Design and implementation of the spectra reduction and analysis software for LAMOST telescope

A-Li Luo[*a], Yan-Xia Zhang[a] and Yong-Heng Zhao[a]

[*a]National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Road, Chaoyang District, Beijing, 100012, China.

ABSTRACT

The Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) will be set up and tested. A fully automated software system for reducing and analyzing the spectra has to be developed before the telescope finished. Requirement analysis has been made and data model has been designed. The software design outline is given in this paper, including data design, architectural and component design and user interface design, as well as the database for this system. This paper also shows an example of algorithm, PCAZ, for redshift determination.

Keywords: LAMOST, data model, architecture, interface, PCAZ

1. INTRODUCTION

The Large Sky-Area Multi-Object Spectroscopic Telescope (LAMOST) has been constructing by the National Astronomical Observatories for several years, and has been planed to set up in coming years. Four thousand fibers will be employed to connect the focus plane to thirty two spectrographs[*], and will yield up four thousand multi-fibre spectra of QSOs, galaxies and stars per field. The LAMOST spectroscopic survey will target over ten million objects chosen from the SDSS photometric survey, DSS-II and other catalogues such as FIRST and ROSAT. Many targets will be selected on the basis of cross-identification between more than one of the above catalogues. The paper describes a fully automated software for data reduction and analysis. It is designed in order to cope with the anticipated flood of spectrographic data.

The software will produce measurements of spectral lines (such as position, equivalent width (EW) or intensity), and computed parameters (such as redshift or velocity). These information can be used for various studies including stars, galaxies, AGNs and cosmology. For example, we can research stellar components of galaxies, star formation rates of galaxies, star formation histories of galaxies, metallicities and dust contents of galaxies, kinematics of our Galaxy, AGN classification, and large scale structure of cosmology. More and more scientific goals and applications of spectral parameters are suggested recently. In addition, the pipeline will also classify objects into different types, by which we can archive the data reasonable and find rare celestial bodies from those unclassified spectra. Thus, the two main tasks of our pipeline are measurement and classification of spectra.

Raw data of CCD spectral image will be feeded to 2-d reduction pipeline, and reduced by traditional scheme to extract 1-d spectra. When 1-d spectra obtained, they will be stored into a hard disk array managed by a storage management server. Then sixteen high performance PC working in Linux environment will read the data files, and process them to obtain calibrated spectra and parameters. Then, each calibrated spectrum will be written to the hard disk array in forms of FITs, and parameters will be stored in the database of LAMOST.

According to software engineering, requirement analysis has been made before designing the software, which also includes data model, functional model and behavioral model. The main scheme of this system and profile of LAMOST data model are shown in this paper; parts of the major spectroscopic attributes are listed; the records and size of each catalogue in the data sets are estimated; principles and rules of designing the software are narrated, especially users interface. An example of 1-d spectral processing algorithm is also be given.

[*]contact: lal@lamost.bao.ac.cn; phone 8610-64841693; fax 8610-64878240

[*]Sixteen spectrographs are for red part of spectra, and others are for blue part. Thus each fibre is connected to two spectrographs, and two hundred and fifty fibres are connected to spectrograph.

