PCA Homework Report

Mohamad Mostafa

291385

1. Introduction:

In this homework, we consider a dataset characterizing galaxy observations, and used to estimate the corresponding redshift. This dataset is given by approximately three thousands of records described by 65 attributes, and it is based on the on the public catalog of Wolf et al. (2004).

2. Extraction of the Working Dataset:

First step was extracting the dataset and saving it in a dataframe, this was done using 'pandas' library, then 2500 samples were extracted randomly from the dataset for training and the remaining was left for testing which was done using the 'numpy' library and then each of the training and test samples were saved separately into different csv files.

3. PCA:

Before starting with the PCA we first notice that few of the columns in the dataset contains missing values and we have a column called 'Nr' which is just an index for the sample which was dropped since it holds no useful meaning in the dataset, so a data cleaner function was defined to clean the training and testing dataframes while replacing the missing values with its corresponding column's mean value, then we separate the variables from the labels. Moreover, we notice that the data needs to be standardized since we have different variations and means of data for each column so using 'scikit-learn' library a standard scaler was used to standardize the training and test data. Then we apply PCA where 8 components were chosen out of the 60 total variables with a total explained variance ratio of 0.84 as we see from Figure 2 and the explained variance of each principal component in Figure 1, and the number of components was chosen by trial to provide the most accurate prediction in the next task. Also, it is important to notice that the first component has almost 42% explained variance, which means that if we look at the entries of the vector, we will see how each of our original

dimensions are represented with magnitude of the principal components vectors' values as seen in Figure 3.

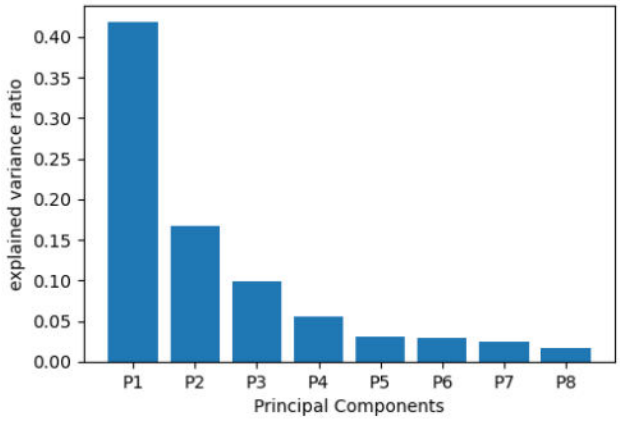


Figure 1

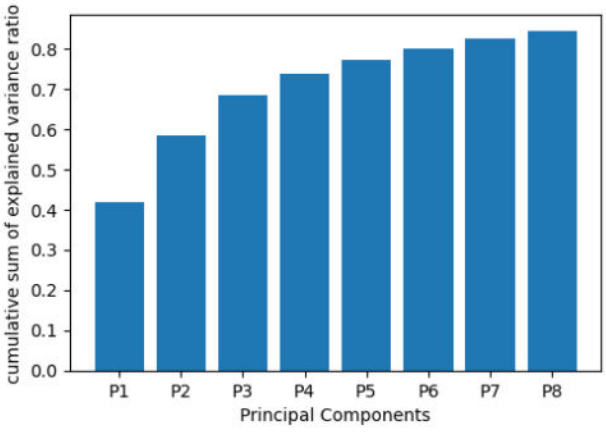
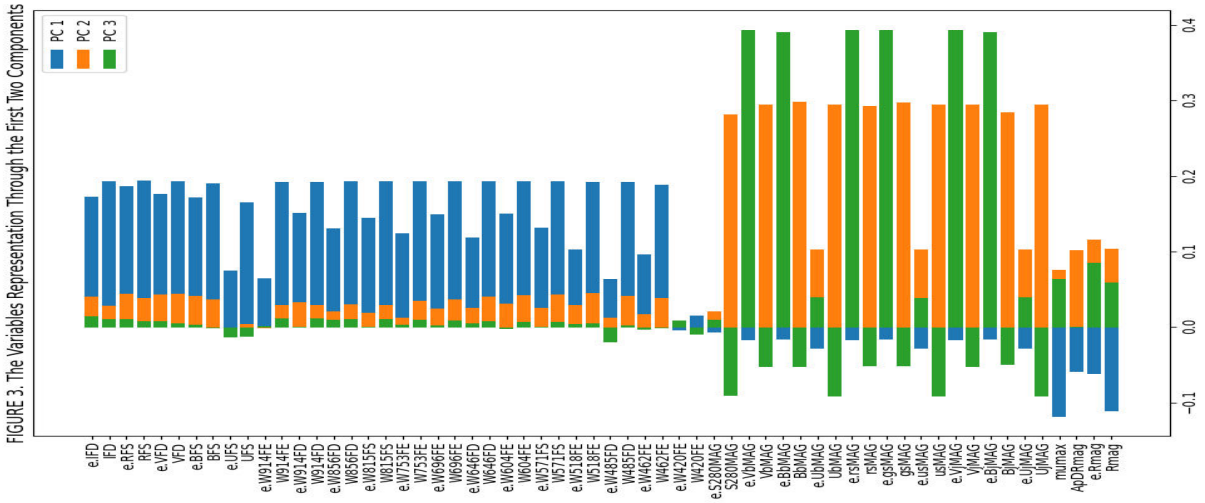


