Draft version December 7, 2024
Typeset using LaTeX **preprint** style in AASTeX631

Photometric redshifts in SDSS

Hanzhang Feng

ABSTRACT

This lab examines a simple photometric redshift model $z_p = b_0 + \sum_{k=1}^{5} b_k m_k$ that estimates redshift of a galaxy from its model magnitudes in {u, g, r, i, z} band. With galaxy data retrieved from SDSS, coefficients $b_k, k = \{u, g, r, i, z\}$ were found to be $0.01455839, 0.08880067, 0.00486052, -0.0258905, -0.02536246$ respectively, and $b_0 = -0.83$. The variance and bias of this model are 0.0024, -0.004 respectively. The root mean square of error ϵ is 0.049, and the mean value of error ϵ is -0.003, corresponds to percentage error of %1.95. Overall, this model successfully predicted the redshift for most galaxies with low error. However, the accuracy of model is lower for larger redshift ($z_s > 0.25$) galaxies, which may be caused by the bias in data sample that contains significantly more low redshift galaxies than high redshift galaxies. Nine representative galaxies were picked from samples that were poorly fit, averagely fit and accurately fit. Their galaxy image and spectrum data were included at the end of this paper as references.

1. INTRODUCTION

Redshift is the increase of emitted wavelength due to motion of object or gravity from nearby object. It can be used to determine velocity and mass of nearby binary stars and measure the expansion of universe. The Sloan Digital Sky Survey (SDSS) is a survey that tracks redshift with multi-spectral imaging and spectroscopic observation. Its latest data release, DR17, was published at April 2022 by Abdurro'uf et al. DR17 is the fifth and final release from the fourth phase (SDSS-IV) that contains SDSS observations through January 2021. The fourth phase SDSS-IV was composed of three main surveys targeting different ranges of scale: 1. APO Galactic Evolution Experiment 2 (APOGEE-2) that maps stars in the Milky Way and nearby satellites; 2. Mapping Nearby Galaxies at APO (MaNGA) that maps nearby galaxies; 3. Extended Baryon Oscillation Spectroscopic Survey (eBOSS) that targets cosmological scale observation (Abdurro'uf et al 2022). Following the SDSS tradition of cumulative data release, DR17 also includes data reduction of previous data releases such as SDSS Legacy (Abdurro'uf et al 2022).

SDSS estimates a redshift for each spectrum. While the spectroscopic data allows the most accurate determination of galaxy redshift, a large portion of galaxies lack such data due to observational constraints, and hence require alternative way to estimate their redshift. One of such ways is to estimate redshift with photometric data. This lab aims to assess the performance of a basic method for photometric redshift estimation. The model adopted in this lab is a simple model that estimates redshift of a galaxy from its model magnitudes in its {u, g, r, i, z} band. This band system was developed by SDSS, corresponding to ultraviolet, green, red, near infrared and infrared band.

The parameters of this photometric redshift model were solved from a training set of galaxy data. These parameters were then applied to a separate set of galaxies to test the accuracy of this model. Both sets of galaxy data were retrieved from Casjobs, which is a SDSS service hosted in SciServer. Casjobs is an online workbench that allows longer catalog queries than regular web search interface. Another SciServer service used in this lab is SkyServer. SkyServer is an interface that retrieves data by SDSS objID. Its Quick Look tool provides galaxy images as well as their spectrum data.

This paper will present, evaluate, and discuss the simple photometric redshift model used. The first part of Method section presents the criterion used to select two sets of galaxy data for training and evaluation. The second part of Method section introduces the model and presents steps to rearrange the model equation and perform reduce chi square. This allows model parameters to be solved from a simple system of linear equations that involves the covariance of observed data only. The Results section tabulates parameters solved from the training set, and presents results of fitting these parameters to the target set by including plots of predicted redshift and their error against true redshift, as well as a histogram of error. Overall performance of this model, such as variance, bias, and outliers are also included in Results section. The Discussion section interprets the resulting fit of target set and discusses possible source of error as well as possible improvements for this estimator.

2. METHOD

2.1. *training set and target set*

Both the training set and the target set were obtained from the latest SDSS data, DR17. Using Casjob query, 787 galaxies were selected into the training set, and 772 galaxies were selected into the traget set. Galaxies in both sets were recorded with their unique IDs, redshifts, 5 band model magnitudes and their corresponding errors. The query used in this lab selects galaxies that are primary objects in SDSS legacy survey. The training sets were selected to be composed of galaxies with their unique ids being an integer multiple of one thousand plus one, or equivalently, galaxies that have their unique IDs ended with "001". The target set were composed of galaxies with ID ending "002". By such selection, two independent sets with reasonable and comparable sample size were randomly selected from the data base.

