Estimating Photometric Redshift using Machine Learning Techniques

December 10 to 21, 2012

Project Guide

Prof. Ninan Sajeeth Philip

Abhijith Varma

Kumar Ayush

Prashansa Gupta

Abhinav Chikhalkar

Alankar Kotwal

Contents

Table of Figures

1. Introduction

"*The important thing is not to stop questioning. Curiosity has its own reason for existing."* From times immemorial, we humans have been extremely curious about knowing the universe surrounding us; understanding the stars and galaxies and making sense out of it in some way. One of the important questions that we have been looking for is to determine how far the objects we see in the sky are? Can we go there? How and with what confidence levels can we answer? What is the map of the universe? Can we make one? This report tries to answer some of these questions, using the basic knowledge of astronomy and merging it with a novel computational tool called ***machine learning***.

The only type of signal we get from astronomical objects is electromagnetic radiation. As a start we are restricting ourselves to only galaxies. Intensive study over the years has given us a lot of insight into the *spectra* of different galaxies. The spectrum is found to have a lot of *characteristic elemental emission and absorption lines* which are quite accurately predicted. These lines are not quite observed at exactly the same wavelengths as expected. The lines are shifted from its exact position to a little higher or lower wavelength, which quite accurately gives us the information about the velocity with which the galaxy is receding from or approaching us. The empirical *Hubble's Law* could subsequently be used to calculate the distance to the galaxies. Measuring *redshift* is the only way to calculate distance to distant unresolved galaxies.

But, this way of measuring has its own downside. The observation time is quite high and the accuracy of measurement is highly dependent on the Signal to Noise Ratio, which limits its use drastically. It is not quite possible to use this method to find redshifts of a large number of objects. A faster though less accurate way of estimating redshift was then developed and adopted viz PHOTOMETRIC REDSHIFT.

Photometry is the science of measurement of light in terms of its perceived brightness to the human eye/ photographic plate. In astronomy it is used as a technique concerned with measurement of flux or intensity of an astronomical object's electromagnetic radiation.

The brightness of the galaxy is viewed through various standard filters which let through some band of wavelengths. Astronomers traditionally use a logarithmic scale to measure the the flux coming through a filter and call it the magnitude of the object in that filter. The difference in magnitudes between two such filters is named color. Thus color in astronomy is different from what it generally represents. We obtained data for large number of galaxies and plotted a color versus color plot. To understand how they evolve with redshift, a color scale was used to represent redshift. It is observed that galaxies with a particular redshift lie in a particular region and the region evolves with redshift. If photometric data of galaxy with unknown redshift is plotted on the same plot; from the position where the point lies in the plot, it is possible to estimate the redshift of the new galaxy. This is the principle of photometric redshift estimation. One can then determine the distance of the galaxy with the help of the Hubble's law.

For photometry, the wavelength bins (filters) are much larger compared to the bins used in spectroscopy. This requires only a short exposure time to reach the same signal-noise ratio. In addition, imaging detectors generally cover a greater area of the sky than multi-object spectrographs. For this reason, photometric redshifts can be measured much faster and in larger quantities than their spectroscopic counterparts.

In the modern world, science has a new frontier ‘Computation’, which tags along with Theory and Experimentation. Though photometric data of Galaxies are apparently related with their redshifts, it is an extremely difficult task to determine the relation using our knowledge of physics and astronomy. Hence, we use computers.

Machine learning is a paradigm in computing where we create programs which ‘learn’. In simple words, they can compare their output with the expected output and modify themselves to perform better. Our project focuses on analyzing different machine learning techniques to determine photometric redshifts.

1. Theory

1.1. Artificial Neural Network (ANN)

It is essentially a network of variables which map our input variables to output variables. We will concentrate on a particular kind of ANN known as 'multi-layer perceptron'.

