Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata
Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata
Ciclo XXV

Automatic classification of galaxy spectra in large redshift surveys

Settore Scientifico Disciplinare FIS/05

Supervisore interno: Professor Marco BERSANELLI

Supervisore esterno: Professor Luigi GUZZO

Coordinatore: Professor Marco BERSANELLI

Tesi di Dottorato di:
ALIDA MARCHETTI

Anno Accademico 2013-2014

Commission of the final examination:

External Referee:
Prof. Andy J. Connolly

External Member:
Prof. Alberto Franceschini

Internal Member:
Prof. Marco Bersanelli

Final examination:

Date: November 28th, 2014

Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

To Giacomo and Filippo

"...e quando miro in cielo arder le stelle, dico fra me pensando: a che tante facelle? Che fa l'aria infinita, e quel profondo Infinito seren? Che vuol dir questa Solitudine immensa? Ed io che sono?"

Canto notturno di un pastore errante dell'Asia, Giacomo Leopardi

Cover illustration:
La notte stellata, Vincent Van Gogh
MIUR subjects:
FIS/05 -

PACS:
98.62.-g

Contents

List of Figures

Introduction

The Universe is overspread by a multitude of galaxies, presenting the more different ages, overall colors and shapes. Some of them, under favorable conditions, are perceivable to the naked eye, as the central region of our own Milky Way, M33 in the Triangulus constellation, or our neighbour M31 in the Andromeda constellation in the north hemisphere, and the two Milky Way's satellite galaxies, the Magellanic Clouds, in the south hemisphere. Hundreds of galaxies are instead within the reach of a small telescope. In 1995 the Hubble Space Telescope took the picture of an impressive number of galaxies (~ 3000) by focusing on a minute (less than a 20-millionth of the entire sky) and apparently empty region of the sky in the Ursa Mayor constellation. Later on HST collected other similar and more resolved images of the deep space, revealing the existence of a multitude of galaxies and galaxy types even at very large redshifts.

This galaxy zoo can be largely divided into two classes: early type galaxies, characterized mainly by old, passively evolving stellar populations, and late type galaxies that show evidence for recent star formation. This dichotomy is displayed in the local Universe under many aspects (**Chapter 1**), for example the morphology of galaxies (Sandage, 1975): younger galaxies usually exhibit a spiral shape with tightly wound gas-rich spiral arms surrounding the central bulge region, while older ones are usually elliptical objects, suggesting the loss of the spiral arm component as a result of evolution or of close encouters (or merging) with other galaxies. The distinction is also evident in the galaxy colors (de Vaucouleurs, 1962): older galaxies, being mainly composed by very old and evolved stars, reflect the color of the prominent star population itself, and are redder than young galaxies. The latter, being gas rich, are forges of newborn and young blue stars, which confer them bluer color.

The population and gas content of the two main types of galaxies also affect their spectral characteristics (Morgan & Mayall, 1957; Madgwick et al., 2002), which decline into different continuum shapes and emission features (which are typical of young galaxies and nearly absent in the older galaxies, leaving the place to the absorption features), as well as their clustering properties (Davis & Geller, 1976; Giovanelli et al., 1986; Guzzo et al., 1997; Norberg et al., 2002; Phleps et al., 2006; Coil et al., 2006; Meneux et al., 2008, 2009; Zehavi et al., 2011): blue galaxies have a more filamentary distribution than red ones, and trace the underlying dark matter distribution with a smaller proportionality parameter, or bias. All the evidence listed here are already present at high redshifts (Brown et al., 2003; Daddi et al., 2003; Coil et al., 2008; Abbas et al., 2010; de la Torre et al., 2011; Coupon et al., 2012) and provide fundamental constraints on galaxy formation and evolution models.

Motivation

The two-class distinction, as specified, is of course a simplification: if the class distinction can be seen as the result of an evolutionary transition, many intermediate galaxy types could be identified from the youngest to the latest galaxy population. Indeed the distribution of galaxy colours can be observed to be bimodal, with two distinct peaks, one in the red and one in the blue (Strateva et al., 2001; Bell et al., 2004; Baldry, 2004; Weiner et al., 2005; Faber et al., 2007; Franzetti et al., 2007). Between these classes lie galaxies with intermediate colours, which are associated to an evolutionary phase called *green valley*: these share in fact the characteristics pertaining to both red and blue classes and are the ones thought to be caught, during the transition from a period of active star formation to quiescence (Bell et al., 2004; Baldry, 2004; Faber et al., 2007; Brammer et al., 2009).