2.2. *photometric redshift model*

The photometric redshift model used in this lab is

$$z_p = b_0 + \sum_{k=1}^{5} b_k m_k, \tag{1}$$

where $k = 1, 2, 3, 4, 5$ represents{u, g, r, i, z} band. $\bar{m_k}$ are the average values of model magnitudes in each band, and b_k are a set of real scalar coefficients to be determined. After a set of b_k is obtained, b_0 can be calculated for each row of data. The final b_0 to be used in the model is be the average of all b_0 for each row. An easier and equivalent way of obtaining b_0 is to calculate it from the average values of redshifts and model magnitudes:

$$b_0 = \bar{z_s} - \sum_{k=1}^{5} b_k \bar{m_k}. \tag{2}$$

The easiest way to solve for b_k is to set up a system of linear equations with $b_k, k = 1, 2, 3, 4, 5$ as independent variables, while using given or calculated values from data as dependent variables and coefficients. This lab set the covariance between redshift and magnitudes as dependent variables, and a matrix of covariance between different band of magnitudes as coefficients. The derivation is shown below.

For each row in data set denoted by index i, Equation 1 can be rewritten as

$$(z_{is} - \bar{z}_s) = b_0 + \sum_{k=1}^{5} b_k(m_{ik} - \bar{m}_k), \tag{3}$$

where z_{is}, m_{ik} are redshift and magnitude given by each row of data, $\bar{z}_s, \bar{m}_k$ are mean value of redshift and magnitude calculated from the whole data set.

Equation 2 separates observational data on the left hand side, while theoretical values on the right hand side. This allows the chi square to be written as

$$\chi^2 = \sum_i \left[\frac{y_{ob,i} - y_{th,i}}{\sigma_i}\right]^2 = \sum_i \left[\frac{(z_{is} - \bar{z}_s) - \sum_{k=1}^{5} b_k(m_{ik} - \bar{m}_k)}{\sigma_i}\right]^2 \tag{4}$$

To minimize χ^2, take its derivative against b_k and set equal to zero. Denoting the expression inside the square bracket as χ_i, this derivative can be written as

$$0 = \frac{d}{db_k}\sum_i \chi_i^2 = \sum_i \frac{d}{db_k}\chi_i^2 = \sum_i 2\chi_i \frac{d}{db_k}\chi_i \tag{5}$$

Substitute the expression for χ_i back to $\frac{d}{db_k}\chi_i$ and eliminate terms independent of b_k to get

$$\frac{d}{db_k}\chi_i = \frac{d}{db_k}\frac{(z_{is} - \bar{z}_s) - \sum_{k=1}^{5} b_k(m_{ik} - \bar{m}_k)}{\sigma_i} = \frac{-1}{\sigma_i}\frac{d}{db_k}\left[\sum_{k=1}^{5} b_k(m_{ik} - \bar{m}_k)\right] \tag{6}$$

Since the term being differentiated with respect to b_k is a sum of k, a change of variable to b_l is useful:

$$\frac{d\chi_i}{db_k} = \frac{d\chi_i}{db_l}\frac{db_l}{db_k} = \frac{-1}{\sigma_i}\frac{d}{db_l}\sum_{k=1}^{5} b_k(m_{ik} - \bar{m}_k)\frac{db_l}{db_k} = \frac{-1}{\sigma_i}\sum_{k=1}^{5}\frac{db_k}{db_l}(m_{ik} - \bar{m}_k)\frac{db_l}{db_k} \tag{7}$$

The differentiation $\frac{db_k}{db_l}, \frac{db_l}{db_k}$ are delta function δ_{lk}. This simplifies $\frac{d\chi_i}{db_k}$ to $\frac{d\chi_i}{db_k} = \frac{-(m_{il} - \bar{m}_l)}{\sigma_i}$. Substitute this expression back to Equation 4 to get

$$0 = \sum_i 2\chi_i \frac{d\chi_i}{db_k} = -2\left[\sum_i \frac{z_{is} - \bar{z}_s}{\sigma_i}\frac{m_{il} - \bar{m}_l}{\sigma_i} - \sum_{k=1}^{5} b_k \sum_i \frac{(m_{ik} - \bar{m}_k)}{\sigma_i}\frac{(m_{il} - \bar{m}_l)}{\sigma_i}\right] \tag{8}$$

which could be written in form of

$$cov\left(\frac{z_s}{\sigma_i}, \frac{m_l}{\sigma_i}\right) = \sum_{k=1}^{5} b_k cov\left(\frac{m_k}{\sigma_i}, \frac{m_l}{\sigma_i}\right) \tag{9}$$