As the name suggests, it has multiple layers of data variables. There is essentially an input and an output layer and an arbitrary number of hidden layers in between based on our application. Each element of n^{th} layer is connected to $(n+1)^{th}$ layer by connections which have weights on them. A signal sent through these connections is modified by the weights as it reaches the next layer. These weights are responsible for mapping the input layer to the output layer. Modification of these weights can vary the output and this is the how an ANN learns. Comparing the output with expected output or target variables, it modifies the weights in order to minimize the error.

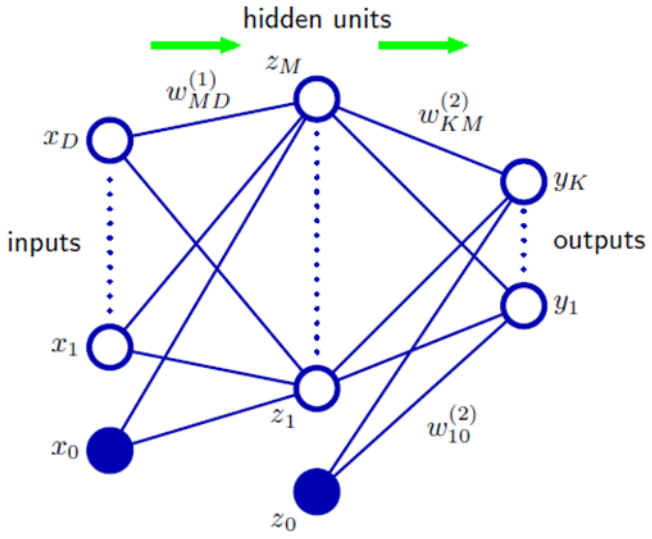

Figure 1-i A multi-layer Perceptron

1.2. Dataset Generation

Machine learning (ANN in our case) as discussed above helps us to predict redshift of different galaxies; and its accuracy of estimation greatly depends on the quality of training dataset used. A completely random dataset encompassing galaxies over a wide range of redshifts is required for the ANNs to be usable. Practically, redshift data of distant – high redshift – galaxies available from spectroscopy also involves a lot of error and hence is not quite reliable; hence the need to synthesize a dataset comes up.

Spectroscopic data is complete data, i.e. all the information about the galaxy, in the form of radiation, is stored in its spectrograph. Therefore, in theory, one could predict the photometric response of the galaxy, given its spectrum and the response of each filter. We used the Sloan Digital Sky Survey (SDSS) data for our study. The SDSS uses five filters – u, g, r, i and z – for the photometric

analysis whose responses are well catalogued (shown in *Figure 2-ii*). A large number of good spectrums, i.e. having high S/N ratio, are available for nearby galaxies. Since, these galaxies have well resolved spectrum, their redshifts are quite well known. Since the Physics of redshift is well understood, we could in principle move this galaxy to any distance in space and predict its spectra. Hence, as first approximation, assuming that:

1. Spectra of similar type of galaxies would remain same, even at extrapolated redshifts
2. Population of different types of galaxies remain at constant ratio with distance,

We could synthesize spectra for higher redshifts. And since the response of filters is also known, galaxy's color could be estimated. Since we don't need the exact color magnitude for a galaxy, we don't use Hubble's Law and hence the deviation of Hubble's constant with redshift has no effect on our calculation.

Additionally, all the observed spectra of different types of galaxy could be red shifted to $z=0$. Since, different types of galaxies lie in different regions, the dataset thus created could be used to classify galaxies into its different types, i.e. once their redshifts have been estimated.