Spectroscopy provides a deeper insight into the physics of galaxies, with respect to average colours, determined from broad-band photometry (**Chapter 2**). For example, selecting red galaxies solely on broad-band colours, does not result in a sample of dead, passive early-type objects but also contains a non-negligible fraction of star forming galaxies and/or dusty starbursts ones (Cimatti et al., 2002; Gavazzi et al., 2003; Franzetti et al., 2007; Graves et al., 2007). Conversely, the high information content of the spectroscopic data sets makes it difficult, in general, to compress and classify all the information contained in a galaxy spectrum in a compact and efficient way. Thus, it's in general very complex to classify this huge variety of galaxies, whose diversity, based not only by different stages of evolution, but also by their mass, their environment and their possible encounters, can be captured in a more complete (even if not exhausting) way by looking at their spectral characteristics. Statistical methods have been successfully used to reduce such complexity by identifying specific features, such as emission line intensities or continuum break strengths (e.g. Madgwick et al. 2003, Colless et al. 2001).

Other alternative methods have been developed to obtain fast spectroscopical classifications; for example, the Support Vector Machine (SVM), a supervised method based on kernel algorithms (Cristianini & Shawe-Taylor, 2000; Shawe-Taylor& Cristianini, 2004) able to recognize structures or patterns within the data, has been siccessfully used to perform class separation (Woźniak et al., 2004; Zhang & Zhao, 2004; Huertas-Company et al., 2008), as star-galaxy separation (Solarz, Pollo & Takeuchi, 2012) or AGN-regular galaxy separation (Małek, Solarz & Pollo, 2013).

Also neural networks, computational models capable of machine learning and pattern recongnition, have been employed for stellar classification (Gulati et al., 1994a,b; von Hippel et al., 1994) ad star/galaxy separation (Odewanh et al., 1992), also joint to Principal Component Analysis (Singh, Gulati and Gupta, 1998), and for spectral classification of galaxies (Sodré & Cuevas, 1997). Learning Vector Quantization (LVQ) was applied to the classification of astronomical objects classification (Zhang & Zhao, 2003). Bayesian Belief Networks (BBN), Multilayer Perceptron (MLP) networks and Alternating Decision Trees (ADtree) were compared for their ability to separate quasars from stars (Zhang & Zhao, 2007). Support vector machines (SVMs) have also been successfully applied to automatic classification (Zhang & Zhao, 2003, 2004). Decision trees, e.g. REPTree, Random Tree, Decision Stump, Random Forest, J48, NBTree and ADTree were investigated to classify active objects from non-active objects (Zhang & Zhao, 2007). In Radom Forest methods, where votes for class membership are polled from a large random ensamble of tree classifiers, has had many successful applications in astronomy (Albert et al., 2008; Gao, Zhang & Zhao, 2009).

The key tool: Principal Component Analysis

Another important method to identify the essential information from complex multi-dimensional datasets is represented by Principal Component Analysis (PCA). Through this mathematical-statistical tool, each galaxy spectrum can be linearly decomposed into fundamental components corresponding to a set of representative templates. The PCA naturally determines the minimum number of templates required to describe the sample, given the noise properties of the spectra. These templates show the features of the spectra that have the most discriminating power (the Principal Components). For astronomical spectra, the Principal Components have been shown to characterize well the spectral shape, and the presence of strong emission lines, allowing the sample to be divided into classes. Often these classes correspond to physical characteristics of the galaxy and can distinguish star-forming, post-starburst and passive galaxies (Connolly et al., 1995; Ferreras et al., 2006; Rogers et al., 2007, 2010).

The PCA has been applied to classify galaxies from the Sloan Digital Sky Survey (SDSS, York et al. 2000) (Yip et al., 2004; Dobos et al., 2012) and the 2 degrees Field Galaxy Redshift Survey (2dFGRS, Colless et al. 2001. The effectiveness of the method was confirmed well before in the separation of broad absorption line QSOs from a full QSO sample (Francis et al., 1993), in stellar classification (Murtagh & Heck, 1987; Storrie-Lombardi et al., 1994), in the classification of spectral energy distributions for stars (Singh, Gulati and Gupta, 1998), or in the classification of other galaxy spectra (Folkes et al., 1996; Sodré & Cuevas, 1997; Bromley et al., 1998; Galaz & de Lapparent, 1998; Ronen, Aragon-Salamanca and Lahav, 1999).

In particular, Folkes, Lahav and Maddox in 1996 investigated low signal-to-noise spectra with the PCA technique and reconstructed the underlying physical information using only 3 components. Combining the results of the PCA with a neural network approach they successfully classified a group of simulated spectra into different morphological classes.

Furthermore, Connolly and Szalay in 1995 carried out a classification of ten template galaxy energy distributions in terms of an orthogonal basis, to estimate the number of significant spectral components that comprise a particular galaxy type, finding a correlation between their spectral classification and those determined from published morphological classifications.

