Homework:
Principal Component Analysis
Computational Linear Algebra for Large Scale Problems
Politecnico di Torino
A.Y. 2020/2021

1st Scarano Nicola
s287908

2nd De Cristofaro Carmine
s291129

***Abstract*—In this homework, we apply to a dataset of galaxy characteristics the PCA, giving interpretation to the principal components(PCs) and the other results; moving first steps from the preparation and the analysis of the dataset, going forward to a visualization part related to the influence of the features on the main PCs. Eventually, we solve a regression task to evaluate the advantages of the Principal component analysis.**

I. Introduction

The considered dataset, *COMBO17*, is characterized by galaxy observation and it is used to estimate the corresponding redshift. In physics, the redshift is an increase in the wavelength, and corresponding decrease in the frequency and photon energy, of electromagnetic radiation due to the fact that galaxies convert interstellar matter into stars with different speeds and so the brightness and color of galaxies change with cosmic time.
Each row of the dataset represents a galaxy observation that is characterized by the following attributes:

- **Nr**: Id number of the object observed;
- **ApDRmag, mumax**: the first is the difference between the total and aperture magnitude in the R band(*red*), while the second is the central surface brightness of the object in the R band. The difference between this two values is an indicator of the galaxy size.
- **Mcz, e.Mcz, MCzml, chi2red**: Mcz and MCzml are two redshift estimates, mean and peak of the distribution, while e.Mcz is Mcz estimated error.
- **UjMAG, e.UjMAG, ..., S280MAG, e.280MAG**: these columns give the absolute magnitudes of the galaxy in 10 bands with their measuraments errors. They are based on the measured magnitudes and represent the luminosities of the galaxies.
- **W420FE, e.W420FE, ..., W914FE, e.W914FE**: these columns observe the brighteness in the 13 bands and each measure is accompanied by an error.
- **UFS, e.UFS, ..., IFD, e.IFD**: observed brightness in the five tradional broad spectral bands, UBVRI (*ultraviolet, blue, violet, red, infrared*).

II. Extraction of the Working Dataset

In the first section is performed the dataset preparation. At first we read the dataset from the given csv file: the original samples are 3642 and they are described by 65 features summarised in the introduction. We set a random seed using the first student number and we generate a random subset of 2500 samples that we are going to use as our working dataset. Then we removed the rows containing null values using *dropna()* method. We also drop the features linked to the redshift and we collect the Mcz features in another list that will be the target of our analysis.

III. Principal Component Analysis

A. Definition

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. Consider an *n x p* data matrix **X**, where each of th *n* rows represent a different repetition of the experiment, and each of the p columns gives a particular kind of feature. The transformation is defined by a set of size *l* of p-dimensional vectors of weights or coefficients

$$w_{(k)} = (w_1, ..., w_p)_{(k)} \tag{1}$$

that map each row vector $x_{\ i}$ o X to a new vector of principal component scores

$$t_{(i)} = (t_1, ..., t_l)_{(i)}, \tag{2}$$

given by

$$t_{(k)} = x_{(i)} \cdot w_{(k)} \tag{3}$$

for i = 1, ..., n and k = 1, ..., l
in such a way that the individual variables t_1, ..., t_i of t considered over the data set successively inherit the maximum possible variance from X, with each coefficient vector w constrained to be a unit vector. l is usually selected to be less than p to reduce dimensionality. Indeed, the main aims of PCA are:

- **reduce** the number of features, **preserving** most of the dataset's information;
- observe **correlations** between starting features, gathering new information.

B. Scaling

Before starting the PCA process, we study the feature distribution. The raw data show there are some Nr values that fluctuate in a range extremely bigger than other, and they could have a wrong impact on the PCA. In order to reduce the impact on the latter, we decide to normalize the data through the z-score normalization. This technique allows us to standardize the features removing the mean and scaling to unit variance. The standard score of a sample x is computed as:

$$z = (x - u)/s \tag{4}$$

where u is the mean of the training samples, and s is the standard deviation of the training samples. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using the *transform* method. The following box plots show the feature distribution before and after the normalization.