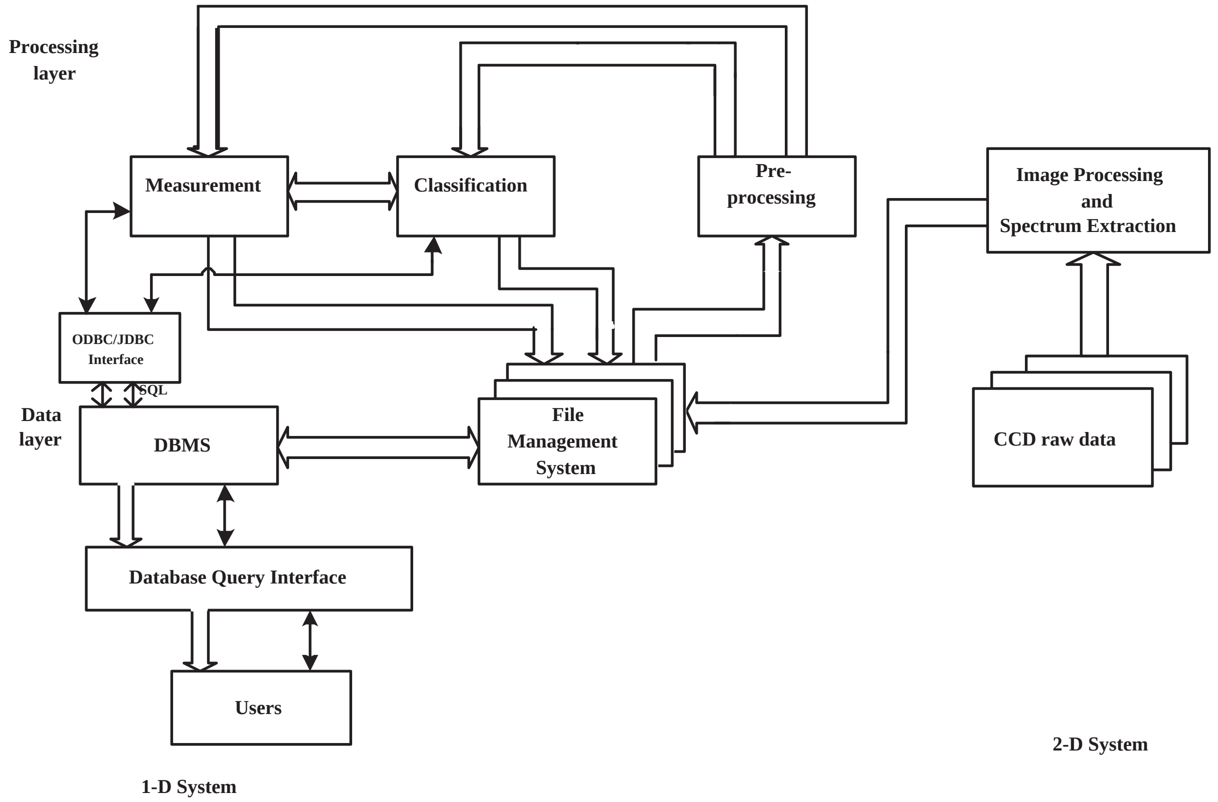

Figure 1. The Main Diagram of LAMOST data analysis software

2. REQUIREMENT ANALYSIS

Goals and objectives of LAMOST spectra reduction and analysis software are to acquire parameters of spectra, and classify those spectra. The input of the software is raw data of CCD spectral image, will be feeded to 2-d reduction pipeline to do traditional reduction to extract 1-d spectra. The output of the 2-d pipeline will be extracted 1-d spectra as FITs files, and they will be stored into hard disk array managed by a storage management server. The pre-processing procedure will read 1-d spectra through standard I/O, and subtract continuum and noise to pick out absorption and/or emission lines. Through the measuring block and the classification block, the output of the software will be calibrated spectra and their parameters. Each calibrated spectrum will be written to the hard disk array as FITs files, and the parameters will be write into LAMOST database through ODBC/JDBC interface. The main scheme is shown as Figure 1.

2.1. Data Model and Description

The software will be based on Object-Oriented technique, and this subsection describes classes for each objects and their major attributes. The classes are linked to each other either by inheritance or by association. Inheritance is a way to specialize a class. Every attribute present in the base class is also present in the inherited class, but the inherited class may have additional attributes. Associations link classes to one another, which are grouped into three forms: one-to-one, one-to-many and many-to-many associations.

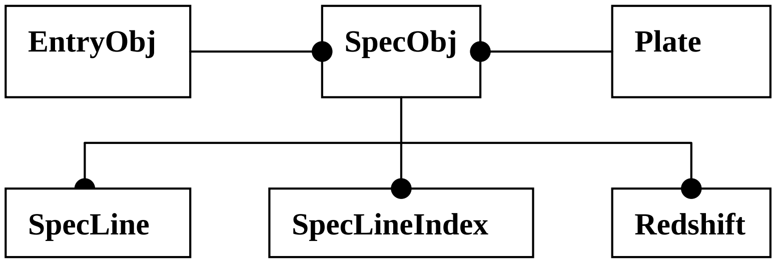

Figure 2. LAMOST Spectroscopic data model

The data of spectra are taken on spectroscopic plates, to each of which we have a corresponding plate object. The spectral objects (named SpecObj) correspond to each spectroscopic survey object, and they have associations to the found and identified spectral lines as well as the redshift measured. The data model is shown in Figure 2.

For each spectroscopic object (in the SpecObj class) there are links to the identified and measured spectroscopic lines, as well as redshifts. SpecObj is the measured parameters for a spectrum stored in this class; SpecLine is found and measured spectral lines; Redshifts and errors are measured using cross correlations and measured lines; SpecLineName is the name of each spectral line; SpecLineIndex is indexed quantities for fast searching of SpecObj. Table 1 lists parts of the spectroscopic attributes in the SpecObj class.