The application of classification methods to observed galaxy spectra presents some challenges. Spectra can be affected by spurious noise features, as positive or negative line residuals due to poor sky subtraction. This is the case of the spectra observed with VIMOS spectrograph at the ESO Very Large Telescope, prior to August 2010, where the fringes, produced by interference of bright sky lines with the CCD surface, resulted in the artefacts listed above. Other features can be the result of zero-order images of bright objects from adjacent spectra. All these features may have been corrected to some extent in the processed spectra, or be still present in the spectra. The many disguises these artefacts can take make it difficult to accurately classify spectral features (Roweis, 1997; Everson & Sirovich, 1995).

My work

I will show here that through the application of PCA I can accomplish the task of cleaning the spectra of noise artefacts while simultaneously obtaining a classification by means of a handful of parameters (**Chapter 4**).

This study is the first performed on the data of the new VIMOS Public Extragalactic Redshift Survey (VIPERS), the largest redshift survey program currently underway at

the European Southern Observatory Very Large Telescope (VLT) (Guzzo et al., 2014). VIPERS is designed to map in detail large-scale structure over an unprecedented volume of the $z \sim 1$ Universe (see **Chapter 3**).

In my thesis I have developed a specific PCA aimed at analysing and classifying the spectra collected by the survey. I will show that the technique is capable of compressing the majority of the observed spectral features into a small number of components, allowing an objective classification of the vast majority of the spectra in the sample.

The reasons for doing this on a survey like VIPERS are manyfold. First, it represents a way to objectively classify the survey spectra according to their spectral features. I shall show, in **Chapter 5**, how true this is by analysing both theoretical models and galaxy templates obtained from observed spectra. I will also show a comparison with a couple of other classification methods, to quantify the completeness and contamination of PCA classification with respect to those robust and widely used methods.

A further, important motivation for using a PCA classification is the possibility to homogeneously define sub-populations of galaxies, to be used for cosmological and evolutionary studies. For example, the analysis of galaxy sub-samples with different bias factors could provide a way to reduce the impact of cosmic variance on the measured cosmological parameters (e.g. McDonald & Seljak 2009). A PCA classification can also separate active and passive galaxies, helping to see the effects of environment on galaxy evolution. Furthermore, the classification could be used to help identify, in the VIPERS redshift range, the progenitors of specific populations of galaxies observed in the local Universe as the Luminous Red Galaxy sample of the SDSS (see for example Wake et al. 2006, Tojeiro & Percival 2010, Tojeiro et al. 2011), or for an analysis of correlation functions in the framework of redshift space distortions (Tojeiro et al., 2012).

A problem with PCA is in general that interesting but rare features can become lost in higher order eigen coefficients. I will address this problem in **Chapter 7** using a Linear Discriminat Analysis (LDA); since the LDA can be used as a complementary classification scheme to PCA, first I searched for the direction, in data space, associated to the maximum variation of two given data features: the intensity of 4000Å break and the [OII] line strength. Thus I defined a separation between data groups, on the basis of the strength of these two parameters for each galaxy. I will apply this LDA to the first 3 VIPERS Principal Components, using the intensity of 4000Å break and of [OII] as a fiducial class separator, and will define different loci for the passive-active-intermediate galaxies.

Then I will exploit LDA to implement a complementary separation between regular galaxies and AGNs, using the existing VIPERS AGN list as a training set: since the peculiarities of this kind of object are not reflected in the 3 principal components, I will apply the LDA to the entire set of (2486) components of the VIPERS dataset.

Another way to exploit to our favour the fact that PCA loses rare object's features, is to compare, in an automatic way, each PCA cleaned spectrum with the original observed one. In **Chapter 6** I will show the preliminary results, attempting to develop a technique to perform an automatic χ^2 comparison between each cleaned spectrum and its observed counterpart, to pick up the members of a very peculiar and hard to select class: the Narrow Line AGNs.

In the **Appendices**, I will show a technical application of the PCA, that performs an automatic cleaning of the VIPERS spectra. I applied the PCA to the same set of spectra, without bringing them to rest frame. This enables finding the principal components of the only features that many spectra at different redshift have in common: the sky signal. From the sky components and with some ad-hoc adjustments (related to VIPERS survey), I developed an automatic technique, able to produce a mask for every strong

sky artefact, and to realistically replace the masked portion of spectrum, again through a peculiar rest-frame PCA.

CHAPTER 1

The galactic zoo

1.1 The blue-red dichotomy

Galaxies can be easily divided into two main classes: the early-type galaxies, composed mainly by an old red stellar population, and late-type ones, showing a bluer color due to the presence of young stars, and also star formation regions. This dichotomy is evident in the majority of the characteristics of galaxies, from their color, to their shape or their spectra, and this holds true up to high redshifts. Of course the separation of galaxies into two groups is not sharp: a number of objects, sharing characteristics with both early and late type object, are identified as *green valley* transition objects. We know that the total stellar mass density of galaxies on the red sequence has roughly doubled over the last 6-8 Gyrs (e.g. Bell et al. 2004), while that of blue galaxies has remained roughly constant. Since new stars form primarily in blue galaxies, this suggests that galaxies are being transformed from the blue to the red population. Thus, we can state that the color of a galaxy is determined mainly by its star formation history, assuming dust extinction is properly taken into account, and that the color distribution of galaxies reflects a distribution in their current specific star-formation rates.