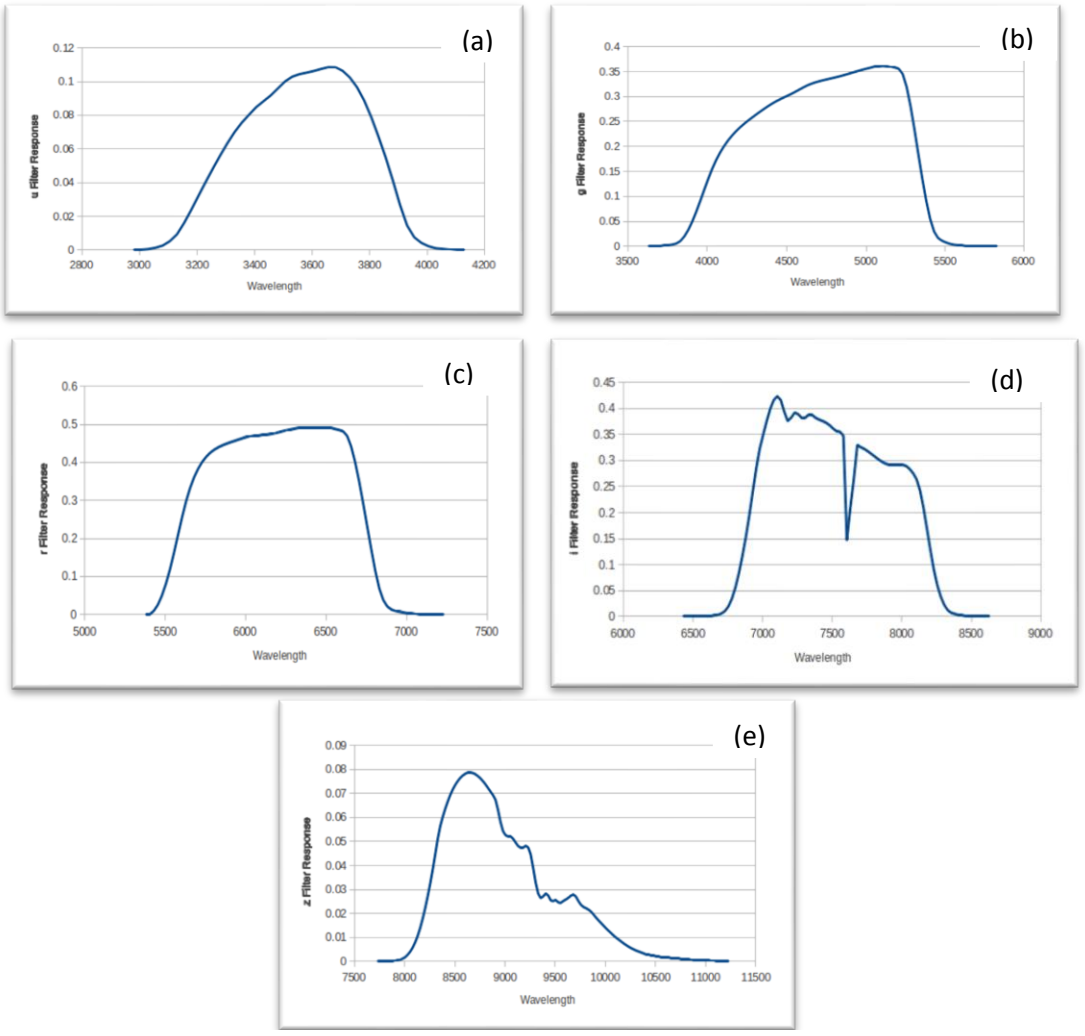

Figure 1-ii Filter responses: (a) u-filter; (b) g-filter; (c) r-filter; (d) i-filter; (e) z-filter

1.3. The Bayesian Network

The Bayesian Method of Machine Learning uses Bayesian probability to give a probabilistic meaning to the value of variables.

It defines conditional dependencies (independencies) among the variables involved.

The bayes theorem is a popular tool to estimate the outcome when there is lot of uncertainty in the inputs, which is very much relevant in the case of astronomical data.

In Astronomy, the only way through which we can make any observations is through *photons.* We can build spectra of each object, but the intensity will decrease as we increase spectral length. Therefore, we try to observe the photons through some *filters* specific to some wavelengths. These filters can be u, g, r, i, z etc.