2.2. Functional and Behavioral Model

This subsection describes three major functions of this software, along with interface, data flow and the behavior of the software.

Description of Function of Pre-Processing(PP): This function is a preparation for measuring and classifying the spectra. The input of this functional block is 1-d original spectra, and the output is spectra with continuum subtracted. The input interface for this function is standard I/O, since each spectrum is a FITs file stored in a file system. The output of this block is connected to measurement block and classification block, and this two blocks use directly the spectra in a data array, which are stored in memory.

Description of Function of measurement: This function is designed to measure the spectra and get spectral line parameters including position, strength, equivalent width, and redishifts of spectra. This functional block will call the spectra in data array which will have been preprocessed by PP functional block. The output of this function will be calibrated spectra and parameters. The parameters will be written into database, so the interface between measurement block and the database is ODBC/JDBC; The calibrated spectra will be passed both into a file system and classification block. The interface between measurement block and the file system is standard I/O, and spectral data will be passed to classification block in form of data array.

Description for Function of classification: This function is designed to classify the spectra into different types according to methods of matching templates and principal component analysis (PCA). This functional block will call the spectra in a data array which will have been preprocessed by both PP functional block and measurement block. It will also read information of measurement from the database through ODBC/JDBC. The output of this function will be classification information of spectra, and the interface between classification block and the database is also ODBC/JDBC.

The software interface to users is the database query interface; the interfaces between computers are standard network (TCP/IP protocol); the whole processing is automated.

In addition to functional model, the behavioral model of the software is also needed to be designed, which includes major events and states. The events (control, items) will cause behavioral change within the system, while states (modes of behavior) will result as a consequence of events. In order to design them easily, we need

Table 1. Parts of the spectroscopic attributes in the SpecObj class

Member Name	Datatype	Description
SpecObj:		
specID	int32	Unique ID
xPlane	float32	X Location of Fiber
yPlane	float32	Y Location of Fiber
RA()	float32	J2000 Right Ascension
DEC()	float32	J2000 Declination
fiberID	int16	Fiber ID
z	float32	Redshifts for galaxies
Velocity	float32	Velocities for stars
specClass	int32	Spectral Classification
PCAClass	float32	Galaxy Classification
sn	float32	Median S/N
mag	float32	Magnitude
found	ManyAssoc	Link to found spectral lines
identify	ManyAssoc	Link to measured spectral lines
lineIndex	ManyAssoc	Link to spectral line indices
SpecLine:		
speclineID	int32	Unique ID
wave	float32	Line Center (Angstroms)
sigma	float32	Sigma of fitted gaussian (Angstroms)
height	float32	Height of gaussian
ew	float32	Equivalent width (Angstroms)
LineFlag	int16	Line found (0) or identified (1)
specobj	OneAssoc	Link to spectrum object
linename	OneAssoc	Link to line name

a state transition diagrams which depict the overall behavior of the system. Here we do not show it because of paper's length.

2.3. Product

The LAMOST data archive will be distributed in two main forms: a spectroscopic catalogue and a set of individual spectra. The former will contain positions, information related to the observations, other measured parameters such as redshifts (or radial velocities), line intensities (or equivalent widths) and positions of identified emission and absorption lines etc. The latter will comprise of one-dimensional spectra for one million quasars, ten million galaxies and one million stars. Catalogue subsets may also be included, and the expected sizes of all the data sets are listed in Table 2.

3. SOFTWARE DESIGN OUTLINE

This section provides an overview of the entire design including architecture, interface and component-level design for the software. Overall goals and software objectives have been discussed in this section, as well as major inputs, processing functionality, and outputs. In this section, we gives the outline of the software design.

3.1. Data design

All data structures, such as internal, global, and temporary data structures are designed firstly.
1. Designing internal software data structure that are passed among components.

Table 2. LAMOST data sets and their expected sizes

product	records	size
Spectroscopic catalogue:		
Raw observational data	-	40TB
Redshift catalog	10^7	20GB
Radial velocity catalogue	10^6	2GB
Observation log and file headers	10^6	10GB
Simplified catalogue	4×10^8	80GB
Individual spectra:		
1D spectra	10^7	1TB

2. Designing global data structure that are available to major portions of the architecture.
3. Designing temporary data structure for interim use.
4. Designing database.