Other explanations could be given to the observed bimodality. A natural one is that the two normal distributions represent different populations of galaxies that are produced by two different sets of processes. In other words, formation processes give rise to two dominant populations that have different average colors and/or color dispersions. An evidence that the color bimodality can be due to this comes from the clustering analysis of Budavari et al. (2003): when the galaxy population was divided into four color bins, the two reddest bins showed similar clustering strengths, as did the two bluest bins, with a sharp transition in properties between them. This can be explained, if the dominant effect is the fraction of galaxies that are part of the red or blue normal distributions, rather than the average color of the galaxies. Galaxies that are part of the red distribution are more strongly clustered.

Finally, there are also several processes related to the environment that may be responsible for the observed bimodality, by transforming galaxies from late to early types, and by truncating their star formation rates. Some of these processes are typical of cluster environment, like galaxy Harassment and Ram-pressure stripping: galaxy Harassment happens when a galaxy encounters another one at high speed velocity, such that the colliding galaxy is impulsively heated; as a consequence, the galaxies become less bound and more vulnerable to disruptions by further encounters and by tidal interaction with the global cluster potential; when a galaxy is instead moving through the intracluster medium, it experiences a pressure that may strip the gas initially associated to the galaxy. Other processes are typical of groups environment, like merging and interactions. Another process that increases the fraction of red galaxies in both groups or clusters is the

strangulation process (Balogh, Navarro and Morris, 2000), that consists of the removal of the hot gas reservoir of infalling galaxies, so that their star formation halts after their cold gas is consumed.

1.2 The morphology of galaxies

Galaxy morphologies became evident when large and effective telescopes began to be used to observe the sky. In the late 18th century, the english astronomer William Herschel built metal reflectors which he used to sweep the sky for anything out of the ordinary, such as the objects called "nebulae". Herschel, and those who followed him, saw different kinds of nebulae: those that appeared to lie within the band of light called the Milky Way (galactic nebulae), and those which were found mainly away from the Milky Way (non-galactic nebulae). The non-galactic nebulae seen by Herschel had a variety of interesting shapes, ranging from round to highly elongated, and showed varying degrees of central brightness. In particular, now we know that each galaxy is composed mainly by these major components: a central ellipsoidal very luminous region, where the most of star formation occurs, a disk (in younger galaxies), where stars, gas and dust gravitate around the bulge, in rings with differential velocity motions, and a dark matter halo, which gave birth to the galaxy, surrounding it and extending over as twice as the luminous matter radius. The bulges are created during a formative evolution phase, where rapid, violent processes, such as hierarchical clustering and merging, led to their formation. Within the disk, the material is slowly arranged through the collective interaction of instabilities, during a secular evolution phase which leads to the formation of structures, as bars, ovals, spirals, rings and triaxial dark matter halos. Older galaxies, where the gas content has been consumed by stars and the angular momentum has been lost, display the shape of an ellipsoid with variable ellipticity from case to case. Albeit those objects represent the goal in the evolution of a galaxy, sometimes they may form as the result of a merger of galaxies. The spiral galaxies can exhibit a variable number of spiral arms, that can go from very wrapped to very loose. Finally galaxies can also be very irregularly shaped.

The morphology of galaxies has a clear correlation with galaxy colors: this has been known for a long time from photoelectric measurements (de Vaucouleurs, 1961), but also lately, for example in the SDSS survey, it was shown a clear bimodality in the distribution of colors that correlates with morphology, at a high degree of significance (Strateva et al., 2001): the red peak includes mainly elliptical and early spiral galaxies (barely sketched or very tightened spiral arms), while the blue one includes mainly open spiral and irregular galaxies. This can of course suggest that the morphology of a galaxy is an indicator of its stage of evolution.

1.3 The role of Active Galactic Nuclei

In some galaxies the central region is observed to outshine all the billions of stars in the galaxy itself. The spectrum is not like that observed from stars, and the emission is observed to be bright at all wavelengths. The luminosity varies on very short timescales, less than a day, and this means that the size of the central region is less than one light-day across (six times the distance from the Sun to Neptune). The most efficient conversion of matter to energy is the accretion by a black hole, and so we infer that it is a Super Massive Black Hole (SMBH) which is causing that emission.

