POLITECNICO DI TORINO

Facoltà di Ingegneria

Corso di Laurea Magistrale in Ingegneria Matematica

Computational linear algebra for large scale problems

HOMEWORK
Principal Component Analysis

Students:
Marco Bottino (s274110)
Maria Margherita Lovera (s278425)

Academic year
2020/2021

Contents

Chapter 1

Introduction

In this homework we consider a dataset characterizing galaxy observations, also used in order to estimate the corresponding redshift. This dataset is given by approximately three thousands of records and each row represents a galaxy observation charaterized and described by 65 attributes.

In the first part of the homework we perform data extraction, data cleaning (if necessary) and we split the dataset into two parts: a development set to train the algorithms and an evaluation set to test them. In the second part we apply the PCA algorithms to the training dataset in order to compute the principal components (PCs) and to give an interpretation of the first PCs. In the third part, a graphical analysis of the PCs is done, taking into account the target value Mcz. In the fourth and last part the test dataset is transformed through the PCA fitted on the training dataset and the mean redshift Mcz is estimated using the k-Nearest Neighbours method in the PCs space.

Chapter 2

Extraction of the working dataset

We first loaded the dataset *COMBO17.csv* and created a new feature called **galaxy_size**, defined as the difference between **Rmag** and **mumax**, since it seemed to be potentially significative in order to analyze the redshift: indeed, the phenomena that can cause the redshift are mainly linked to space and, so, to space dimensions (i.e. objects moving apart or closer together in space or space expansion). We also observed that there were 24 rows with NaN values, so we dropped them obtaining a dataset of shape 3438x66. Furthermore, we noticed a column containing the indeces of the rows (i.e. **Nr**) and we chose to remove it since we know it would not have been useful for our goal. Then we sampled randomly 2500 rows to create two new datasets:

- *COMBO17pca_278425.csv* (2500x66), the development set containing the extracted rows
- *COMBO17eval_278425.csv* (938x66), the evaluation set with the remaining rows

Since our analysis consists in analyzing the redshift, we removed from both the datasets the features containing informations about the redshift

- **Mcz**
- **e.Mcz**
- **MCzml**
- **chi2red**

and we stored **Mcz** as our target variable.
For each dataset we z-normalized the features by applying Python StandardScaler() from scikit-learn and compared the variance explained from the features before and after the standardization. We chose not to perform any feature selection in this step, since to do so we would have chosen arbitrarily a threshold of explained variance under which the features would have to be dropped. We preferred to postpone this kind of analysis on the principal components.

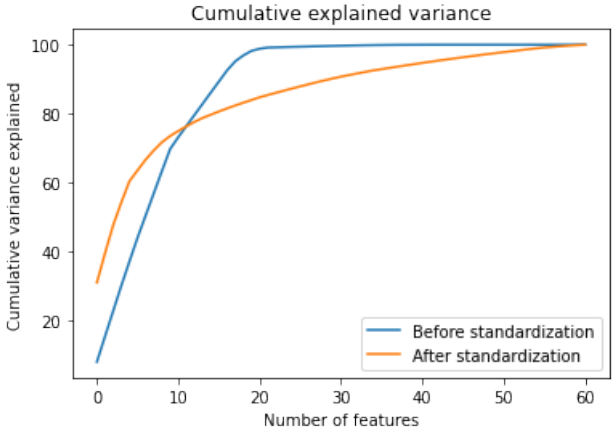


Figure 2.1: Before and after z-normalization

By looking at Figure 2.1, we observed that before the z-normalization the 20 most variable features already explain more than 98% of the variance in the dataset. After the normalization, though, the variance is more distributed along features.

Before normalization		
Feature	**Var. explained**	**Cum.var. explained**
BjMAG	7.79034	7.79034
rsMAG	7.4371	15.2274
VbMAG	7.35502	22.5825
VjMAG	7.35275	29.9352
gsMAG	7.06986	37.0051
After normalization		
Feature	**Var. explained**	**Cum.var. explained**
mumax	30.8434	30.8434
W696FE	8.56802	39.4114
gsMAG	8.12881	47.5403
Rmag	6.60996	54.1502
e.UjMAG	6.17908	60.3293

Table 2.1: 5 most variable features before and after normalization

By looking specifically at the features in Table 2.1, it is possible to see that after the z-normalization there is a feature that explains alone more than 30% of the variance: it's the central surface brightness of the object in the R band, *mumax*.
To get a better understanding of the dataset we chose to plot the correlation matrix as a heatmap:

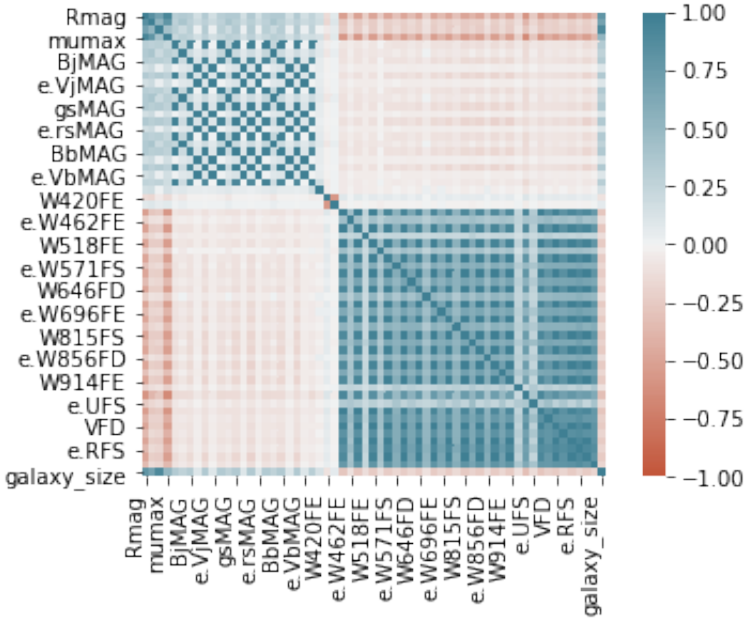


Figure 2.2: Correlation matrix

We can observe from the correlation matrix that there are three main groups of features:

- the first features, which estimate the galaxy size through the magnitude measured in the red band
- the generic measures of magnitude and their errors in the different bands (xMAG)
- the measures of brightness through photon flux density and their errors (xFE, xFD, xFS)

It seems like the measure of brightness in band 420 is an exception in this division since it's uncorrelated with the rest of the features.
We then tried to reduce this redundancy in the features by performing dimensionality reduction with PCA.

Chapter 3

Principal Component Analysis

After having preprocessed the dataset, we performed PCA to extract the most relevant features from the dataset, since it seemed to be quite redundant. We first performed PCA with a number of components equal to the number of features, in order to be able to plot the variance explained. Furthermore, even though in our analysis we consider the normalized data, we compared again the results with the ones obtained by applying PCA on the dataset without the normalization (see Figure 3.1): once again the variance is better distributed after the normalization.

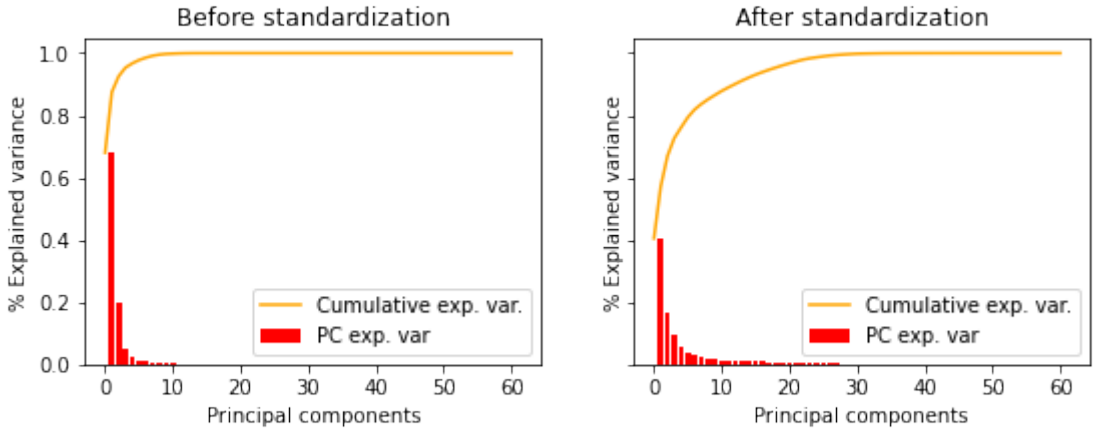


Figure 3.1: PCA performed before and after standardization

Since we are interested in the normalized data, we look at the plot on the right: it is possible to observe that with only 13 components we explain more than 90% of the variability of the dataset, against the original 61 features (see also Table 3.1). For this reason we chose to consider 13 principal components so to represent the evaluation set into the 13-dimensional PC space (see Chapter 4).

N° of components	Cum. var. explained
1	0.40496
2	0.571517
3	0.669818
4	0.726774
5	0.76093
6	0.793663
7	0.818701
8	0.836545
9	0.851718
10	0.86526
11	0.878319
12	0.88986
13	0.900862

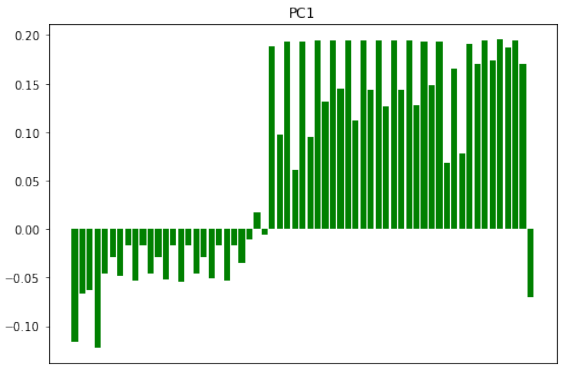
Table 3.1: First 13 principal components with respective cumulative explained variance after standardization

3.1 Principal components interpretation through loadings

By looking at the loadings, i.e. the coefficients representing the relation between the original features and the new principal components, we tried to give an interpretation to the meaning of the single components.

PC1

We can observe positive values for the metrics of brightness and negative values for the metrics of galaxy magnitude.
We therefore assumed that this component captures the general photon flux density among the different bands. We can call it *photon flux density* (PF DENS).