Assuming the photometric data includes only four colors u-g, g-r, r-i, i-z., since the SDSS data gives only five u, g, r, I, z filters.

We would like to estimate the probability distribution of redshift given a set of colors for an unknown object by computing the Bayesian Probabilities.

2. Procedure and Results

2.1. ANNs

We coded an artificial neural network based on the Back-propagation algorithm. Our neural network had just three layers-input, output and one hidden layer. It is more convenient to say that the network has two layers-since it has two layers of connections which can be modified. The network was trained on a dataset of size 13,156. We compared the calculated redshifts with the spectroscopic redshifts of the galaxies in the dataset to train our neural network. In order to judge how our network has performed, two graphs were plotted.

1. Photometric Redshift v/s Spectroscopic Redshift

   We aim for a straight line of slope 1

2. Error Number Frequency Distribution

   Histogram of error -- Photometric Redshift-Spectroscopic Redshift

Two important parameters of a neural network which is to be decided by us in the simplest scenario are 'learning rate' and 'number of nodes in hidden layer'. We tested our neural network on various learning rates and number of nodes in hidden layer. We show the best two of our results

Results

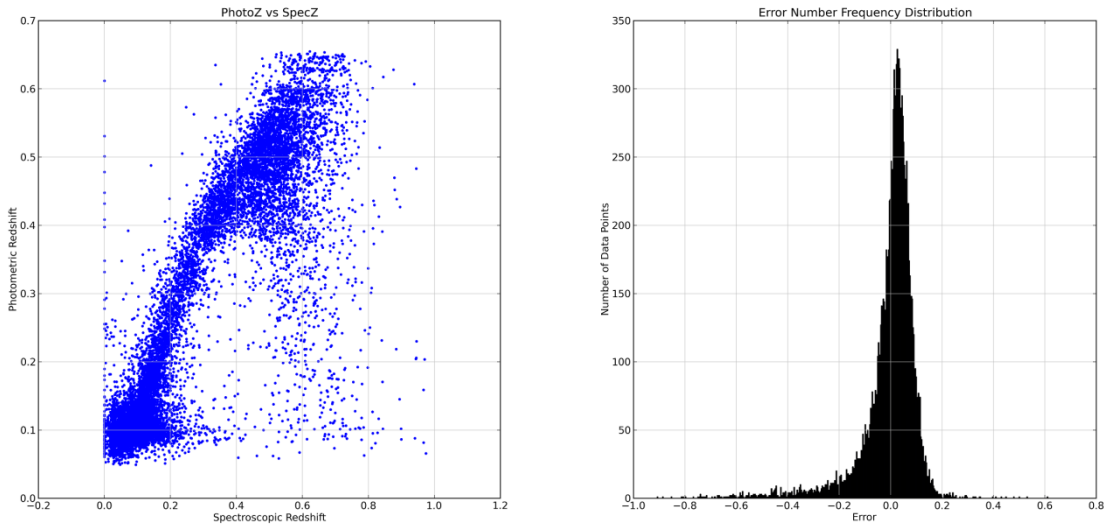

Figure 2-i learning rate: 0.05, Number of nodes in hidden layer: 1

