Theses of the dissertation titled

Determining Physical Properties of Galaxy Populations

and

Virtual Observatories

László Dobos
physicist MSc

Eötvös Loránd University – Faculty of Natural Sciences
Doctoral School of Physics – Particle physics and astronomy program
school and program leader: Prof. Dr. Ferenc Csikor

advisor: Prof. Dr. István Csabai

Eötvös Loránd University – Faculty of Natural Sciences
Department of physics of complex systems

Budapest, 2011.

Introduction

Modern day cosmological research is based on large volume sky surveys. Observations of even larger volumes and even more wavelength ranges are necessary to precisely determine the cosmological parameters and investigate the evolution of galaxies. Large samples allow extensive statistical analysis of the data, consequently yield better estimates of the physical parameters. On the other hand, processing large data volumes is computationally challenging. Today, taking up these challenges is part of astronomical research, just like solving other kinds of technical difficulties related to observations. For this, this thesis consists of three parts, out of which two are strictly about astronomy and the third addresses the frontiers between astronomy (and other basic disciplines) and computer science.

In Part 1 I review the basics of extragalactic spectroscopy, including observational instruments, observational techniques and the reduction of spectra. I introduce some basic methods of evaluation of spectra, including stellar population synthesis models and the systems of absorption indices. I briefly touch the subject of dimensionality reduction techniques, which are unavoidable to deal with high-dimensional data like spectra. Finally, I write about some aspects and methods of spectroscopic classification of galaxies. Concerning my own research, I introduce the robust principal component code that I implemented and present the atlas of composite spectra that I compiled using this PCA code.

Part 2 is about the extreme value and order statistics of luminous red galaxies. I review the basic methods of order statistics and show its efficiency by applying it to the luminosity function of the luminous red galaxies. I discuss the subject of brightest cluster galaxies which is one of the key topics of galaxy cluster research.

Part 3 is about the novel techniques of handling astronomical (or any scientific) data and about my research in this field. Although this part poses mostly computational questions, it is still tightly connected to astronomy, since the newest sky surveys supply data in such volumes, that familiarity with the techniques of data intensive research is indispensable.

1 Determining physical parameters of galaxy populations

In order to get a appropriately detailed and comprehensive view of galaxies, and of the dark matter, stars and gas that they consist of, spectroscopic observations have to be performed. To determine the exact redshift of galaxies – which are fundamental for investigating the large-scale structure of the universe – using spectroscopy is inevitable. Since investigating the properties of large-scale structure inherently requires the recording of numerous spectra, modern optical band sky surveys – like the Sloan Digital Sky Survey – focus on the collection spectroscopic data of such quality, that, beyond determining redshifts, can be used to analyze other physical properties of the galaxies as well. The about half a million galaxy spectra observed by SDSS can not only be used to investigate individual galaxies, but also to analyze the properties of complete galaxy populations statistically. By averaging the spectra that are individually relatively noisy on can derive high resolution composite spectra of such high signal-to-noise ratio which would be the equivalent of thousands of hours of observation.

The observable optical spectra of galaxies are determined by three basic components: the *stellar continuum*, which is characterized mainly by absorption lines but essentially a thermal radiation; the *discrete emission lines*, which come from the interstellar gas excited by the radiation from nearby hot stars or incidentally by the active galactic nucleus; and the interstellar dust, which is around 100 K and is responsible for the near-UV and optical absorption, as well as for the near-IR thermal emission. By modelling the contribution of these components to the observable spectra of galaxies on can get detailed information about the age and metallicity of the stars of the galaxies, about the mass–luminousity ratio, the star formation rate, the amount of dust within the galaxies and, by taking the emission line ratios into account, one can determine the primary source of the radiation ionizing the interstellar gas. Based on the latter, galaxies can be classified as star-forming or galaxies with active nuclei.

Our results in the field of spectroscopic analysis of galaxies:

1.1 The author of the dissertation implemented the robust, iterative principal component algorithm published by Budavári et al. (2009). We adapted the algorithm to the application to SDSS spectra, paying special attention to the initialization of the algorithm when a significant amount of the spectra are gappy. |2,6|

1.2 We compiled a high-resolution, high signal-to-noise ratio composite spectrum atlas of SDSS galaxies. Galaxies were classified based on color indices, star-forming and nuclear activity. The composites for every galaxy class were calculated by using the robust PCA

algorithm. The composites of the atlas are primarily intended to be used for template-based photometric redshift estimator algorithms. |6|