The high energy and radio emission is direct, coming from the central regions around

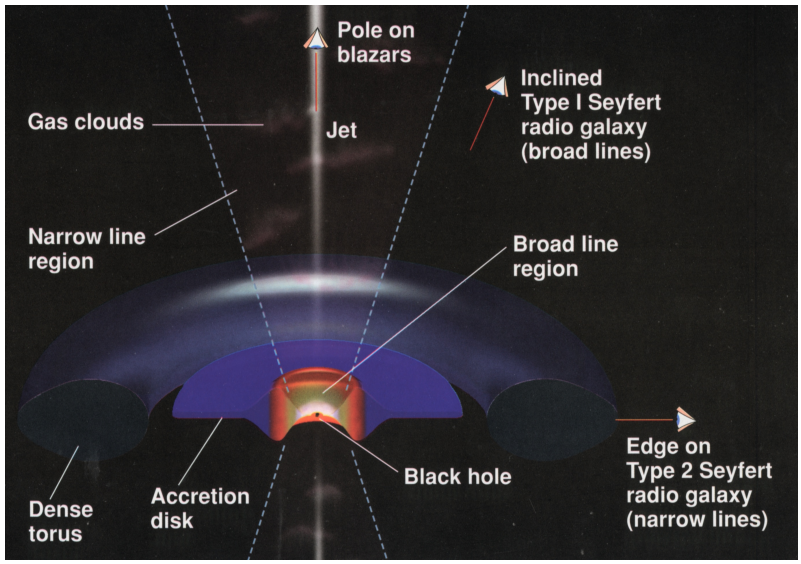


Figure 1.1: Explicative picture of the AGN unified model (Lessons of Professor Margaret M. Hanson, University of Cincinnaty, `www.physics.uc.edu`)

the black hole itself. In the optical and infrared wavebands the emission is not direct - the light has been absorbed and then is re-radiated (possibly at a new wavelength) by clouds of gas and dust, which surround the central "engine". Perpendicular to the accretion disk, two relativistic jets remove high energy plasma from the AGN, through to the magnetic field; the jets can extend up to more than 10 kpc from the central black hole. This is what is called an Active Galactic Nucleus (AGN).

Actually, most of the galaxies host a supermassive black hole at their center, but if the black hole is quiescent, i.e. it is not accreting matter, then the galactic nucleus is not active and the object is not classified as an AGN.

There are a whole menagerie of AGNs, depending on their brightness, their inclination w.r.t. the line of sight, presence or absence of jets and their observed spectrum, and the classification schemes used have been built up over the years (see §6 for a more detailed description). The lower luminosity AGN are called Radio Galaxies and Seyfert Galaxies; there are two types of Seyfert galaxies: Seyfert I show broad and narrow lines, while Seyfert II only exhibit narrow lines, due to the presence of a torus of dust and electrons in the line of sight direction. The more powerful AGNs are called Quasars (from Quasi-Stellar Objects or QSO, as they looked like stars in early telescopes), and blazars. Not all AGNs are strong radio sources, but many are, and they were discovered because they were radio bright objects, which looked like stars but were at much larger distances. As they are so bright that they can be seen across the Universe, they are a useful cosmological tool to measure the evolution of the Universe.

According to the AGN unification paradigm 1.1, all the different features an AGN can display, can be explained as a consequence of the inclination of the object with respect to the line of sight: a Seyfert I is an AGN whose disk is nearly perpendicular to the line of sight, a Seyfert II is near to parallel, so that the broad line features are hidden by the torus, a Quasar or a Blazar are orientated such that the relativistic jet is directed towards the observer.

1.4 Hints of galaxy evolution

Galaxies were born within cold dark matter halos: baryonic matter collapsed into the potential wells of dark matter, which began to accumulate, before the electromagnetic decoupling, from small primordial overdensities. From that moment, in about 1 billion years, the protogalactic clouds formed, and the protogalaxies, still made up of gas and dark matter, gave birth to the first generation of stars. After a few million years, the supernovae generated by the first generation stars seeded the galaxies with the first heavy elements, and heated the surrounding gas. This heating slowed the collapse of the gas, and the outer gaseous region of the galaxy settled down into a rotating disk, with a spheroidal hotter and denser bulge at its center. This happens when the protogalactic cloud has low gas density and high spin, thus, less star formation and fast rotation. If the protogalactic cloud is gas rich, and has little or absent angular momentum, this results in a quicker cooling and a faster star formation before the gas has time to settle in a disk: this may lead to originating an elliptical galaxy. These elliptical galaxies, having experienced an enhanced intial phase of star formation, look redder than the spiral ones, or late-type ones, since their population is dominated by older, more evolved red stars, and contain much less gas than the spiral ones. Actually an elliptical galaxy can also form from a merger of two spiral galaxies, or from the evolution of a spiral galaxy, that loses its gas rich spiral arms, consuming the gas to create stars, which gradually become old and red.

From the beginning the galaxies go on storing matter through interactions, fusion of structures, and accreting matter, flowing along the filaments towards the dark matter halos. According to the passive galaxy evolution paradigm, excluding temporarily the interactions and mergers that can largerly affect the evolutionary history of a galaxy, after the star forming-gas consuming phase, which implies bluer colors, due to the young stellar population, the evolution of a galaxy, which ends up composed mainly by old stars, or early-type galaxy, is mainly driven by the evolution of the single stars themselves, and the aging of a stellar population naively implies redder colors.