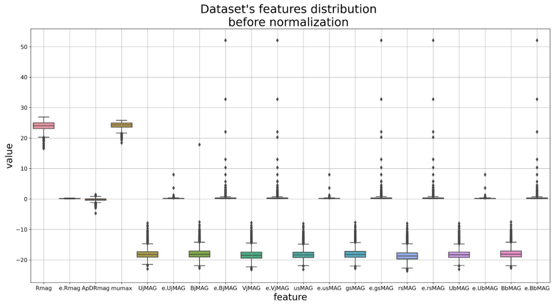

Fig. 1. feature distribution before normalization (first 20 features)

C. Choice of number of component

The application of the PCA class needs an hyperparameter set-up. The main parameter that must be set is the *n_components* alias *m* parameter. We choose the default Singular Value Decomposition solver that allows us to set the *n_components* as a float number in the *(0,1)* range. In this way we select the number of component such that the amount of variance that needs to be explained is grater than the percentage specified by *n_components*. We notice that to represent the 90% of the variance we need more than ten components but we also find out the first four components are by far the most important. Printing the cumulative and the explained variance ratio we give credibility to our choices.

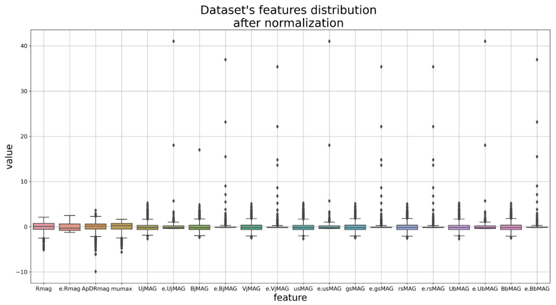

Fig. 2. feature distribution after normalization (first 20 features)

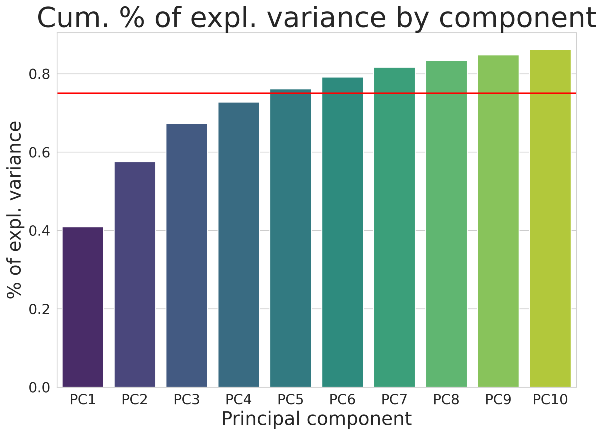

Fig. 3. cumulative explained variance by component

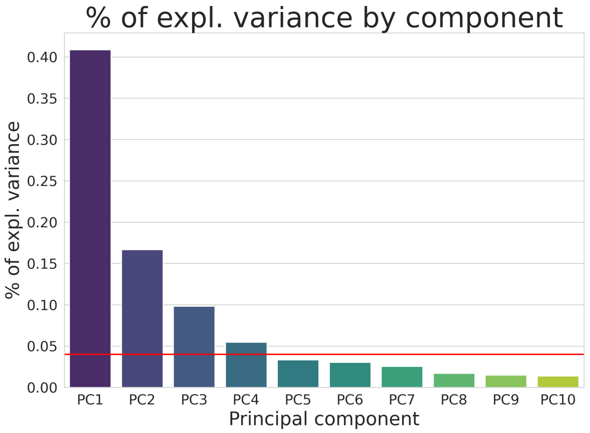

Fig. 4. explained variance by component

D. Interpretation of the first four principal component

In this section we show how the four main principal components are distributed, and we try to give them an interpretation.