3.2. Architectural and component-level design

1. Program structure design needs an architecture diagram, which is a pictorial representation of the architecture.
2. Design of each software component contained within the architecture, including components, interfaces, algorithm, restrictions/limitations and local data structures etc.
3. Design of the software's interface(s) to the outside world, including interfaces to other computers, to other systems, products and/or networks.

3.3. User interface design

There are two different user interfaces for this software: one for internal user and one for normal user. The internal user interface is designed for the LAMOST team, which will be written in Python. The normal user interface is web-based, and users can easily access the LAMOST database through a friendly user's interface. The interface will be designed in XML, and available to users by simple browsers (such as IE, Marzilla etc.). It will support two main kinds of query.

Firstly, there will be a simple search tool enabling one to retrieve data subsets based on search limits chosen by the user. For example, the user will be able to search for all objects lying within a specified field on the sky, by entering positional limits interactively. The interface will also permit interactive "advanced" searches as well as interactive "refined" searches of data subsets.

The second kind of query will give users the option of supplying their search criteria in standard SQL[1](a widely used database language). Users will work using Views[†] rather than with heavily-indexed base tables. To speed access, indices are helpful to manage those most frequently accessed attributes, and an SQL query will automatically use those indices covering the most important attributes. Aided by different indices, users will be able to retrieve observational information as well as spectral parameters from LAMOST database.

4. AN EXAMPLE OF ALGORITHM

In general, there are two ways of the determining redshift of galaxies. One method is to measure spectral lines and identify them, then calculate the wavelength shift of these lines. The other method is to match observational spectra with templates. Luo et al.[2–5] applied the wavelet technique to find the spectra line, dividing the galaxy spetra into emission-line spectra and non-emission-line spectra. For non-emission-line spectra, they identified

[†] View and index are concepts of database.

automatically the spectral lines by 4000Å break, and then determined the spectra redshifts. While for those spectra that are difficult to identify, we have to use the method of cross correlation to obtain redshifts.

Glazebrook et al.[6] suggested a cross-correlation based method for redshift measurement, named "PCAZ" after Principal component analysis (PCA), which is a widly used technique in feature extraction, data compression and classification.[7, ?–9] PCA is also used for spectral classification of large survey, such as Las Campañas Redshift Survey, the two degree Field (2dF) Redshift Survey and the Sloan Digital Sky Survey (SDSS). But all this classification is limited to spectra with redshift reduced. Connolly et al.[10] gave the systematic error arising from the influence of redshift on PCA classification, and Glazebrook was enlightened to put forward a method to determine redshift using PCA.

4.1. Algorithm and Improvement

"PCAZ" generalizes the cross-correlation approach by replacing the individual templates with a simultaneous linear combination of orthogonal templates. The standard fast Fourier transform (FFT) method can be used to speed the computation. With flexible linear combination of eigen templates, linear combination of orthoganal template is easy to match observed data. The requirement of PCAZ is that the standard spectra templates is enough to cover all types of spectra.

Details and equations about PCAZ can be referred to the paper of Glazebrook et al..[6] The method can not be applied directly to LAMOST spectra, for whose spectra are observed in red-band spectrograph and in blue-band spectrograph, respectively. Though combining the spectra from two bands by such methods as wavelet, it can not avoid the seam in spectra, which will possibly leads to the addition of false emission-lines or absorption-lines. Thus, mistakes will occur in cross correlation. In addition, not all spectra can be linearly represented by templates. If applied to the whole wavelength, some problems would arise. Because blue band spectra of galaxy are contributed mainly by hot stars, and red band are mainly by cold stars, the spectra of different bands possibly match different templates. Therefore it is necessary to process the spectra of different bands separately to avoid this problem.

For PCAZ method, the wavelength range of templates is the same as that of observed spectra. When the templates shift in wavelength, the overlapped wavelength coverage used for computing likelihood function will change. As the redshift is larger, the result will not be accurate. If we shorten the template range while the wavelength range becomes narrow, this not only loses the spectral line information of templates, but also the spectra of large redshift can not be measured. In addition, the accuracy of measurement turns low.

Considering problems above, we improved PCAZ as follow:
1.We construct the spectra templates of red and blue band separately. Thus we needn't combine the spectra from red and blue band to avoid the error caused from combination. The spectral redshifts of red and blue bands are computed separately and compared, and taken as the candidates of final redshift.
2. We make the wavelength range of templates as long as possible, meanwhile, the blue wavelength of templates is shorter than that of observed spectra in order to keep the information of templates as much as possible and measure the spectra of large redshift.