Take uniform weighting $\sigma_i = 1$, Equation 9 can be written in matrix form:

$$\begin{bmatrix} cov(z_s, m_u) \\ cov(z_s, m_g) \\ cov(z_s, m_r) \\ cov(z_s, m_i) \\ cov(z_s, m_z) \end{bmatrix} = \begin{bmatrix} cov(m_u, m_u) & gu & ru & iu & zu \\ ug & gg & rg & ig & zg \\ \vdots & & \ddots & & \vdots \\ \vdots & & \ddots & & \vdots \\ uz & & \dots & & zz \end{bmatrix} \begin{bmatrix} b_u \\ b_g \\ b_r \\ b_i \\ b_z \end{bmatrix} \tag{10}$$

Coefficients b_k can be obtained by solving this system of equations.

3. RESULTS

Solving with data from training set, coefficients b_k are

$$\begin{bmatrix} b_u \\ b_g \\ b_r \\ b_i \\ b_z \end{bmatrix} = \begin{bmatrix} 0.01455839 \\ 0.08880067 \\ -0.00486052 \\ -0.0258905 \\ -0.02536246 \end{bmatrix} \tag{11}$$

and $b_0 = -0.83$.

By fitting this model with magnitudes in target set, a list of predicted photometric redshifts z_p were obtained. To test the accuracy of this model, a list of errors between predicted redshift and true redshift $\epsilon_i = z_{is} - z_{ip}$ was calculated. A plot of predicted redshifts z_p against true values of redshift z_s for the target set is shown on top left panel in Figure 1. A line of gradient 1 and 0 y-intercept was plotted in orange as a reference. The plot of error ϵ against true redshift z_s is shown on top right panel in Figure 1. A line of gradient 0 at $\epsilon = 0$ was plotted in orange as a reference. Bottom panel of Figure 1 is a histogram of ϵ.

With outliers defined to be galaxies with redshift error $\epsilon > 0.13$, nine outliers in the target set were found. The outliers are marked red in predicted redshift plot and error plot in Figure 1. Their unique IDs are tabulated below.

1237659346953110002	1237664669518201002	1237654880198394002
1237658300595110002	1237668270844151002	1237652946916606002
1237657628431680002	1237662246595789002	1237655369287336002

The variance and bias of this model are 0.0024, -0.004 respectively.

4. DISCUSSION

This model successfully estimates redshifts for most galaxies with low error.

The root mean square of error ϵ is 0.049, and the mean value of error ϵ is -0.003, corresponds to percentage error of 1.95%. The shape of histogram of error ϵ follows Gaussian distribution and centered at 0.

While the negative bias indicate that this model tend to underestimate redshifts, the top left panel of Figure 1 shows an overestimate of redshifts for low redshift galaxies, and an underestimate of redshift for higher redshift galaxies.

The precision of this model decreases for galaxies with higher redshifts. Beyond $z_s > 0.25$, almost all redshifts are underestimated. One possible source of error is that the data sample is mostly composed of low redshift galaxies. In both the training set and the target set, 88% of galaxies have redshift $z_s < 0.25$. Training with a predominantly low redshift data set, the resulting model might have parameters that better reflects low redshift galaxies, and hence make less accurate estimation for higher redshift galaxies. To eliminate this error, a traning set with samples evenly distributed across redshifts should be adopted.

Figure 2 shows galaxy images for nine representative galaxies retrieved from SkyServer. The top row of Figure 2 shows three galaxies that were fit poorly; middle row shows averagely fit galaxies and bottom row shows accurately fit galaxies. The spectrum data of these nine galaxies are shown in Figure 3 with the same order. There are no significant trend that shows difference between poorly fit galaxies and accurately fit ones in both galaxy image and spectrum data. However, there seems to be a subtle trend indicating that poorly fit galaxies have more fluctuations in spectrum data than that of accurately fit galaxies.