Figure 2

4. PCA Graphical Representation:

As we observe in Figure 3, we see that our first principal component has high positive values for the variables regarding the brightness of different band and low negative values in the other variables which we can interpret that PC1 represent the brightness of the galaxy observation while for PC2 we can conclude that it represents the magnitude of the bands of the galaxy observation and PC3 has almost the same behavior of PC2 except it does addition for the error variable of magnitude band and subtraction for the magnitude band regarding a positive vector.



In Figure 4,5 and 5 we observe the plot of our observations in the basis of our principal components so in Figure 4 for example we see that our observation is not quite clustered well as in Figure 5 with 2 PC where we can see better how the data are divided in clearer way and even a little better in Figure 6 with 3 PC where some of the data plotted in green are better separated across PC 3 from the blue ones.

Figure 4. Score Plot Using PC 1

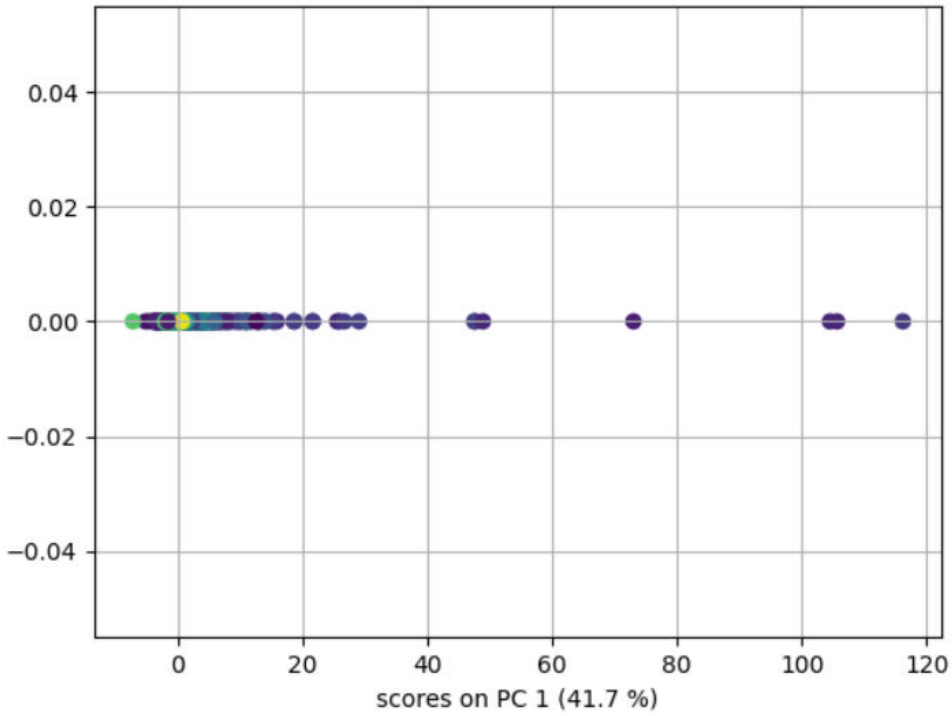


