difference from the old space and the new space.

It actually seems that the old space is easier to see than the new space. However, The new space gives a visual representation that there seems to be more observations than the old space where the data is overlapped with each other. This suggest that the new space has a better variability between the data.

In addition, the scale of U and Ur are different. Ur is closely resemble to the original data set. Where U has a smaller interval, different range. That is because Ur is the scaled version where it takes U and multiply S

One of the reason that I can think of why is there no clear distinction between the flowers is because they do not follow a linear assumption. The relationship might be quadratic or cubic but we assumed them to be linear

Another reason might be because we do not have enough of observations or we need to find more features that better represent the flower to clearly distinct them

3D Plots

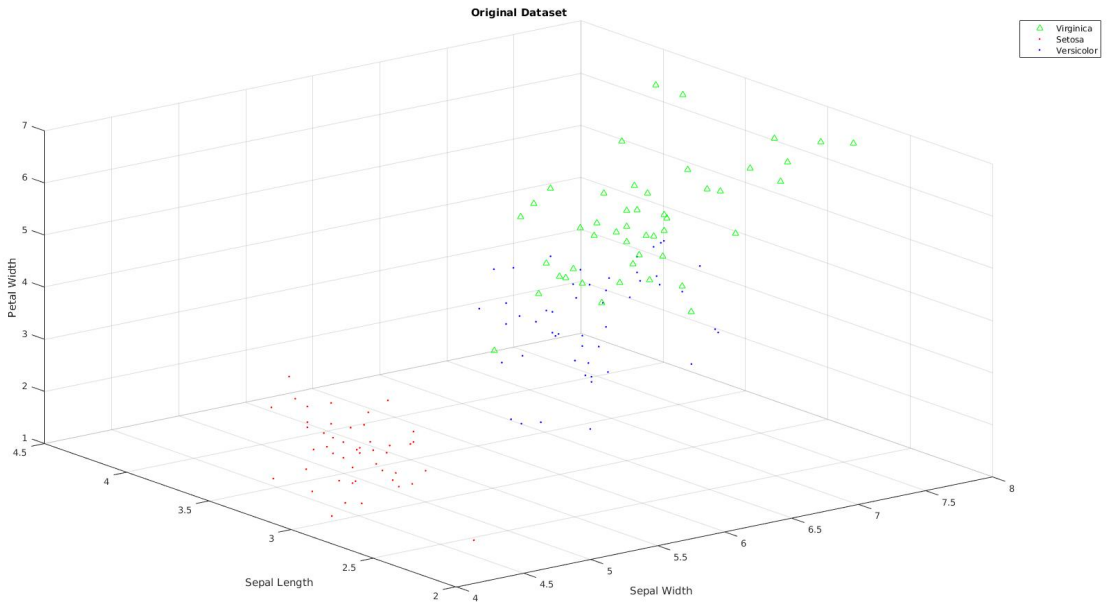

Illustration 12: 3D Scatter Plot of the original data set

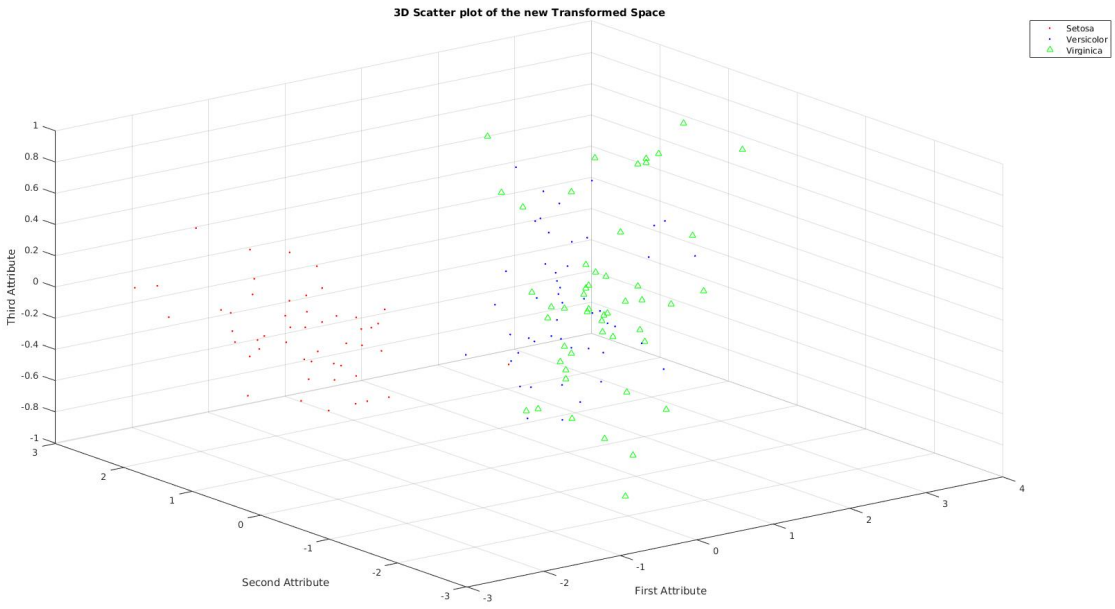

Illustration 13: 3D Scatter Plot of the transformed Data set

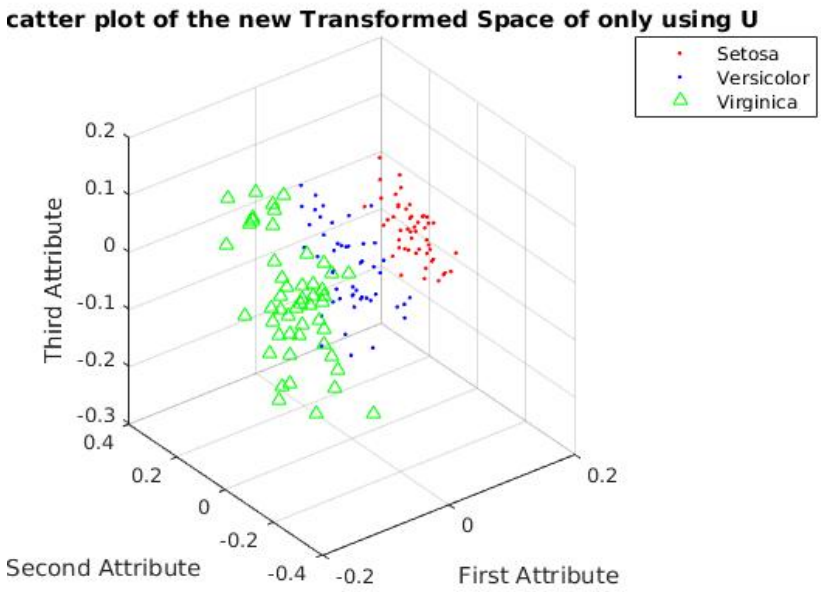

Illustration 14: 3D Scatter Plot of using U

For the 3D plots, I decided to use only the 1st,2nd and 3rd index since they capture the most information about the new space

The analysis is similar to the 2D plots where I do not see much difference in the transformed data set versus the transformed space. However, since this is 3D image, I used Matlab to rotate them in a 3 dimension up until an angle where I see that there is a clear distinction between the flowers. Overall, I see that the newly transformed space after rotating have a better separation than the original data set.

The only difference I see between using U and Ur is the scale in which Ur is the scaled version of U in which the axis have different range.

To sum up, to have better classification, I think we will need more observations and collect more features that contribute more to the classification of the flowers.