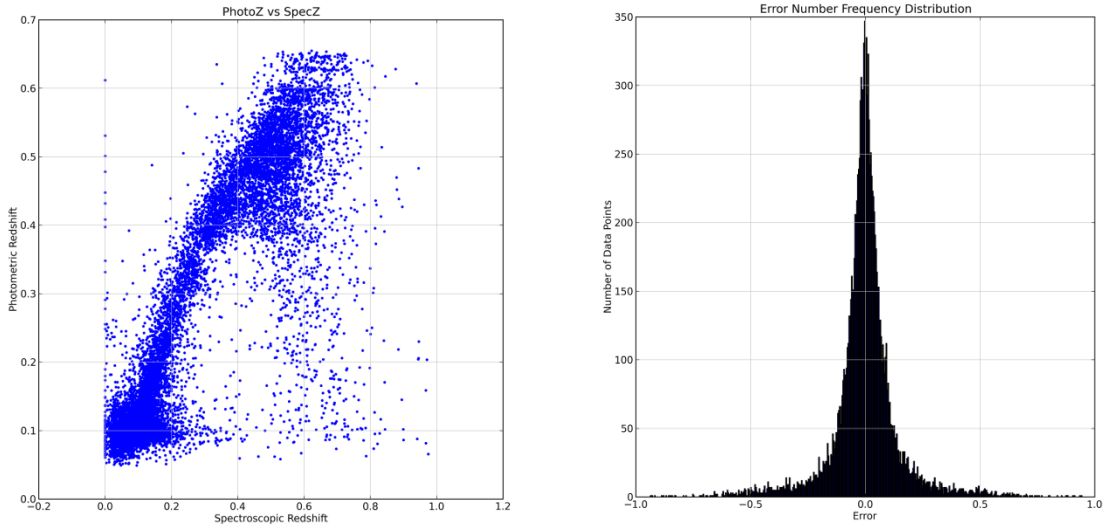

Figure 2-ii Learning rate: 0.25, Number of nodes in hidden layer: 4

2.2. The Bayesian Way

Our aim is to estimate the probability distribution of redshift given a set of colors for an unknown object by computing the Bayesian Probabilities.

First, we compute the *posterior* probability of any one color, say u-g. This is basically the probability distribution of redshift given particular u-g
This *posterior* probability distribution for u-g becomes *prior* probability distribution distribution for g-r, which can be in turn used to compute *posterior* probability distribution for g-r.
This process can be carried out for all the colors taken into consideration.
Finally, the thus obtained probability distribution signifies *the probability distribution of redshift given a set of photometric data.*

In the present situation, we are interested in calculating the Probability Distribution of Redshifts in a given color say u-g.

- We bin the u-g data and the Redshift data in the training set.
- Then estimate the likelihoods for each redshift value for a given u-g value.
- Finally plot u-g versus Redshift, coloring different values of probability differently.

Similar procedure can be followed for all the colors separately.
The probability plots are as shown. The bin sizes of the colors and redshift can be further reduced so as to increase the accuracy of the results.
In all the following graphs the accuracy of estimated redshift is 0.1

