Empirical and spectral template based approaches in the analysis of galaxy data

Theses of the PhD Dissertation

Róbert Beck

Eötvös Loránd University, Faculty of Science
Doctoral School of Physics, Particle Physics and Astronomy Program

Prof. Tamás Tél
Professor, DSc
Head of the Doctoral School

Prof. László Palla
Professor emeritus, DSc
Program Leader

Supervisors:

Dr. István Csabai
Professor, DSc

Dr. László Dobos
Assistant lecturer, PhD

Eötvös Loránd University, Faculty of Science
Department of Physics of Complex Systems

Budapest, 2017
Introduction

Recent sky surveys such as the Sloan Digital Sky Survey (SDSS, York et al., 2000) provide astronomers and astrophysicists with a deluge of data, both on Galactic and extragalactic sources. With the large amount of data come new challenges in terms of processing it efficiently, but also novel possibilities via statistical analysis. The SDSS photometric catalogue contains broad-band photometric measurements of over 200 million galaxies in five filters, giving a low-resolution representation of the incident light, however, the SDSS spectroscopic catalogue incorporates detailed spectra for just over 2 million galaxies due to the longer observing time required. Moreover, sky surveys that are upcoming or currently in progress, such as LSST (Ivezic et al., 2008) or Pan-STARRS (Tonry et al., 2012), will only perform photometric measurements, with no planned spectroscopic follow-up that matches their scale.

The relative lack of detailed spectroscopic measurements is significant because generally only two spatial coordinates of extragalaxies are directly observed, i.e. their position on the celestial sphere. Many physical applications would require the third coordinate, distance, e.g. to find the absolute magnitude of a galaxy, or to construct a three-dimensional map of the Universe for a large scale structure or weak lensing study, and thus refine cosmological models. Given a detailed galaxy spectrum with identifiable spectral lines, it is relatively straightforward to obtain its redshift, and compute its distance from us using the adopted cosmology. However, when no spectrum is available, one must turn to the technique of photometric
redshift estimation.

Photometric redshift (photo-z) estimation is an attempt at mapping the redshift — broad-band colour relationship, estimating redshifts with only photometry available. It is therefore a vital tool in astronomy, and its importance will only grow as upcoming sky surveys release data. There are two main approaches to photo-z estimation, the empirical and the spectral template based approach, both with their respective advantages and disadvantages (Csabai et al., 2003; Hildebrandt et al., 2010; Dahlen et al., 2013). The former utilizes a training set with known redshifts to build an empirical model, mostly using various machine learning techniques. The latter starts from a set of template galaxy spectra, simulates the observation process, and attempts to find the best-fitting spectrum to the data. The templates can be assembled from observations, or can come from semi-analytical or theoretical models. From the point of view of applications, the main current challenge of photo-z estimation is providing accurate and unbiased measures of photo-z uncertainty, especially for parameter regions that are not well-covered by the available data.

Stellar population synthesis models are the state of the art in galaxy modelling (Bruzual & Charlot, 2003; Maraston & Strömbäck, 2011; Vazdekis et al., 2012). They involve simulating the evolution of an entire population of stars, leading to model spectra that contain the stellar emission of the population (including absorption lines), and potentially even extinction (and reemission) caused by dust. However, incorporating emission lines into the models is far from straightforward, and usually requires introducing numerous new parameters. This is because those lines are emitted by
a separate component, ionized interstellar gas clouds, and the ionization source is not necessarily the hot young stars, but can be an active galactic nucleus (AGN). By studying relationships within observed emission-line galaxy populations, perhaps the models can be enhanced to be better representations of reality, even without many additional free parameters.

Results

The dissertation focuses on my recent research into the areas outlined above, namely the machine learning analysis of emission-line galaxies, and photometric redshift estimation, both empirical and spectral template based. The work utilised primarily the SDSS spectroscopic and photometric galaxy catalogues, more specifically Data Release 7 (DR7, Abazajian et al., 2009) and Data Release 12 (DR12, Alam et al., 2015). My thesis statements are presented below:

1. In a sample of emission-line galaxies from the SDSS DR7, using a machine learning method based on k-nearest neighbours, local linear regression (LLR), I empirically estimated emission line equivalent widths from stellar continua with reasonable accuracy ($\sigma_r \approx 0.35$ relative flux error), for both star-forming galaxies and AGN hosts (Beck et al., 2016a). This approach allows empirically adding emission lines to stellar population synthesis models. The algorithm itself was implemented by me in C++.

2. I proposed a stochastic model to perform the same function as pre-
presented in statement 1., i.e. adding emission lines to stellar continua, but more practically, without having to process an extensive training set (Beck et al., 2016a). The model involved “encoding” the training set into 60 multivariate Gaussian classes with the help of the k-means clustering algorithm of R.

3. In the same SDSS DR7 emission-line galaxy sample as in statement 1., I used a supervised machine learning algorithm, a support vector machine (SVM, specifically the R implementation) to classify galaxies into star-forming and AGN host categories based on their stellar continua and emission lines (Beck et al., 2016a). This approach recovered the simpler, empirically derived separation line of Kauffmann et al. (2003) with only 6% discrepancy, reinforcing the earlier result.

4. I assembled the official photometric redshift catalogue of the SDSS DR12 (Beck et al. 2016b). The DR12 spectroscopic catalogue was extended with cross-matched redshift data from other surveys to build a sizable training set, and I adopted an empirical approach, LLR, to utilize it. I also followed up with a minimum chi-square spectral template fitting step after the photo-z estimation itself to yield more physical information. Photometric redshifts were provided for 208 million galaxies. To create the catalogue, I significantly improved and extended the previous version of the C++ estimation code.

5. To supplement the SDSS DR12 photometric redshift catalogue, I introduced photometric error classes and published a redshift error map
(Beck et al., 2016b). These additions help users in selecting galaxies with reliable photo-z accuracy estimates, and in filtering out data that do not meet a given quality standard because of large photometric errors or inadequate training set coverage.

6. To address some technical limitations of existing public solutions, I developed and published a new spectral template fitting photometric redshift estimation implementation, Photo-z-SQL (Beck et al., 2017a). The code was developed in C#, fully supports parallelism and variable photometric filter sets, and can be directly integrated into a database server running Microsoft SQL Server, bringing photo-z computations straight to the data. These features could prove to be useful to upcoming sky surveys dealing with petabytes of data.

7. The Photo-z-SQL code was thoroughly evaluated using two recent public photometric redshift testing datasets (Beck et al., 2017a). I adopted successful approaches from the literature including the Bayesian photo-z method, public galaxy template sets and priors. Also, I added refinements such as iterative filter zero point calibration based on a training set, and an extra photometric error term representing template mismatch. Accordingly, the estimation results were comparable to those of the better-performing codes in the literature.

8. I assembled a photo-z testing dataset using data from SDSS DR12 and other spectroscopic surveys to replicate the photometric error
distribution differences between spectroscopic and photometric samples (Beck et al., 2017b). Also, I provided the initial data for another dataset that focuses on testing extrapolation capability. The datasets were published to allow a comparison of photo-z methods, and facilitate research on accurate photo-z validation.

9. I evaluated the performance of the empirical LLR method (statement 4.) and the template-based Photo-z-SQL code (statement 6.) on the two datasets described in statement 8. (Beck et al., 2017b). The approaches performed reasonably well in their respective categories compared to competitors in both tests, i.e. regarding extrapolation capability and handling of different photometric error distributions.

Publications supporting the thesis statements


Additional publications


References


Csabai I. et al., 2003, AJ, 125, 580


Hildebrandt H. et al., 2010, A&A, 523, A31