1.3 The composites determined by averaging are considered physical in the sense that they are derived as non-negative linear combinations of the spectra of real stellar populations. Hence, it makes sense to determine the physical parameters of the composites. By applying widely-used methods of the field, we determined the most important spectral indices of the composites, which can be used to infer the physical parameters. We showed that these indices overlap with the distributions of the same indices of individual galaxies, and that these indices cover much of the parameter space spanned by the individual galaxies. Hence, we proved that the atlas suitably represents all common galaxy types. |6|

2 Statistics of luminous red galaxies

The SDSS paid special attention to the spectroscopic observation of luminous red galaxies (LRGs). These galaxies are the most massive objects of the universe and are excellent proxies of the matter distribution (thus, the large-scale structure). The spectral features of LRGS (especially the strong break at 4000 Å) make their spectroscopic redshifts easily measurable to very faint magnitudes. The sample available to us is complete to redshift $z = 0.38$, but the most luminous objects were observed up to $z > 0.5$. In such a broad redshift range, the evolution of the stars constituting the galaxies is significant. According to the accepted models, these massive galaxies were formed around $z \geq 2$, almost at one time, and since than their stellar populations have been evolving passively without considerable star formation. Using an appropriate evolutionary model, one can determine the K and evolution corrected absolute magnitudes of these galaxies, based on which – by also correcting for the slow growth of their co-moving number density – a universal luminosity function can be determined which is valid in the whole $0 \leq z \leq 0{,}5$ redshift range.

The most massive elliptic galaxies are in the centers of galaxy clusters. It has been an open question whether the brightest cluster galaxies (BCGs) can be considered as extremes of the same luminosity function which describes the magnitude distribution of the rest of the red galaxies constituting the clusters, or BCGs belong to a separate population. One important argument in favor of the latter is that, looking from cluster to cluster, a magnitude gap M_{12} is measurable between the BCG and the second brightest galaxy amounting to $M_{12} \simeq 0.8$ mag on average. Such a large gap cannot be explained by taking the magnitude gap between BCGs and satellites of random samples generated based on a common Schechter luminosity function. The gap in this case is only $M_{12} \simeq 0.2$ mag.

Our results in the field of the statistics of luminous red galaxies:

2.1 Based on one of the newest spectral evolution model, we determined the universal luminosity function of LRGs and analysed it using the methods of extreme-value and order statistics. By dividing the sample into redshift shells, we were looking for the relationship between the magnitudes of the k^{th} brightest galaxies and the co-moving number densities of galaxies of the redshift shells. We showed that the expectation value of the magnitude of the k^{th} brightest galaxy can be estimated from the galaxy density, and – by reversing the method – the density of the galaxies can be estimated pretty well based on solely the magnitude of the brightest galaxies. |5|

2.2 Using order statistics, we analyzed how good standard candles the k^{th} brightest galaxies of the redshift shells are. We concluded that the variance of the magnitude of the k^{th} brightest galaxy decreases fast with the order k, consequently the less luminous (but still well observable) galaxies are much better standard candles than the BCGs. |5|

2.3 The methods of order statistics are very well applicable to the luminosity function of LRGs if the galaxies are binned by redshift. What can we say about the luminosity distribution of the galaxies of individual clusters? We showed that the measured size of the magnitude gap M_{12} can be got from the universal Schechter distribution once we don't consider simply the expectation value of the brightest magnitude, but also impose the condition that the sample must contain an extremely bright galaxy. |5|

3 Virtual Observatories

The biggest astrophysical observations and simulation projects of the last decade yielded data on such a large scale which fundamentally reformed the methods of processing, storage, distribution, analysis and visualization of scientific data. Beside the three former main scientific paradigms – experiment, theory and simulation – a new method emerged, the data-driven research. Alexander Szalay and his research group at the Johns Hopkins University in Baltimore reformed the handling of astrophysical data by building a database and a data warehouse from the measurement of SDSS that is accessible by anyone via the Internet. By developing on the database ideas and tools of the SDSS, the Virtual Observatory was born, which has become a wide international collaboration by today.

The main idea of the Virtual Observatory is that the data warehouses of the big research centers housing the data collected by the multimillion dollar projects should be designed such a way that the data become available to anyone via the Internet using standardized protocols and data formats. By going one step further: data warehouses should be interconnected and be prepared to cooperate automatically to help the work of researches using data from multiple sources. Projects like these can significantly reduce the time

spent on data-handling issues of those researchers who work with data from multiple instruments, so they can concentrate on scientific questions instead of dealing with the computer related problems.