CHAPTER 2

Aims of the work

2.1 Galaxy classification methods overview

2.1.1 Photometric classification

A galaxy can be classified on the basis of its photometric properties, e.g. its brightness profile, its luminosity, its scale radius, or its spectral energy distribution (SED). With a modern telescope it's possible to register the image of a galaxy on the focal plane of a telescope as a matrix of pixels. Then the number of counts in each pixel can be converted to physical units of specific flux, from which the surface brightness of the object can be inferred. Galaxies typically have surface brightnesses with a regular behaviour, with a peak at the center decreasing towards the outer regions. If one considers the isophotes of a galaxy, i.e. the level curves at a constant surface brightness, it is possible to define the physical dimension of a galaxy; in facts, the isophotes are usually fitted by simple ellipses, at least in presence of spheroidal or thin disk galaxies. From the surface brightness profile $\Sigma(r)$ it is also possible to characterize the structure of galaxies and classify them. Elliptical galaxies are characterized by a de Vaucouleurs profile

$$\Sigma(r) = \Sigma_0 \exp\Big[-\Big(\frac{r}{r_0}\Big)^{\frac{1}{4}}\Big],$$

while the spiral ones have a de Vaucouleurs profile for the bulge and a surface brighness for the disk which follows

$$D(r) = D_0 \exp\Big(-\frac{r}{r_0}\Big),$$

where r_0 identifies the radius at which the luminosity is half of the central one, and Σ_0 and D_0 the central surface brightnesses.

Integrating $\Sigma(r)$, one obtains the total flux F, and from F the total bolometric luminosity of the galaxy. In general, for the spiral galaxies, the central surface brightness is nearly constant, while in the elliptical it is strictly related to the luminosity. Since the color of the light emitted by a galaxy is dominated by the more luminous stars in it, in a late-type spiral galaxy the emitted light is blue, since the more luminous stars are the blue principal sequence stars; in an elliptical galaxy the more luminous stars are the rarefied red giant, which determine the red color of these kind of galaxies. A photometric visual classification of galaxies on the basis of their rest-frame colors can be easily accomplished: the observed color of a galaxy is given by the difference of magnitudes in two different bands; this quantity must first be corrected to evaluate the true intrinsic colors of the stars. In particular, to recover the intrinsic color, the extinction due to interstellar absorption, the Earth atmospheric absorption and the signal loss inside the observing instrumental system, must be corrected to the observed color. Then the color

of each galaxy in the sample can be plotted against another color, typically (B-V) vs. (U-B), which easily separates red from blue galaxies into two distinct blobs (see Fig. 5.6).

2.1.2 SED fitting

Galaxies emit electromagnetic radiation over the full possible frequency (wavelength) range. Analysis of this radiation is the main means through which to study distant galaxies and thus learn about their formation and evolution. The distribution of energy over wavelength/frequency is called the Spectral Energy Distribution (SED).

Integrated spectral energy distributions (SEDs) are the primary source of information about the properties of unresolved galaxies. SED fitting can be used effectively to derive a range of physical properties of galaxies, such as redshift, stellar masses, star formation rates, dust masses, and metallicities. Indeed, the different physical processes occurring in galaxies, all leave their imprint on the global and detailed shape of the spectrum, each dominating at different wavelengths. Detailed analysis of the SED of a galaxy should therefore, in principle, allow us to fully understand the properties of that galaxy. SED fitting is thus the attempt to analyze a galaxy SED and to derive one or several physical properties simultaneously, from fitting models to an observed SED.

Galaxies emit across the electromagnetic spectrum. Excluding those galaxies dominated by an accreting supermassive black hole at their nucleus (the AGNs), the ultraviolet to infrared spectra of all galaxies arises from stellar light, either directly, or reprocessed by the gas and dust of the surrounding interstellar medium (ISM). Thus the UV-to-IR spectral energy distribution or SED contains a large amount of information about the stars of a galaxy, such as the stellar mass to light ratio, and the surrounding ISM, such as the total dust mass. However, to extract such information, models are necessary in order to connect physical properties of the galaxy with the observed SED.

In its simplest sense, a galaxy is a population of stars ranging from numerous, low-luminosity, low-mass stars, to the bright, short-lived, massive OB stars. On closer examination, these stars are distributed in both metallicity content and age, ranging from when the galaxy first formed to those newly born. The method of creating a model synthetic galactic spectrum through the sum of the spectra of its stars is called stellar population synthesis or SSP (Tinsley, 1972). A simplification for the modelling of galactic SEDs is that the emitted light can be represented through a sum of spectra of simple stellar populations (SSPs) with different ages and element abundances. Here a SSP is an idealized single-age, single-abundance ensemble of stars whose distribution in mass depends on both the initial distribution and the assumed age of the ensemble.