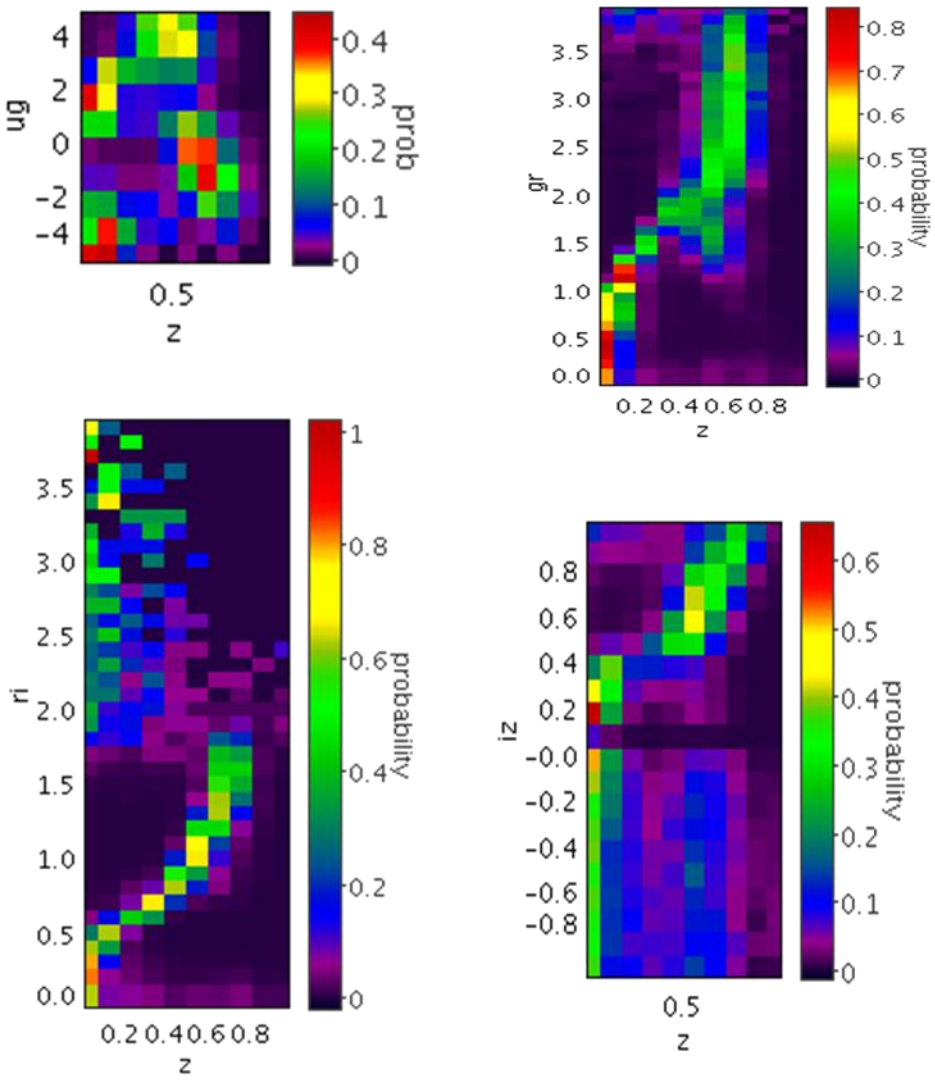

Figure 2-iii Colour vs. redshift plot with gradient being the probability

2.3. Dataset Generation

Spectrum data of Intensity (F_λ) and lambda (λ) were taken from SDSS along with their redshifts (z). Magnitudes of some galaxies were found out using Equation 2-1 to check whether similar trend is observed. Since the SDSS uses other method of calculating synthetic magnitudes, the values are not expected to match but the trend should remain similar.

$$M_i = -\frac{2.5}{\ln 10}\left(\sinh^{-1}\frac{\int_{\lambda_i}^{\lambda_f} F_\lambda(\lambda)\cdot R_\lambda(\lambda)\cdot d\lambda}{2ab} + \ln b\right) \qquad 2\text{-}2$$

R_λ being the response of the i^{th} filters. a and b are taken to be 1810 and $1.4\text{x}10^{-9}$ for u, 3730 and $0.9\text{x}10^{-9}$ for g, 4490 and $1.2\text{x}10^{-9}$ for r, 4760 and $1.8\text{x}10^{-9}$ for i and 4810 and $7.4\text{x}10^{-9}$ for z respectively. Spline of first order is considered while integration.

Since, u and z bands have a lot of error involved due to atmospheric absorption at lower lambda and limited data at higher lambda, essentially only three magnitudes g, r, and i are considered and hence only two colors: g-r and r-i.

Results thus obtained by shifting all the data to z=0, 0.1, 0.2, 0.3, 0.4 and 0.5 are shown. The data of actual galaxies with different redshifts is also plotted (*Figure 3-vFigure 3-iii*). The synthetic data matches the observed data quite well for low redshifts, but for higher redshifts tend to differ. This could primarily because the evolution of spectra of different galaxies over time is not considered. Given enough amount of time we could in principle account for the evolution of spectra using presently accepted models and synthesize data for higher redshifts too. This not only would help us predict redshifts more accurately, but also test for our spectra evolution theories in the nearby universe.