At the beginnings of modern astronomy, journals were sufficient to publish data. These data are still easily available today, either in the original printed format, either electronically. It is a big question, what will be the fate of the large databases that are only available in electronic format. The American National Science Foundation requires all supported researchers to make all collected data publicly available after the projects have finished. The big data warehouses, beside storing and serving raw data, should support the processing of the data as well. Since the bandwidth of networks does not grow according to the amounting data, it is absolutely necessary to move the computations to the data and not the data to the processors doing the calculations.

Data-intensive research requires a general knowledge about the actual scientific field and experience with computer technologies addressing the data handling problems at the same time. While a hundred years ago it was inventing new mathematics that counted as serious results in theoretical physics, today developing novel data handling techniques plays the same role. Although the academic institutes of the world do not consider *e-science* an individual discipline yet, together with my advisor and colleagues we believe that it will reach its rightful status in a couple of decades and will become a multidisciplinary field of science on its own right. E-science will be indispensable for the every day work of physicists and astronomers.

Our results in the field of Virtual Observatory:

3.1 The author of the dissertation developed the Spectrum Services for the Virtual Observatory, which organizes the spectra of SDSS into a database and makes the data searchable through a web interface and a web service. The program offers numerous spectrum processing functions which not only makes preprocessing spectra easier but are usable for scientific analysis. |7,8,12,13,19|

3.2 We developed a method to answer similarity queries on databases of galaxy spectra. The method is based on two important components: First, we determine a principal component basis, on which we expand all galaxy spectra and organize the expansion coefficients in a data table. Second, we index the data table of the principal component coefficients with a special spatial index which supports fining the nearest neighboring data points of a query point. When searching for similar spectra to a given spectrum, the following is done: The query spectrum is expressed on principal component basis to get the eigencoefficients. Using the spatial index, a few data points being the closest to the eigen-

coefficients of the query spectrum are looked up. These spectra will all be very similar to the query spectrum. |1,10,13|

3.3 When cross-matching different extragalactic catalogs, it is very important to know the exact sky coverage (the so called footprint) of the surveys. The geometry of the sky coverage of certain instruments can be fairly complex, whereas their exact description is absolutely necessary for various services and software tools of the Virtual Observatory. We implemented the Footprint Service which organizes the footprints of the most widely used catalogs. Footprints can be queried via a web interface and a web service. Users can draw and upload the sky coverage of their own observations as well. |9,18|

3.4 Automatic and fast cross-matching of survey catalogs containing observations made at different wavelengths would be an important service of Virtual Observatory, but the solution to this problem is hard. Based on the Bayesian statistical method developed by Budavári & Szalay (2008), we developed the prototype of a cross-match system which uses database server clusters. With our solution, the cross-match problems which would require manual work and would have taken weeks earlier can be done automatically in a couple of minutes. |15|

3.5 A big deficiency of relational database systems used to store scientific data is their lack of support to natively handle array-based data. Once we want to organize the results of cosmological simulations in databases, it is necessary to extend the database servers with program modules capable of handling arrays. For this purpose, the author of the dissertation developed an extension to Microsoft SQL Server 2008 which allows handling of multidimensional data structures. Beside simply storing multidimensional data, several basic array handling functions are also implemented. Also, widely-used mathematical libraries (e.g. LAPACK, FFTW etc.) were made callable directly from the database management system. |16, 17|

Papers in refereed international journals

1. Csabai, I., Dobos, L., Trencséni, M., Herczegh, G., Józsa, P., Purger, N., Budavári, T. and Szalay, A. S.: „Multidimensional Indexing Tools for The Virtual Observatory", 2007, Astronomische Nachrichten, **328**, *852*

2. Budavári, T., Wild, V., Szalay, A. S., Dobos, L. and Yip, C.-W.: „Reliable Eigenspectra for New Generation Surveys", 2009, MNRAS, **394**, *1496-1502*

3. Yip, C. W., Connolly, A. J., Vanden Berk, D. E., Scranton, R., Krughoff, S., Szalay, A. S., Dobos, L., Tremonti, C., Taghizadeh-Popp, M., Budavári, T., Csabai, I., Wyse, R. F. G., Ivezić, Ž.: „Probing Spectroscopic Variability of Galaxies and Narrow-Line Active Galactic Nuclei in the Sloan Digital Sky Survey", 2009, AJ, **137**, *5120-5133*

4. Yip, C.-W., Szalay, A. S., Wyse, R. F. G., Dobos, L., Budavári, T. and Csabai, I.:„Extinction in Star-forming Disk Galaxies from Inclination-dependent Composite Spectra", 2010, ApJ, **709**, *780–790*