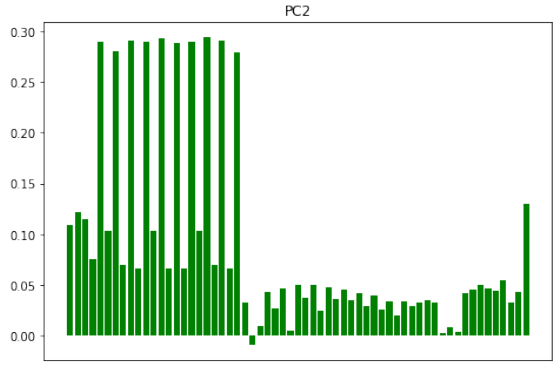


Top features for PC1		Bottom features for PC1	
Feature	**Component**	**Feature**	**Component**
RFS	0.19526	BbMAG	-0.050579
VFD	0.194994	gsMAG	-0.0517626
W753FE	0.194943	VjMAG	-0.0534122
W604FE	0.194927	VbMAG	-0.0534274
W646FD	0.194787	rsMAG	-0.0544538
W815FS	0.194624	ApDRmag	-0.0634369
W696FE	0.194539	e.Rmag	-0.0667143
W571FS	0.194501	galaxy$_size$	-0.0695302
IFD	0.194126	Rmag	-0.116371
W856FD	0.194039	mumax	-0.122592

We applied the same kind of reasoning to the following components.

PC2

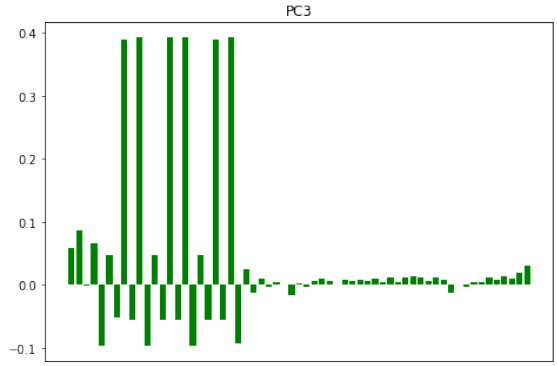
Positive values for all the measures of magnitude, in particular of the absolute magnitude in all the bands: the component measures the magnitude in general of a galaxy. We can call it *magnitude* (MAG).



Top features for PC2		Bottom features for PC2	
Feature	**Component**	**Feature**	**Component**
BbMAG	0.293424	e.W462FE	0.027149
gsMAG	0.292461	e.W753FE	0.0256388
VjMAG	0.289897	e.W571FS	0.0244721
VbMAG	0.289872	e.W815FS	0.0203215
UjMAG	0.289531	e.W420FE	0.00924275
UbMAG	0.289479	UFS	0.00885969
usMAG	0.28944	e.W485FD	0.00532819
rsMAG	0.288382	e.UFS	0.00418835
BjMAG	0.27999	e.W914FE	0.00256042
S280MAG	0.278868	W420FE	-0.00900767

PC3

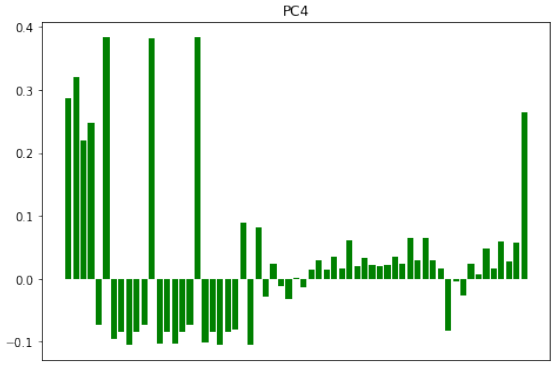
High positive values for the measurement errors of galaxy magnitude, negative values for the absolute magnitudes: the component measures the relative uncertainty of the measures of magnitude. We can call it *magnitude uncertainty* (MAG ERR).



Top features for PC3		Bottom features for PC3	
Feature	**Component**	**Feature**	**Component**
e.gsMAG	0.392051	BjMAG	-0.0527636
e.VjMAG	0.391988	gsMAG	-0.0548867
e.VbMAG	0.391988	BbMAG	-0.0554703
e.rsMAG	0.391865	rsMAG	-0.0556566
e.BjMAG	0.389814	VbMAG	-0.0561147
e.BbMAG	0.389813	VjMAG	-0.0561175
e.Rmag	0.0860953	S280MAG	-0.0941081
mumax	0.0657535	usMAG	-0.0968732
Rmag	0.0590012	UbMAG	-0.0971981
e.UbMAG	0.0468107	UjMAG	-0.0973158

PC4

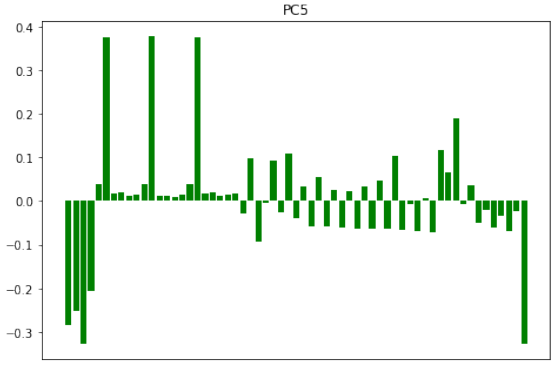
The highest positive values are the errors of measurement in the magnitude ultra-violet band error and the features related to the galaxy size: the component can be an indicator of the *size* (SIZE).