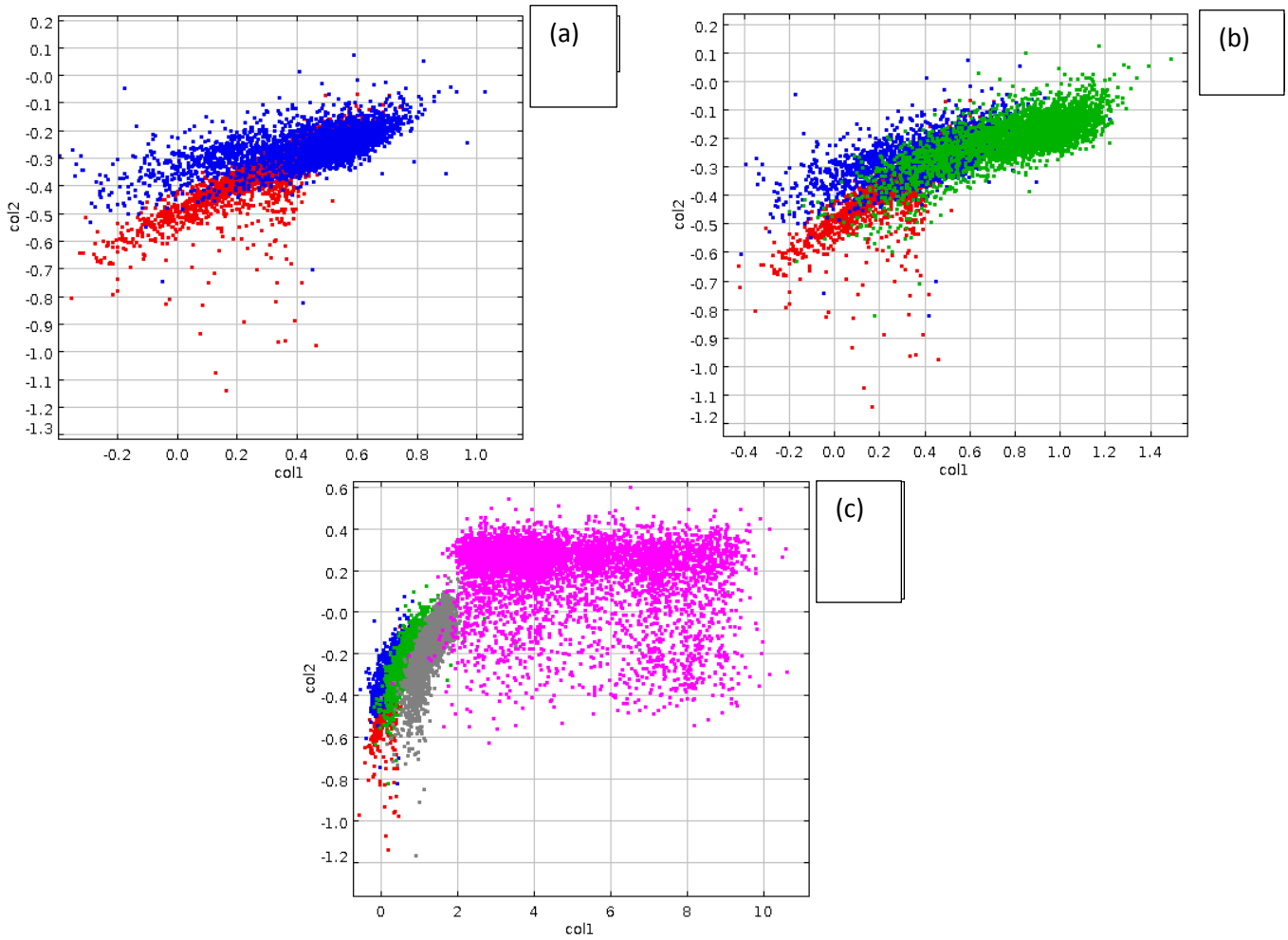

Figure 2-iv Sythetic data on a colour-colour plot: (a) redshift of 0(red) and 0.1(blue); (b) redshift of 0, 0.1 and 0.2 (green); (c) redshifts of 0, 0.1, 0.2, 0.3 (grey) and 0.5 (pink). The synthetic data with redshifts of 0.5 and above didn't quite match