5. Dobos, L. and Csabai, I.: „Order Statistics of the Early-Type Galaxy Luminosity Function", 2011, MNRAS, **414**, *1862-1874*

6. Dobos, L., Csabai, I., Yip, C-W., Budavári, T., Wild, V. and Szalay, A. S.: „A High Resolution Atlas of Composite SDSS Galaxy Spectra", 2011, MNRAS submitted

Papers in proceedings of international conferences

7. Dobos, L., Budavári, T., Csabai, I. and Szalay, A. S.: „Spectrum and Bandpass Services for the Virtual Observatory", Astronomical Data Analysis Software and Systems (ADASS) XIII 2004, Astronomical Society of the Pacific Conference Series, **314**, *185*

8. Dobos, L., Budavári, T., Csabai, I. and Szalay, A. S., „New Features in the Spectrum Services for the Virtual Observatory", Astronomical Data Analysis Software and Systems (ADASS) XV 2006, Astronomical Society of the Pacific Conference Series, **351**, *471*

9. Budavári, T., Dobos, L., Szalay, A. S., Greene, G., Gray, J. and Rots, A. H.: „Footprint Services for Everyone", Astronomical Data Analysis Software and Systems (ADASS) XVI 2007, Astronomical Society of the Pacific Conference Series, **376**, *559*

10. Dobos, L., Csabai, I., Trencséni, M., Herczegh, G., Józsa, P. and Purger, N.: „Spatial Indexing and Visualization of Large Multi-Dimensional Databases", Astronomical Data Analysis Software and Systems (ADASS) XVI 2007, Astronomical Society of the Pacific Conference Series, **376**, *629*

11. Mátray, P., Csabai, I., Hága, P., Stéger, J., Dobos, L. and Vattay, G., „Building a Prototype for Network Measurement Virtual Observatory", 2007, Proceedings of the 3rd Annual ACM Workshop on Mining Network Data, MineNet 2007, San Diego, California, USA, *23–28*

12. Dobos, L., Budavári, T., Csabai, I., Szalay, A. S.: „Spectrum Services 2007" in „Astronomical Spectroscopy and Virtual Observatory", 2008, Proceedings of the Euro-VO Workshop, *79*

13. Dobos, L., Budavári, T., Csabai, I., Szalay, A. S. and Herczegh, G.: „Improved Search in Spectrum Databases", Astronomical Data Analysis Software and Systems (ADASS) XVII 2008, Astronomical Society of the Pacific Conference Series, **394**, *389*

14. Dobos, L., Csabai, I., Budavári, T. and Szalay, A. S.: „Building a Database from the SDSS Imaging Data", Astronomical Data Analysis Software and Systems (ADASS) XVIII 2009, Astronomical Society of the Pacific Conference Series, **411**, *366*

15. Simmhan, Y., Barga, R. S., van Ingen, C., Nieto-Santisteban, M. A., Dobos, L., Li, N., Shipway, M. P., Szalay, A. A., Werner, S., Heasley, J.: „GrayWulf: Scalable Software Architecture for Data Intensive Computing", 2009, Proceedings of the 42st Hawaii International International Conference on Systems Science (HICSS-42), *1-10*

16. Dobos, L., Szalay, A. S., Blakeley, J.A., Budavári, T., Csabai, I., Tomic, D., Milovanovic, M., Tintor, M., Jovanovic, A.: „Array Requirements for Scientific Applications and an Implementation for Microsoft SQL Server", 2011, Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases

17. Falck, B., Budavári, T., Cole, S., Crankshaw, D., Dobos, L., Lemson, G., Neyrinck, M., Szalay, A. and Wang, J.: „The Indra Simulation Database", 2011, American Astronomical Society Meeting Abstracts **#218**, *131.04*

Book chapters

1. Budavári, T., Szalay, A. S., Fekete, G., Dobos, L., Greene, G., Gray, J. and Rots, A. H.: „Chapter 9: Web-based Tools – Footprint Services in the Virtual Observatory" in „The National Virtual Observatory: Tools and Techniques for Astronomical Research", 2007, Astronomical Society of the Pacific Conference Series, **382**, *75*

2. Dobos, L. and Budavári, T.: „Chapter 17: Web-based Tools – Spectrum and Filter Services for the VO" in „The National Virtual Observatory: Tools and Techniques for Astronomical Research", 2007, Astronomical Society of the Pacific Conference Series, **382**, *147*