Top features for PC4		Bottom features for PC4	
Feature	**Component**	**Feature**	**Component**
e.UjMAG	0.383025	e.BbMAG	-0.0843996
e.UbMAG	0.383025	e.BjMAG	-0.0843997
e.usMAG	0.382632	e.rsMAG	-0.0844636
e.Rmag	0.319672	BjMAG	-0.0960889
Rmag	0.287376	BbMAG	-0.100669
galaxy$_size$	0.264263	gsMAG	-0.102932
mumax	0.247066	rsMAG	-0.103734
ApDRmag	0.220032	VjMAG	-0.104757
e.S280MAG	0.0899764	VbMAG	-0.104758
e.W420FE	0.0810237	W420FE	-0.105148

PC5

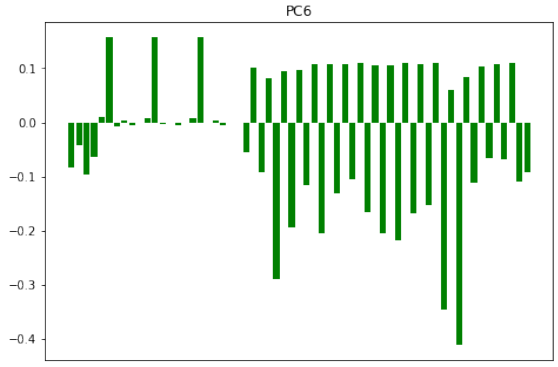
The values of the errors in the ultra-violet are still high (and so will be in the next component), whereas the loadings of the measures of galaxy on the opposite to the previous one are negative. We can call it *ultra-violet relative error*, since the bigger is the galaxy more it drops (UERR REL).



Top features for PC5		Bottom features for PC5	
Feature	**Component**	**Feature**	**Component**
e.usMAG	0.376885	W856FD	-0.0672026
e.UjMAG	0.376469	IFD	-0.0682732
e.UbMAG	0.376469	W914FD	-0.0699657
e.UFS	0.189556	W914FE	-0.0703797
e.W914FE	0.116528	e.W420FE	-0.0943184
e.W485FD	0.107741	mumax	-0.205789
e.W815FS	0.102398	e.Rmag	-0.251786
W420FE	0.0976125	Rmag	-0.285231
e.W462FE	0.0933531	ApDRmag	-0.327135
UFS	0.0649676	galaxy$_size$	-0.327849

PC6

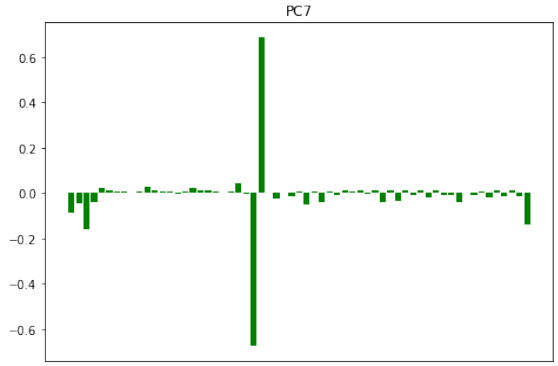
Positive values for the absolute measures of photon flux density, negative values for the errors of the same measures. On the opposite of the previous components, this one measures the *accuracy of photon flux density measure* (PFD ACC).



Top features for PC6		Bottom features for PC6	
Feature	**Component**	**Feature**	**Component**
e.UjMAG	0.156687	e.W914FD	-0.154022
e.UbMAG	0.156687	e.W696FE	-0.165787
e.usMAG	0.156684	e.W856FD	-0.168486
W696FE	0.110108	e.W485FD	-0.194765
W856FD	0.110018	e.W753FE	-0.205044
W914FE	0.109764	e.W571FS	-0.205768
IFD	0.10871	e.W815FS	-0.217624
W646FD	0.108611	e.W462FE	-0.290759
W914FD	0.108548	e.W914FE	-0.347475
W604FE	0.107257	e.UFS	-0.412191

PC7

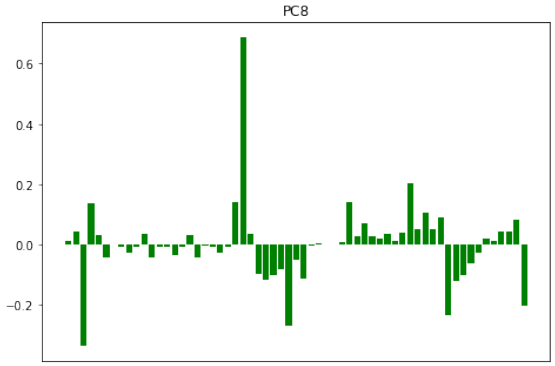
Very high values for the error of measure of W420FE and very low for its absolute value. This feature represents an exception even in the PCA representation. We can call this component *uncertainty of W420FE* (W420FE UNC).