Thanks to the just showed bar plots, we can see which features have the strongest impact in defining the main principal components. Since the initial data is highly dimensional, also the main principal components are heavily influenced by a relatively large number of features. For the sake of simplicity, we do not analyse all the features but just the ones with the highest value for each component. The features that affect more each component ensure a feasible interpretation of them:

- the first component is strongly correlated to information about observed brightness of a galaxy in the five traditional broad spectral bands how features 56, 54 and 32 (IFD, RFS, W571FS) suggest more. Moreover, the features 0, 1, 2, 3 (Rmag, e.Rmag, ApDRmag, mumax), that are related to the aperture magnitude and the central surface brightness of the Red band, are subctrated from the component how we can deduce from their negative values in the fig.5;
- the second component is focused on the measured magnitudes of the galaxies in 10 bands and represent their luminosity. The features 18, 12 and 8(BbMAG, gsMAG, VjMAG) are the ones with the higher positive impact on the component and only features 24(W420FE) has negative value, since is related to the brighteness of a band not taken in account for this component;
- the third component, instead, is linked to the error relative to the brightness of the galaxy in each of the traditional spectral bands because features 13, 21 and 9(e.gsMAG, e.VbMAG, e.VjMAG) are the ones with the higher positive value and they all are error measurements;
- eventually, the fourth component is affected by the error on the measured magnitudes(e.UjMAG, e.UbMAG, e.usMAG). This component is the one with the higher number of features with negative values because they are all the features added to the preceding components.

E. Graphical Representation

In the previous subsection we note that quite all the original features are represented in that four PCs. This is an expected result because the main four components represent in percentage large part of the variance. To better understand the disposition of the data in the PC space we build the following scatter plots.

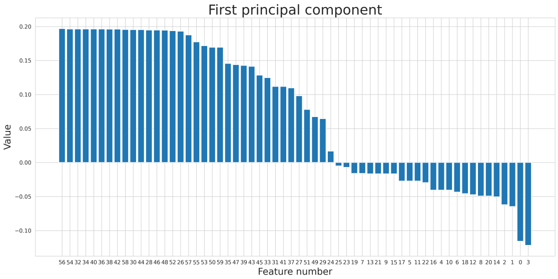

Fig. 5. first principal component with relatives features

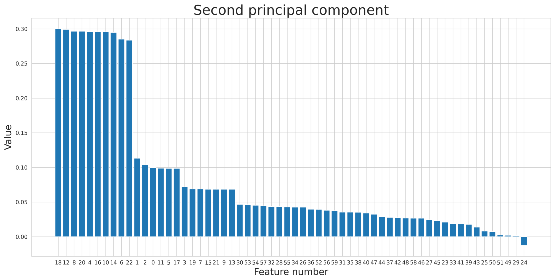

Fig. 6. second principal component with relative features

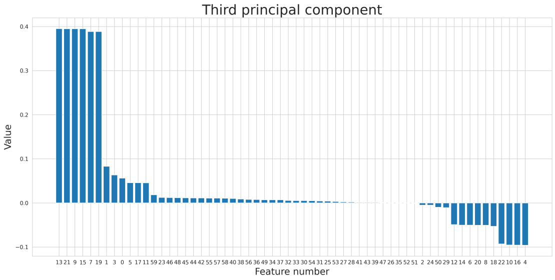

Fig. 7. third principal component with relative features

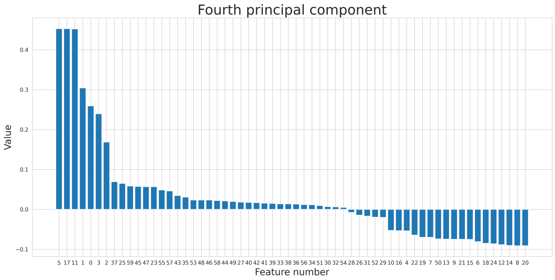

Fig. 8. fourth principal component with realive features

The distribution of the Galaxies in the PC1-PC2 space, fig.9, highlights how the two components are able to divide,with good results, the points with respect to Mzc values, the target of the dataset. Moreover, it is clear that points with similar target values are quite clustered in the space while outliers are well separeted.

In the three dimension representation, fig.10, we can deduce similar conclusions. Indeed, points with similar Mcz values are more grouped and less overlapped; so the increasing of the dimensions in this case make the data more distinguishable.