Once obtained the model SEDs, there are different statistical methods to fit a spectrum with them, and the main results of the SED fitting are, for example, as hinted above, the determination of the photometric redshift of a galaxy, the inferring of its stellar masses, of its attenuation by dust or its dust emission, or its star formation rate.

2.1.3 Morphological classification

The fact that the bimodality in the galaxy distribution (see Chapter 1) is manifest in many parameters in a similar way, is not surprising: in fact the galaxy parameters related to structural and stellar population, as explained above, are known to be well correlated. This is actually the basis of classification schemes such as the Hubble Sequence (Hubble, 1926; Sandage, 1961). Hubble introduced the classification scheme illustrated in the Fig. 2.1, also called Hubble *tuning fork*, which separates most galaxies into elliptical, normal spiral, and barred spiral categories, and then sub-classifies these categories with respect

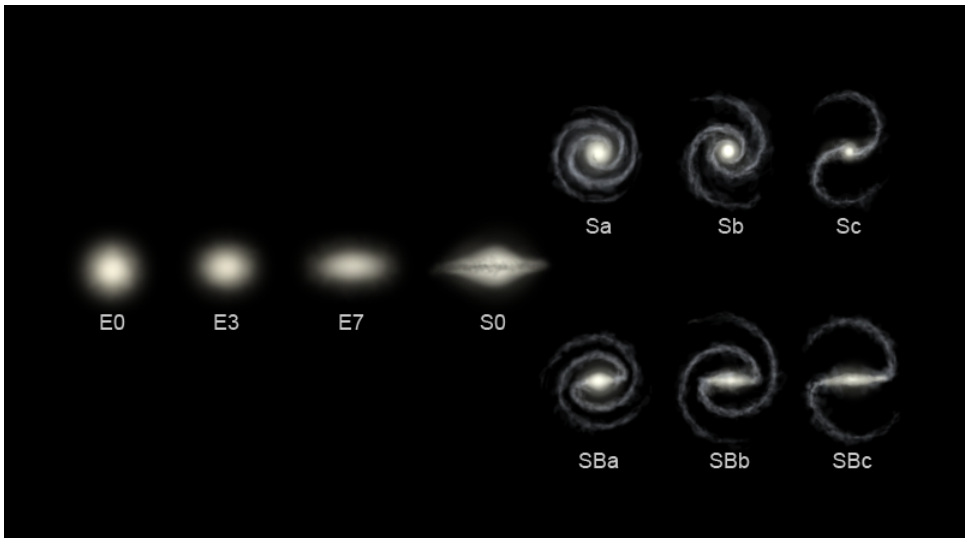


Figure 2.1: Hubble-Sandage morphological classification scheme

to properties such as the amount of flattening, for elliptical galaxies, and the nature of the arms, for spiral galaxies. The galaxies that do not fit into these categories are classified separately as irregular galaxies. In particular, the diagram is roughly divided into two parts: elliptical galaxies and spiral galaxies. Hubble gave the ellipticals numbers from zero to seven, which characterize the ellipticity of the galaxy - *E0* is almost round, *E7* is very elliptical. The spirals were assigned letters from a to c, which characterize the compactness of their spiral arms. Sa spirals, for example, are tightly wound whereas Sc spirals are more loosely wound. Also it is worth noting that the sizes of the round central regions in spirals - the bulges - increase in size the more tightly the spiral arms are wound. There are indications pointing to a very close connection between the bulges of certain galaxies (Hubble types $S0$, Sa and Sb) and elliptical galaxies.

The spiral galaxies are sub-divided into two groups - normal spirals and barred spirals. The most important difference between these two groups is the bar of stars that runs through the central bulge in barred spirals. The spiral arms in barred spirals usually start at the end of the bar instead of from the bulge. Barred spirals have a B in their classification. An SBc is thus a loosely wound barred spiral galaxy. $S0$, or lenticular galaxies, are in the transition zone between ellipticals and spirals and bridge these two types. Hubble found that some galaxies are difficult to put in the context of the tuning fork diagram. Those include irregular galaxies which have odd shapes, dwarf galaxies which are very small, and giant elliptical galaxies which are very large elliptical galaxies residing in the centers of some clusters of galaxies. For a time the Hubble tuning fork was thought to be an evolutionary sequence - that galaxies might evolve from one type to another progressing from left to right across the tuning-fork diagram. Hence E, $S0$, Sa and SBa galaxies were called early-type, while Sc and SBc were called late-type. Astronomers still use this nomenclature today, though the initial concept was later found to be an over-simplification. Galaxy evolution is a far more complex process than Hubble imagined, involving the conditions of the galaxy's initial collapse, collisions with other galaxies, and the ebb and flow of internal star birth.