Top features for PC7		Bottom features for PC7	
Feature	**Component**	**Feature**	**Component**
e.W420FE	0.685932	e.W571FS	-0.0386188
S280MAG	0.041168	e.W753FE	-0.039108
usMAG	0.0249242	e.UFS	-0.0397636
UbMAG	0.0247426	mumax	-0.0410094
UjMAG	0.0242955	e.Rmag	-0.045797
W914FE	0.0132948	e.W518FE	-0.0526791
W815FS	0.0132783	Rmag	-0.0885026
IFD	0.0129356	galaxy$_s ize$	-0.139705
W856FD	0.0122654	ApDRmag	-0.158372
e.usMAG	0.011928	W420FE	-0.675641

PC8

From this component on we weren't able anymore to find any significant pattern into the loadings, so from this point on we decided not to go on with loadings anaysis.



Top features for PC8		Bottom features for PC8	
Feature	**Component**	**Feature**	**Component**
e.S280MAG	0.685899	e.W420FE	-0.0968841
e.W856FD	0.204491	e.W462FE	-0.102828
e.W646FD	0.141732	BFS	-0.103204
S280MAG	0.139676	e.W518FE	-0.114053
mumax	0.138312	W462FE	-0.117973
e.W914FD	0.107428	e.UFS	-0.120185
e.W914FE	0.0899463	$\text{galaxy}_{s}ize$	-0.203952
e.IFD	0.0806566	UFS	-0.235917
e.W696FE	0.0716574	e.W485FD	-0.268118
W914FD	0.0509516	ApDRmag	-0.336429

3.2 PCA graphical representation

We represented the scores of out dataset on the first principal components extracted by using scatter-plots in one, two and three dimensions. The observations are colored according to the value of the target Mcz, allowing us to understand which PCs are more useful to separate data with different red-shift values.

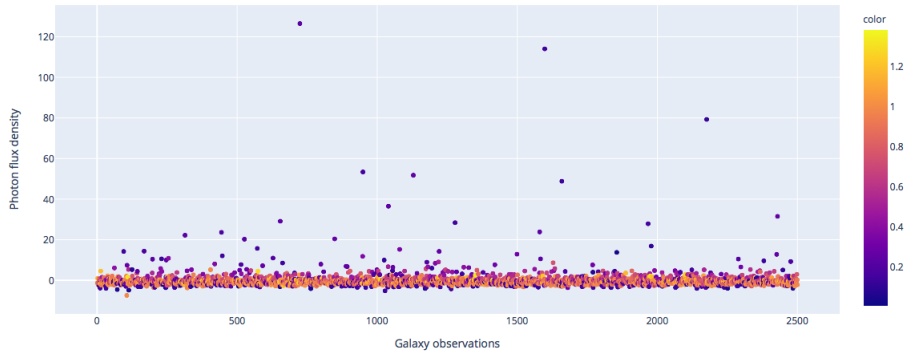


Figure 3.2: The photon flux density doesn't seem relevant in separating the data according to the target.

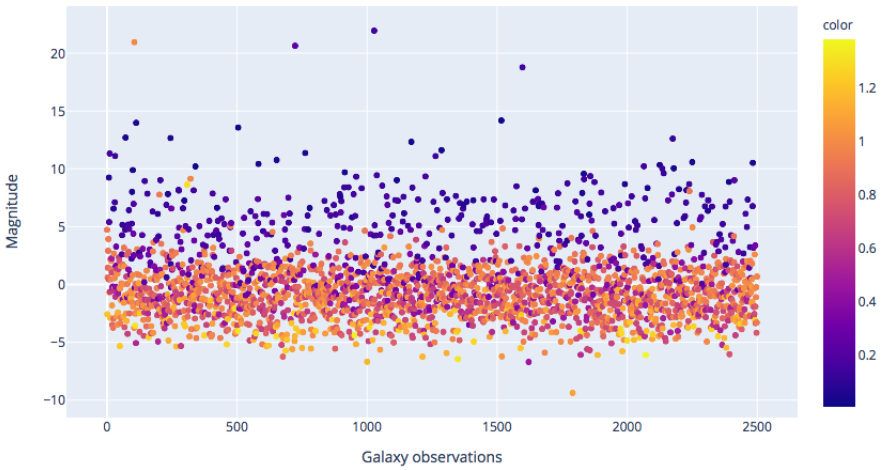


Figure 3.3: The general magnitude seems to be a good measure for this purpose: we can see that low values of magnitude correspond to high values of red-shift