5. FIGURES

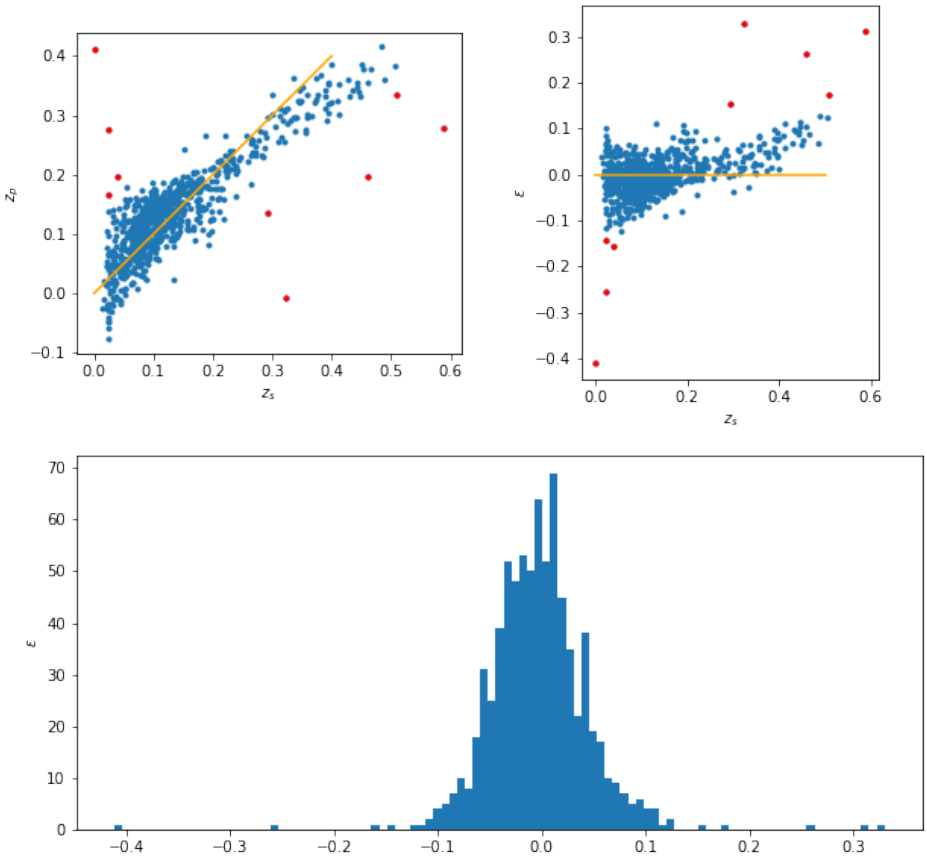

Figure 1. Top left: plot of predicted redshifts from model against true redshifts for target set. Outliers are marked red. Top right: plot of errors between predicted and true redshifts for target set. Outliers are marked red. Botton: histogram of errors for target set.

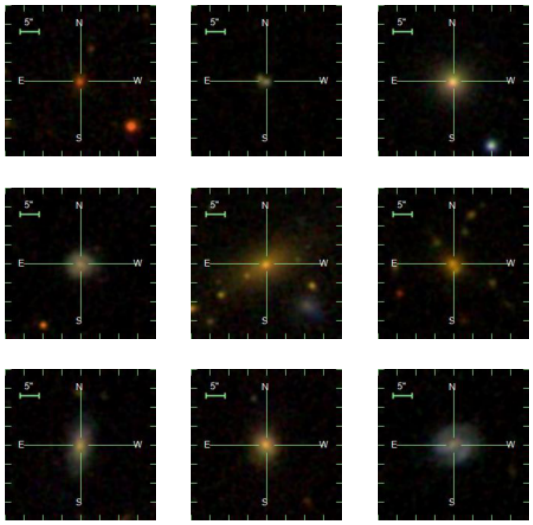

Figure 2. Representative galaxy images. Top row: redshift poorly estimated. Middle row: average accuracy. Bottom row: accurately estimated.

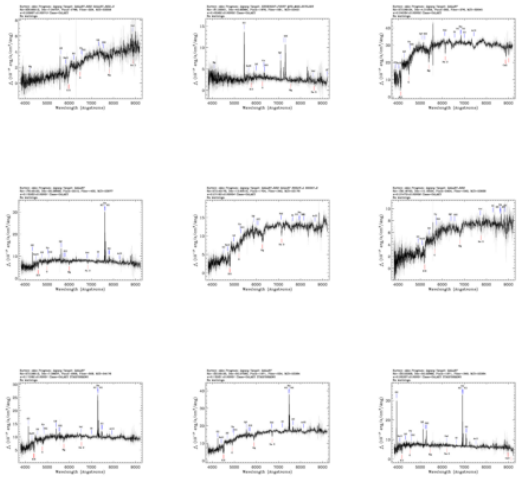

Figure 3. Representative spectra data. Top row: redshift poorly estimated. Middle row: average accuracy. Bottom row: accurately estimated.