Loránd Eötvös University
Faculty of Science

# Improving photometric redshift estimation techniques

Theses

**Norbert Purger**

Supervisor
**Dr. István Csabai**
Department of Physics of Complex Systems

Assigned Head of PhD School: Dr. Ferenc Csikor
Head of PhD Program: Dr. Ferenc Csikor

Budapest
**2011**

# 1. Introduction

The thesis provides insights into the techniques of photometric redshift estimation. With the ability to explore the positions and other properties of the more than 100 billion galaxies we could get key information about the structure and evolution of the Universe. The current observational techniques on the other hand limit the opportunity for such exploration. The Sloan Digital Sky Survey (SDSS-I) project created images of the quarter of the entire sky during its operation and identified around 230 million extended objects, galaxies. But the same project measured the “spectroscopic fingerprint” and distance indicator – the spectra and its redshift – of only about 1 million galaxies. This less than half percent ratio demonstrates very well the technical limits, the time and cost limitations of the measurement procedure. Photometric redshift estimation methods were developed to amend this efficiency: Enabling the estimation of the redshifts solely based on the photometric information available from object and reference spectral templates, models or spectra measured earlier..

What previously has also been called the “poor man’s spectroscopy”, now it has become an elemental part of the largest ongoing and planned imaging astronomical surveys. The ambitious plan is that one will be able to give photometric redshift estimates with a characterized limited precision – for example in the case of LSST – for up to 10 billion galaxies. Even the knowledge of the approximate redshifts gives important statistical information in the case of large datasets. Based on better spatial separability, galaxy clusters can be selected more properly, estimate distances of supernovae host galaxies can be determined. Small distortion of galaxy images, like statistically measurable gravitational weak lensing, together with the separation of the distances into buckets enables to determine precise correlations, or even mass profiles of galaxy clusters. Barion acoustic oscillation measurements could also profit from photometric redshifts. Such techniques and the potential measurement opportunities are expected to contribute substantially to the understanding of dark matter and dark energy – the components that stand for the 95% of the Universe.

The aim of my PhD program was the improvement of the photometric redshift estimation methods - especially the so-called empirical methods - in order to help fulfilling the requirements of the large astronomical surveys of the future.

## 2. Applied methods

The catalog of the Sloan Digital Sky Survey provides an excellent environment to examine photometric redshift estimates, since both large amount of photometric and many follow-up spectroscopic measurements are available. The SDSS and other large astronomical catalogs can only be handled and accessed efficiently using database technology. In order to handle the measured datasets, the competence of handling and developing based on such database management systems was required. As specific methods, the cross-matching of objects based on data available from two different surveys and efficient search methods with multidimensional photometric data were used intensively. For the former I applied searches based on angular distances and the nearest neighbor (HTM and zone index). For the latter I participated in the development of an efficient search method implementation based on the kd-tree data structure. This search technique also played a key role in the empirical redshift estimation performed on large data sets. The creation of the complementary photometric redshift catalogs for SDSS DR6 and DR7 required skills for loading data into a customized database. While working with the repeated observations of the Stripe 82 catalog, I applied the co-addition of observation measurements based on flux.

The template-based photometric redshift estimation required the handling of spectra and stellar population synthesis models. Here and for the case of the hybrid photo-z technique I used K-correction when transforming into the objects' comoving system. During the application of the empirical photometric redshift estimation I had the opportunity to develop and use many different methods from the area of machine learning and other, non-parametric statistical techniques. For example polynomial fitting, k-nearest neighbor (k-NN) methods, iterative linear fitting and other k-NN regression techniques, kernel regression, neural networks, support vector machine (SVM) methods. I used tools of descriptive statistics to compare the results of the different estimations.

# 3. Theses

1. I have investigated methods to handle astronomical measurements data efficiently. With the application of photometric redshift estimation, I have demonstrated the efficiency of an indexing technique that runs inside a database server, and uses a k-d search tree based on magnitude parameters [1], [9], [10]. Within the confines of the Virtual Observatory, I have prepared many large astronomical catalogs for simple and quick cross-matching based on coordinates with results from other surveys [7]. These catalogs are publicly available.

2. I have created photometric redshift (photo-z) catalogs based on the data of SDSS DR4, DR5 [8] and DR6 [2] using spectrum template-based photometric redshift estimation. For the calculation of the estimates I have also included the measurement results of the special photo-z spectroscopy plates created as part of the SDSS Southern Survey. The "`Photoz`" catalog of the SDSS DR6 contains photometric redshift and other physical parameters of 200 million galaxies.

3. I have developed a hybrid method that compounds the advantages of the empirical photo-z method - based on iterative linear fitting of k-nearest neighbors - and the earlier, template-based method. Using this method I have created a new photometric redshift catalog for the 230 million galaxies of the SDSS DR7. This way, based on the photometric redshifts, the catalog still provides distance modulus, K-correction and absolute magnitude values. The same time the expected error of the photometric redshift estimates calculated for the SDSS DR7 galaxies decreased significantly, close to the half of the value given by the estimates for the DR6 case [3], [5].