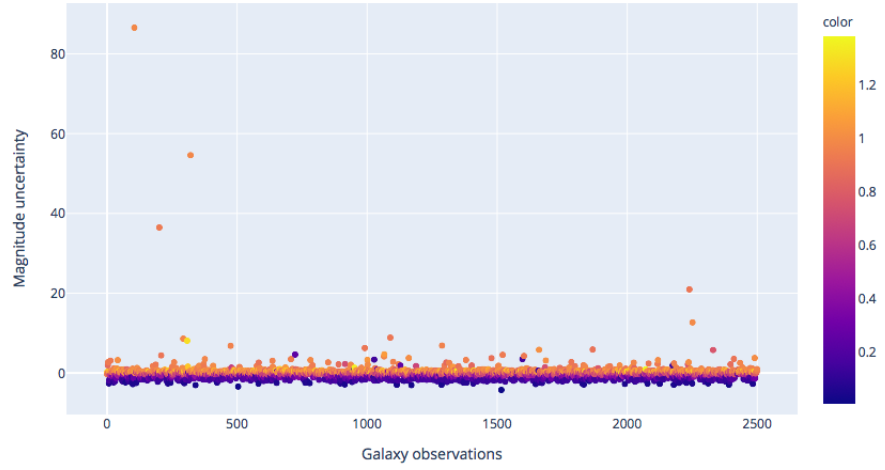


Figure 3.4: The uncertainty in the measurement of magnitude also contains information about the red-shift value: they seem to be proportional. We can also notice that a few observations have very high uncertainty, they may be outliers that would have to be monitored if one would want to perform more advanced analysis on the dataset.

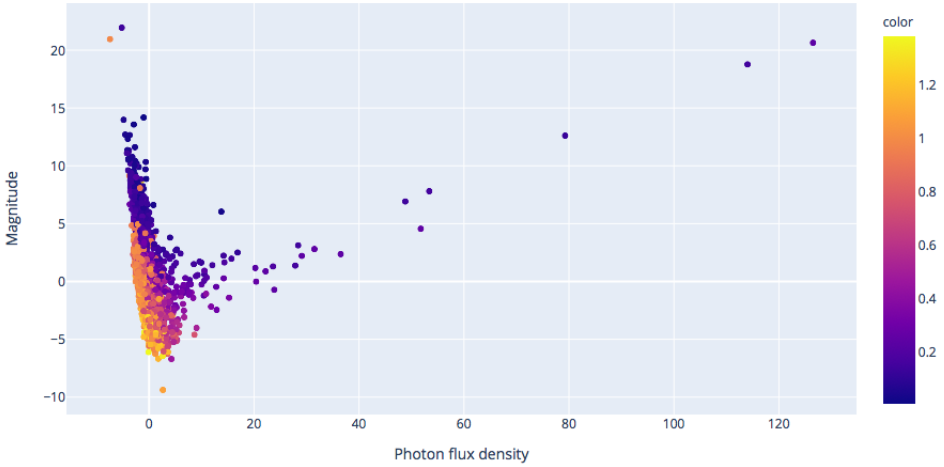


Figure 3.5: The 2D plots give us more information of how these components interact: there seems to be a negative correlation between the first two components. The highest values of red-shift can be found for low values of photon flux density and low values of magnitude.

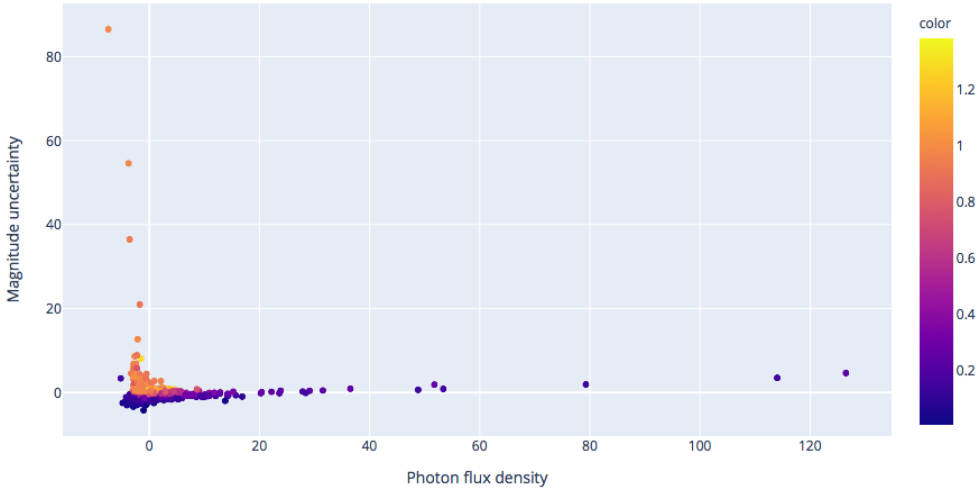


Figure 3.6: Photon flux density and magnitude uncertainty seem to be uncorrelated: it's interesting to highlight how when one has positive values, the other usually is near to zero, and vice versa.