Figure 5. Score Plot Using PC 1 and PC 2

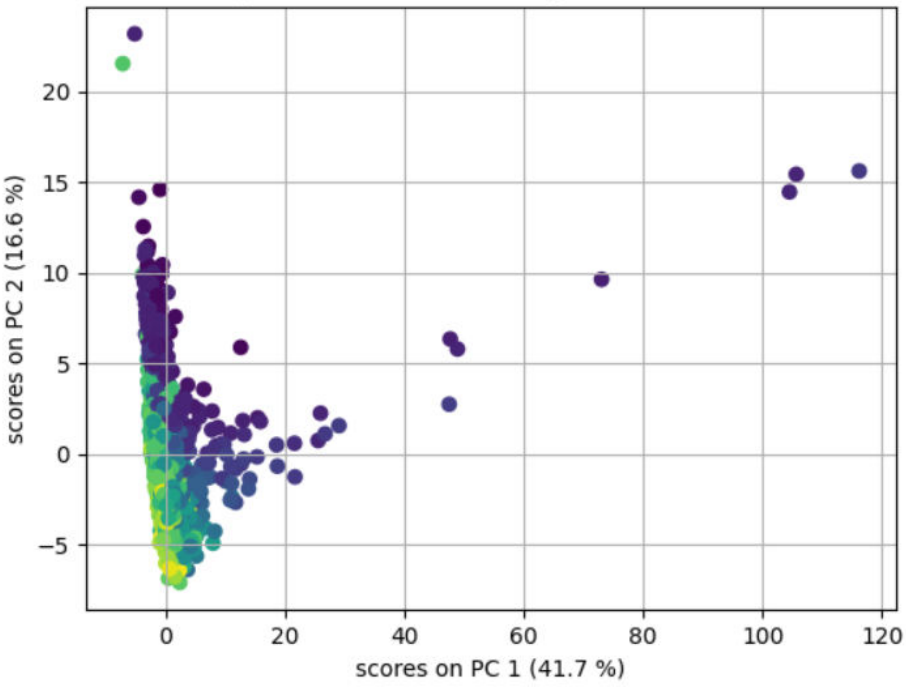
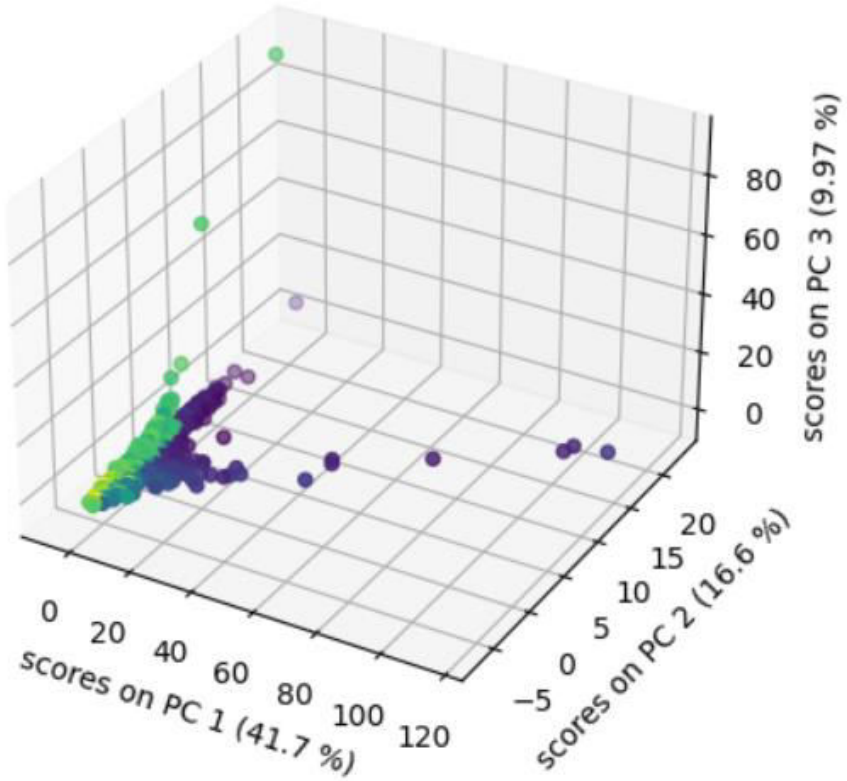


Figure 6. Score Plot Using PC 1,2 and 3



So, we see how the first 3 components have a good presentation for the brightness and magnitude for different bands which explains the galaxy observations we have.

5. PCA and k-NN:

To predict the values of the Mcz for the evaluation dataset we used the k nearest neighbor regressor on the dataset in the basis of principal components and the metrics used to evaluation the error of the prediction is the mean absolute error and the mean relative error.

Mean Absolute Error (MAE):

$$\frac{1}{N}\sum_{i=1}^{N}|\widehat{Mcz_i} - Mcz_i|\,.$$

Mean Relative Error (MRE):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{Mcz_i} - Mcz_i|}{|Mcz_i|}\,.$$

Using the kNN algorithm on the evaluation dataset after scaling and in PC basis we get value of MAE = 0.045 and MRE = 0.078, this was achieved by using the best n umber of components which is 8 with lowest error predicting the redshift.