Starting from the blue band of templates, we take the wavelength range of template spectra similar to those of the blue band and the red band section by section (for example, 5 points as a step for each section), and orthogonalize them by PCA, respectively. For blue band and red band of any observed spectrum, we apply PCAZ to them and get two sets of redshift values, corresponding to the redshift measurement of the two bands, respectively. Comparing the two sets of redshift values, the most approximate value among two band will be the true redshift. Adopting the logarithmic coordinate of wavelength, the redshift is given by

$$z = 10^{r\times(step\times j+\delta)} - 1 \tag{1}$$

where r is the logarithmic coordinate resolution of observed spectra, usually 0.0001. j is the number code of template matching the observed spectra. The step of section spectra is represented as *step*. δ is the value got by PCAZ. z is the final computed redshift value.

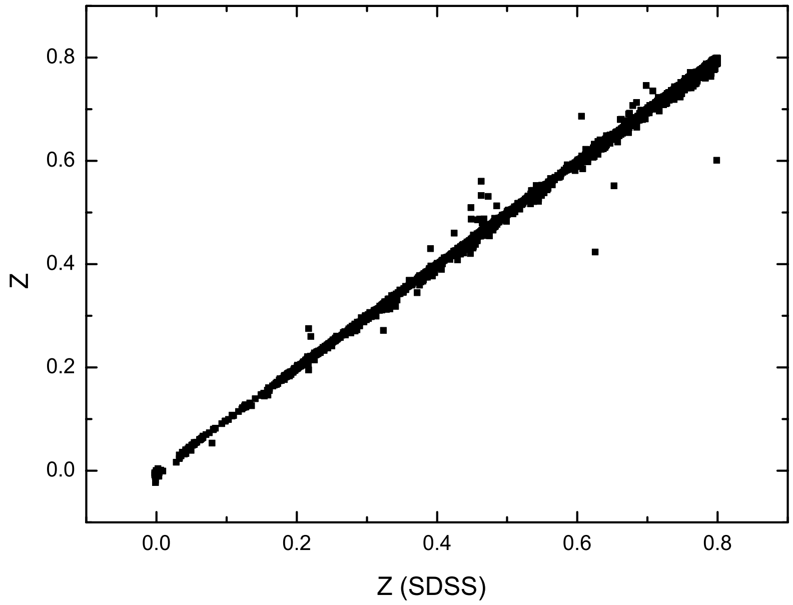

Figure 3. Example of redshift measurement

4.2. Data

The elemental templates are adopted from Kinney et al. (1996), which include the spectra atlas of quiescent and starburst galaxies (1 spectrum of elliptical galaxy, 4 spectra of spiral galaxies and 6 spectra of starburst galaxies). The galaxy spectra of the atlas of Kinney et al. (1996) cover the ultraviolet to near-infrared spectral range from 1100 up to 12,000Å. The quiescent galaxies were grouped according to morphological type: E, S0, Sa, Sb and Sc, respectively. The spectra of starburst galaxies were grouped according to increasing values of the color excess $E(B-V)$: from SB1, with $E(B-V) < 0.1$ to SB6, with $0.61 < E(B-V) < 0.7$. The eleven templates of galaxy spectra were used to construct orthogonal templates.

For the shapes of continuum heavily affect the template matching procedure, we need subtract the continuum from both template spectra and observed spectra before data processing. The method to solve the problem was described by Luo & Zhao.[3] The released data of SDSS DR1 are very similar to the future observed data of LAMOST. So we choose the data to test our method. From the data of thirty sky area, the high S/N galaxy spectra are picked out and the number of these galaxy is 4006. The galaxy redshift of the data covers the range from 0 to 0.8. The part of galaxy spectra occupies 99.8% in all the spectra.

4.3. Results

Based on the templates given by Kinney et al. (1996), we apply principal component analysis on the eleven templates to get orthogonal templates. From the orthogonal templates of blue band and red band, it is clear that they are divided into two classes: normal template spectra and non-normal template spectra. The former are characterized by absorption lines, while the later are typical of emission lines. Our goal is to determine the galaxy redshift of LAMOST. The galaxy spectra of SDSS DR1 cover the redshift range from 0 to 0.8, which is consistent with the data of LAMOST. Therefore we test our approach with the data of SDSS DR1. The computed values of redshift are compared with those of redshift given by SDSS. As shown in Figure 3, most of data points lie on the diagonal line. Only a few deviate the line, which may possibly arise from the uncomplete templates. The result of dealing with 4006 spectra shows that the accuracy adds up to 97%, and standard error σ is 0.0058. With such high accuracy and so small standard error, the method proves to be very reasonable and applicable to measure the redshift of the high S/N spectra.