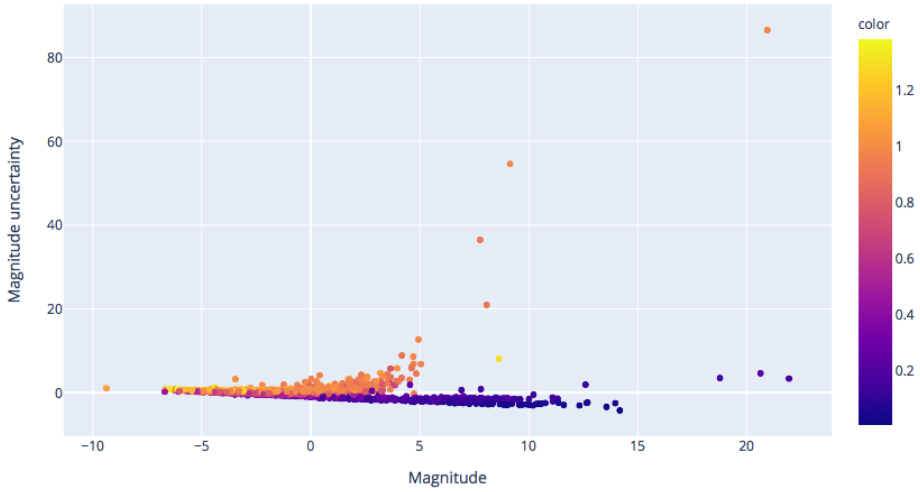


Figure 3.7: The two components considered seem to have some kind of correlation. We can find the highest values of red-shift once again for low values of magnitude, but it's interesting to notice how high values of the magnitude uncertainty seem to rise the red-shift independently from the value of the magnitude.

Figure 3.8: Biplot of loadings and scores among the first 2 principal components.

By looking at Figure 3.8, we observed that the biplot doesn't add much information to what we learnt by looking at the previous plots, but we can observe that the features are distributed among three main directions:

- horizontally on the right we can observe all the features regarding the measurements of the photon flux density

- vertically (slightly oriented to the left, just like the scores) we find the measurements of magnitude
- more on the left we find the measures of galaxy size

We can therefore assume that the measures of magnitude are linked to the ones of size (which makes sense, since one of the measures of size is the magnitude on the red band) and almost opposite to the one of photon flux density, highlighting the difference between the two kinds of measurement.

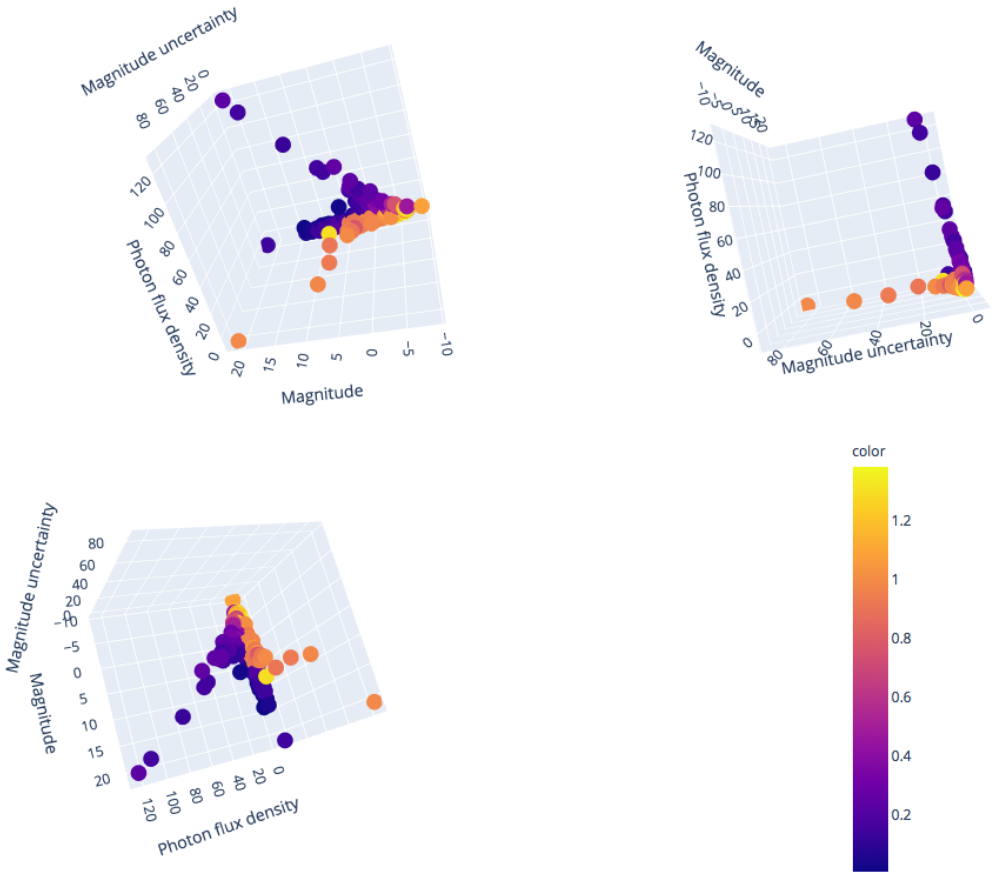


Figure 3.9: The 3D plots confirm our hypothesis about how the components interact: high values of the magnitude uncertainty rise the red-shift independently from the value of the magnitude (top left), the highest values of red-shift can be found for low values of photon flux density and low values of magnitude (top right), photon flux density and magnitude uncertainty are uncorrelated (top right) and the magnitude uncertainty is proportional to the red-shift values (bottom).

Chapter 4

k-NN classifier

To complete the assignment, we applied the k-Nearest Neighbour classifier to estimate the target value **Mcz** for the galaxy observations contained in the evaluation set. In order to do this, we first represented the data into the m-dimensional Principal Component space, where m = 13, and then we used the knowledge we obtained on **Mcz** of the development set so to compute the values in the evaluation set.