IV. PCA AND K-NN

To test the efficiency of the approach describe until now, this data are used to perform a Regression task using the KNN. The target that must be predicted is the MCz value. The data used as training set are the ones co-trained in the COMBO17pca_287908.csv file, while

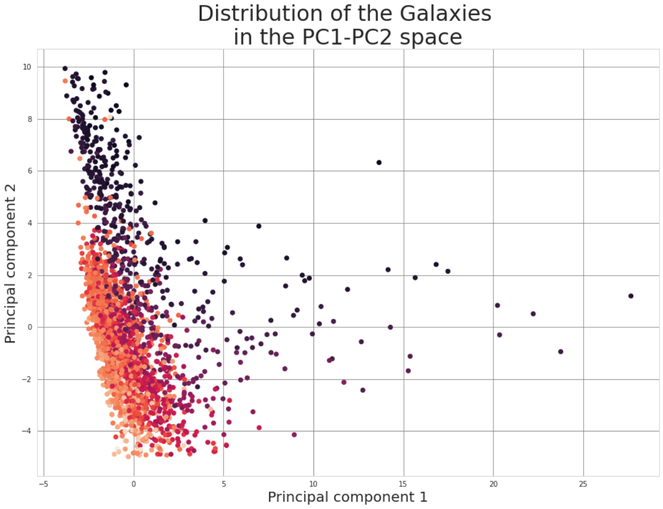

Fig. 9. distribution of the galaxies in two-dimensional space

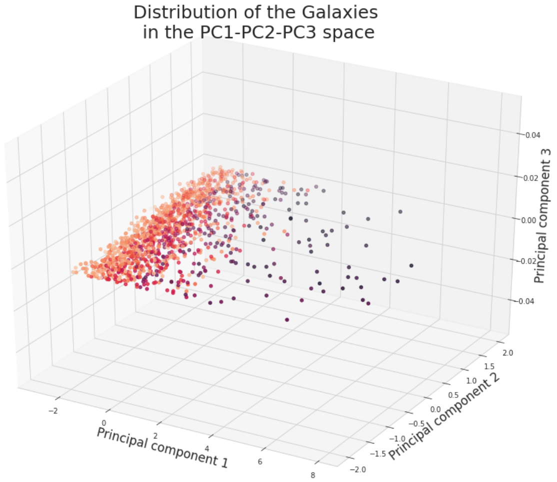

Fig. 10. distribution of the galaxies in three-dimensional space

COMBO17eval_287908.csv is used as test set. Both training and test sets are normalized using the z_score normalization. Afterwards, we run *KNeighborsRegressor* applying PCA with different number of components to get the minimum errors. We achieve this results with eight components that ensure a mean absolute error equal to *0.013* and a mean relative error equal to *0.021*.

The obtained results could refute the assumptions and the procedure followed so far, since we get the minimum errors with eight components instead of four. However, analysing the graph, Fig.11, we can appreciate how the first four components have the strongest impact on the reduction of the errors, looking to the slopes. Therefore, this conclusion well supports what has been shown so far.

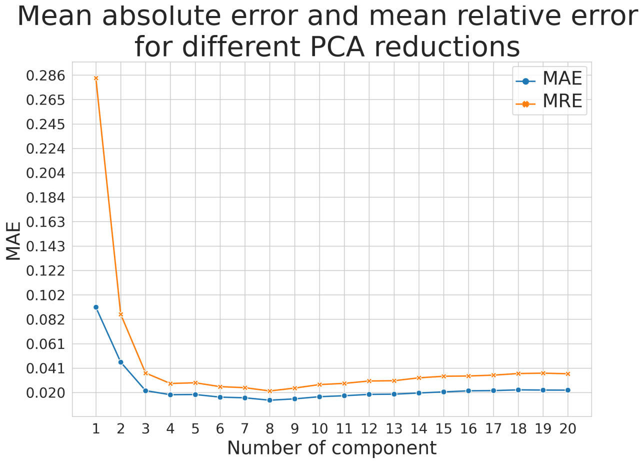

Fig. 11. mean absolute error and mean relative error related to the number of components