4. I have extended the method based on the k-NN regression with the inclusion of other, similar empirical estimators, like the modification of the local estimator, the distance function, and the calculation based on kernel regression. Based on the comparisons using these implemented methods and the neural network model, I have stated that with our method one can

simply and quickly give robust estimates, that also compare very well with the precision of the estimations based on the neural network model [5].

5. I have investigated the effects of the special photo-z measurements of faint, blue galaxies created during the SDSS Southern Survey, and the high precision, "coadded" Stripe 82 photometry on the precision of the photometric redshift estimates [6]. The photometric redshifts calculated using the reference set that also contained the calibration data of the blue galaxies led to 30% more accurate estimates in the case of faint objects with normal photometry. The more precise, deeper coadded photometry uniformly improved the quality of the estimates, decreasing the error of the photometric redshifts by up to 50%.

6. Using the "*coadded"*, deeper photometry of the SDSS Stripe 82 and the individual observations I have investigated the expectable improvement of the photometric redshift estimation precision in the case of the large, future sky surveys that build upon on the addition of many repeated observations [6]. For the tests I have cross-matched the results of many large galaxy redshift surveys with the photometric set, thus creating a deeper and representative reference set. The redshifts estimated this way show very good precision - compared to the original z~0.5 depth of the SDSS spectroscopic set - now up to z~1 with the new data. The data of the created catalog are available. I have simulated the measurements with the intermediate precision based on less repeated observations. To do this, I have averaged the properly matched individual observations by flux in several steps. These results show the improvement of the quality of the redshift estimates based on real data, depending on the number of repeated observations used in the addition.

7. As part of the "Photo-z Accuracy Testing" programme I have applied the template-based and empirical k-NN regression methods that I used in the case of the SDSS [4]. Both methods performed according to the expectations on the idealized, but noisy test dataset. Technically close – comparable to a shot noise - to the reference level of the code that was used to create the test set.

# 4. Publications

*Publications connected with the thesis*


[1] I. Csabai, L. Dobos, M. Trencséni, G. Herczegh, P. Józsa, **N. Purger**, T. Budavári, A.S. Szalay “Multidimensional indexing tools for the virtual observatory”, Astronomische Nachrichten, 2007 Sep, Volume 328, 852-857.

[2] Jennifer K. Adelman-McCarthy, …., **N. Purger** and the other members of the SDSS collaboration. “The Sixth Data Release of the Sloan Digital Sky Survey”, The Astrophysical Journal Supplement Series, 2008 April, Volume 175, 297-313.

[3] Kevork N. Abazajian, …., **N. Purger** and the other members of the SDSS collaboration, “The Seventh Data Release of the Sloan Digital Sky Survey”, The Astrophysical Journal Supplement Series, 2009 June, Volume 182, 543-558.

[4] H. Hildebrandt, S. Arnouts, P. Capak, L. A. Moustakas, C. Wolf, F. B. Abdalla, R. J. Assef, M. Banerji, N. Benítez, G. B. Brammer, T. Budavári, S. Carliles, D. Coe, T. Dahlen, R. Feldmann, D. Gerdes, B. Gillis, O. Ilbert, R. Kotulla, O. Lahav, I. H. Li, J.-M. Miralles, **N. Purger**, S. Schmidt and J. Singal “PHAT: Photo-z Accuracy Testing”,2010, Astronomy & Astrophysics, 523 A31

[5] **N. Purger**, M. Trencséni, I. Csabai, “K-nearest neighbor regression for estimating photometric redshifts , 2011, Astronomy & Astrophysics, under submission

[6] **N. Purger**, M. Trencséni, I. Csabai, “Photometric redshifts for SDSS Stripe 82”, 2011, Monthly Notices of the Royal Astronomical Society, under submission


*Other conference publications connected with the thesis*


[7] **N. Purger**, T. Budavári, A. S. Szalay, A. Thakar, I. Csabai, „Build Your Own SkyNode!”, Astronomical Data Analysis Software and Systems (ADASS) XIII, Proceedings of the conference held 12-15 October, 2003 in Strasbourg, France. Edited by Francois Ochsenbein, Mark G. Allen and Daniel Egret. ASP Conference Proceedings, Vol. 314. San Francisco: Astronomical Society of the Pacific, 2004., p.201

[8] **N. Purger**, I. Csabai,T. Budavári, Z. Győry, “Photometric Redshifts For The SDSS Data Release 5”, The Virtual Observatory in Action: New Science, New Technology, and Next Generation Facilities, 26th meeting of the IAU, Special Session 3, 17-18, 21-22 August, 2006 in Prague, Czech Republic.

[9] L. Dobos, I. Csabai, M. Trencséni, G. Herczegh, P. Józsa**, N. Purger**, „Spatial Indexing and Visualization of Large Multi-Dimensional Databases”, Astronomical Data Analysis Software and Systems XVI ASP Conference Series, Vol. 376, proceedings of the conference held 15-18 October 2006 in Tucson, Arizona, USA. Edited by Richard A. Shaw, Frank Hill and David J. Bell., p.629

[10] I. Csabai, M. Trencséni, L. Dobos, P. Józsa, **N. Purger**, T. Budavári, A. S. Szalay “Spatial Indexing of Large Multidimensional Databases.”, CIDR 2007, 207-218.