4.1 k-NN

Considering the evaluation set ("red-shift features" excluded of course), we represented it in the 13-dimensional PC space, where m = 13 was selected by working on the PCA of the development set. We then applied the k-NN classifier on the evaluation set represented in 13-dimensional PC space, considered again without the redshift features since our aim was to predict the values of **Mcz**. By performing a grid search in order to choose the best combination of weight and number of neighbors for the classifier, we obtained the following results:

MAE	MRE	R2 score
0.0561	0.0996	93,73%

Table 4.1: k-NN calssifier on evaluation data represented in 13-dimensional PC space with weight = 'distance' and n_neighbors = 7

In order to make it more general, since the number m = 13 was chosen empirically, we decided to implement the model for all possible numbers of components between 2 and 15 and we performed everytime a grid search in order to select for each number of components the best parameters concerning the weight and number of neighbors of the k-NN. We made this choice also because it is known that in general the k-Nearest Neighbor works better in low-dimensional datasets, so it is possible that 13 is not the ideal number of PCs to achieve a good approximation of the red-shift with this classifier.

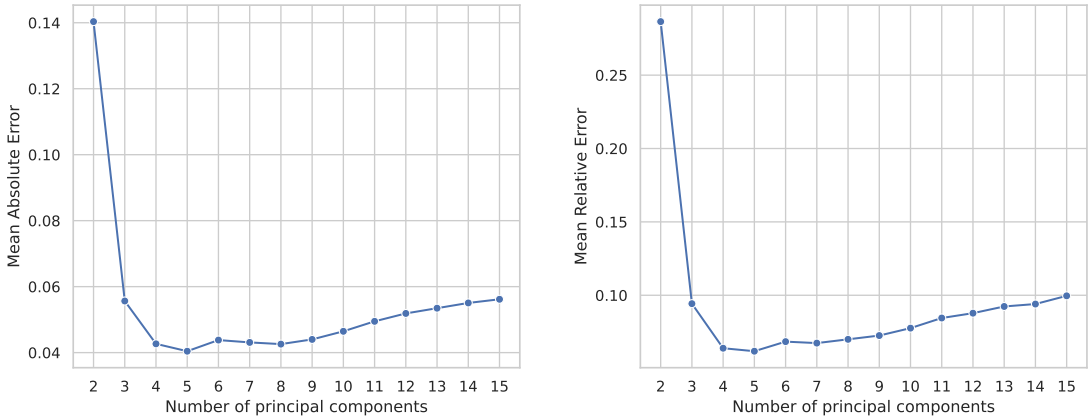


Figure 4.1: Mean Absolute and Relative Error

By looking at Figure 4.1 is it possible to observe that the best value for both the Mean Absolute Error and the Mean Relative Error is reached when we choose 5 principal components (with n_neighbors = 7 and weight = distance). Furthermore, with 5 PCs we also have the best performace of the model: the R2 score reaches the 96,53% (see Figure 4.2 below).

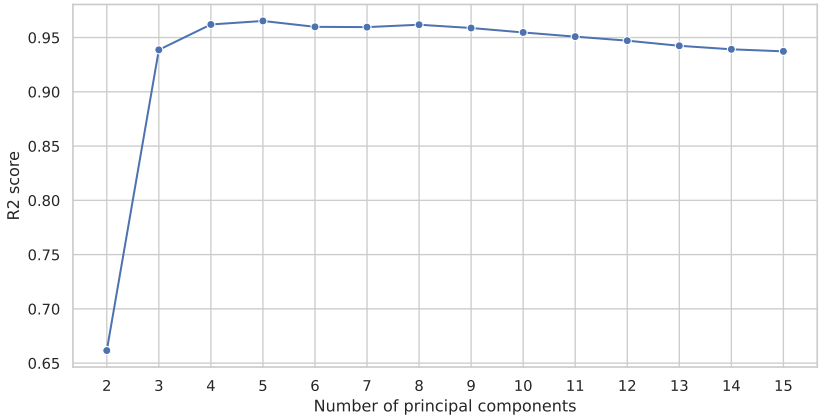


Figure 4.2: R2 score

Chapter 5

Conclusion

When a dataset contains redundant features, applying techniques of dimensionality reduction and feature extraction like PCA can be fundamental to highlight the useful information. The purpose of the work can also influence the number of components considered: if one wants to perform an explorative analysis of the dataset it's important to mantain most of the variance explained by the original dataset, whereas for most specific target (like classification or regression of a target) it can be useful to further reduce the number of features considered, since many of them can be not significant for the task.
In this case our dataset had 61 features, but almost all of them were measures of the brightness of the galaxy, the only variation being in the metric used or in the bandwidth considered. This redundancy has been highlighted by verifying that almost all of the variance of the dataset is explained by less than 15 of the original 61 features considered. We tried to give an interpretation to these new components, but since the division is performed according to the variance explained sometimes they may not follow the human natural interpretation of the data.
Some of these components were already really good to separate the data according to the red-shift value, thus this suggested us that we could consider even fewer components for our classificator: it turned out that the best results on the evaluation set can be accomplished with just five components, therefore working on a new dataset more than ten times smaller compared to the original one.