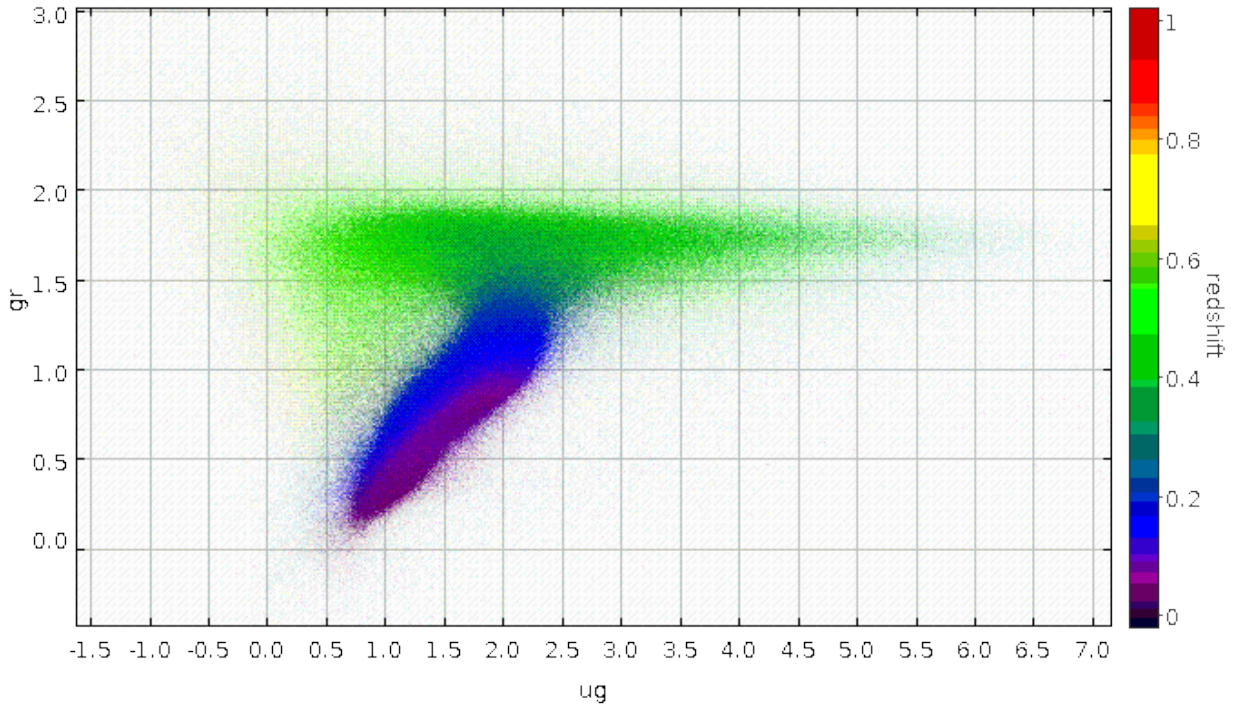

Figure 2-v Colour-colour of approx. a million galaxies whose redshifts are known

3. Acknowledgments

First and foremost we would like to acknowledge our guide, Prof. Ninan Sajeeth Philip, who was extremely helpful and always full of vigour. He gave us full liberty to develop the problem and guided us through.
We are grateful to our coordinator at IISER Mohali, Dr. Harvinder Kaur Jassal for being such a wonderful host and our local coordinator Dr. Aniket Sule for giving us this excellent opportunity.
We are thankful to Dr. Sivarani Thirupathi for helping us with the code and sharing ideas during the span of the project.
Last but not the least we are indebted to our parents for always supporting us and instilling in us a will power to overcome our problems.