4.4. Discussion

Tonry & Davis proposed the cross-correlation approach to automatedly determine the spectra redshift.[11] In the method, a series of templates consisting of different types of galactic spectra, individually tested, is not necessarily the optimal template set to use. Glazebrook et al. generalizes the cross-correlation approach as PCAZ by replacing the individual templates with a simultaneous linear combination of orthogonal templates.[6] Based on PCAZ and according to the spectral characterization of LAMOST , we developed our method to measure redshift automatically for LAMOST spectra.

The method proposed by us has several advantages as follows:

1: The method keeps the information of template as much as possible and may be applicable to measure the larger redshift. In theory, the blue wavelength of template is smaller, the capability of redshift measurement can be larger.

2: It can self-verify the measured value of redshift. For each section of template, it gives two measured redshift values of the red band and the blue band. By comparing the two of values, the accuracy of measured redshift can be determined. If the two values are more close, the value is more approximate to the true redshift value.

3: It has rather high precision. It measures the redshift more than one time and takes the value more approximate to the true redshift as the final value. Thus it makes up for the loss caused by that the reduction of data points arises the error of Fourier transform to become large. Finally, it is reliable and robust to measure the redshift by our method.

5. SUMMARY

This paper gives more general idea of designing the LAMOST spectral analysis software. Many details can not be explained, for example, subfunctions of each function are not presented. We have finished design, definition and algorithm research, and now on the stage of coding, the first complete version will be finished in coming 4 months. In this paper, we also gives an example of redshift measurement algorithm. It is not the only one algorithm to measure redshift of a galaxy spectrum, and the system will decide the final redshift measured by this algorithm or other algorithms and write it in LAMOST database as an attribute of that spectrum.

ACKNOWLEDGMENTS

The author would like to thank Ms. Hui-Xian Li (NAOC) for much help in drawing figures in this paper.

REFERENCES

1. J. Bowman, S. L. Emerson, and M. Darnovsky., *The Practical SQL Handbook: Using SQL Variants, 4th Edition*, Addison Wesley Professional, 2001.
2. A. L. Luo and Y. H. Zhao, “Astronomical spectral lines auto-searching using wavelet technology,” *Acta Astrophys. Sinica* **20**.
3. A. L. Luo and Y. H. Zhao, “Steps towards a fully automated classification and redshift-measurement pipeline for LAMOST spectra. I. Continuum level and wavelength estimation for galaxies,” *Chinese Journal of Astronomy and Astrophysics* **1**, pp. 563–572, 2001.
4. A. L. Luo and Y. H. Zhao, “Wavelet-based automatic methods to get spectral classification information of galaxies,” *II NUOVO CIMENTO B* **116**, pp. 879–888, 2001.
5. A. L. Luo, “Pattern-recognition methods in automatic techniques for astronomical spectral analysis,” *The Publications of the Astronomical Society of the Pacific* **114**, p. 789, 2002.
6. K. Glazebrook, A. R. Offer, and K. Deeley, “Automatic redshift determination by use of principal component analysis. I. Fundamentals,” *Astrophysical Journal* **492**, p. 98, 1998.
7. Y. X. Zhang and Y. H. Zhao, “Classification in multidimensional parameter space: Methods and examples,” *The Publications of the Astronomical Society of the Pacific* **115**, pp. 1006–1018, 2003.

8. S. R. Folkes, O. Lahav, and S. J. Maddox, "An artificial neural network approach to the classification of galaxy spectra.," *Mon. Not. R. Astron. Soc.,* **283**, pp. 651–665, 1996.
9. H. P. C. F. C. B. e. a. Francis, P. J., "An objective classification scheme for qso spectra," *Astrophysical Journal* **398**, pp. 476–490, 1992.
10. A. J. Connolly and A. S. Szalay, "A robust classification of galaxy spectra: Dealing with noisy and incomplete data," *The Astronomical Journal* **117**, pp. 2052–2062, 1999.
11. J. Tonry and M. Davis, "A survey of galaxy redshifts. i - data reduction techniques," *Astrophysical Journal* **84**, pp. 1511–1525, 1979.