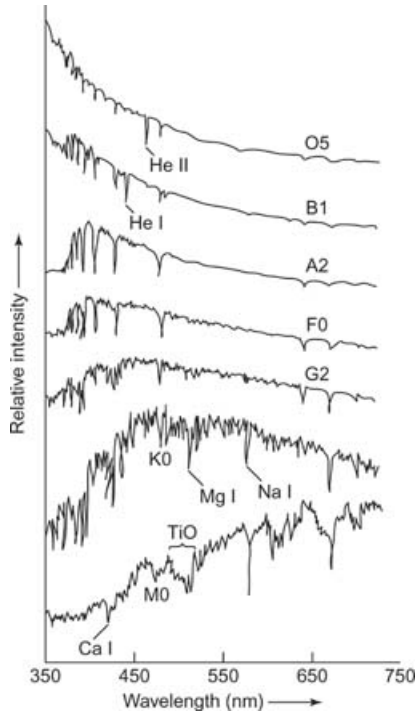


Figure 2.2: Model galaxy spectra from latest-type (top) to earliest-type (bottom), with absorption lines indication. In the red region of blue spectra the type of stars responsible for the shape of the continuum and absorption lines are marked.

2.1.4 Spectroscopic classification

Galaxy can be also classified on the basis of their spectroscopic properties. In facts, as explained above, each galactic spectrum has a shape and features that reflects its average stellar poulation, together with the gas and dust components. In particular, a galaxy spectrum can be ideally decomposed into two main components: the continuum and the line features. The line features in turn can be separated into absorption and emission feature, which have a different physical origin one from the other. The continuum component is caused by a combination of a range of blackbody emitters (the stars) spanning a given range of temperatures, producing a nearly flat (see later in the section) overall spectrum.

In an elliptical (typically early-type) galaxy, the spectrum shows a strong "step" at 4000Å, in the galaxy's own reference frame (Fig. 2.2), called 4000Å break; the position of the break, in wavelength, is marked by the presence of the two strong absorption lines of Ca II, produced by the atmospheres of old stars. The spectrum, to the red side of the break, is almost twice as bright as the blue side: this effect arises because of an accumulation of absorption lines of mainly ionized metals at the left of the break, and by a deficiency of hot, blue stars. As the opacity increases with decreasing stellar temperature, the 4000 angstrom break gets larger with older ages, and it is largest for old and metal-rich stellar populations, behaving as an indicator for the age of a galaxy. There are absorption features superimposed on the continuum: they are due to the absorption of the atoms (metals) and the molecules in stellar atmospheres, that absorbe specific wave-

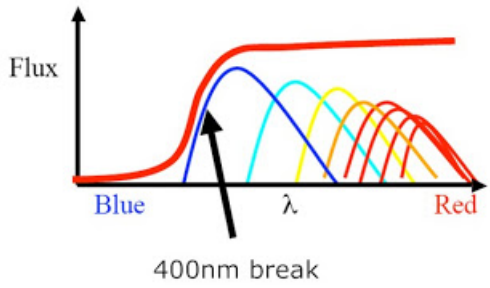


Figure 2.3: Cartoon showing the superposition of stellar black body spectra, to produce the galaxy spectral continuum. In the absence of the 4000Å break, due to increase in opacity of stellar atmospheres, the slope would be increasing steadily towards the bluer wavelengths, producing the typical spiral or late-type continuum.

lengths, and can be also due to cold, interstellar gas clouds in the interstellar medium, which can extract energy from the passing radiation at key frequencies. The presence of absorption features implies the presence of old stellar populations, which are typically found in elliptical galaxies and in the bulges of spiral galaxies. Key features include the Calcium H and K lines (found at 3934Å and 3969Å), the G-band (4304Å), and Magnesium (5175Å) and Sodium (5894Å) lines.

Besides the continua and absorption features there are also emission features, which are much more intense and frequent in the late-type galaxy spectra (Fig. 2.2): these are due to gas being heated and then re-radiating energy at specific wavelengths. Young stars form within gas clouds, which they then ionize, remaining often embedded in their cloud. Emission features thus pertain to very hot gas and hot young OB type stars, so they are typical of the disks of spiral galaxies, and of irregular galaxies. Key emission features include the [OII] doublet (3737A), [OIII] (4959A and 5007A), and the Balmer series (6563A, 4861A, 4340A, 4103A, ...).

The continuum of a late-type galaxy shows very weak or totally absent 4000Å break. Thus, the continuum has an opposite slope with respect to an elliptical galaxy one, and reflects only the typical shape of the composition of various blackbody spectra, which is not perfectly flat: the radiation of hotter stars have higher black body peak intensities and shorter wavelength ranges, viceversa for the colder stars, producing a decreasing slope of the continuum towards the red region of the spectrum (Fig2.3).

From spectroscopy it is possible to associate an accurate distance to a galaxy: from the evaluation of the redshift of the position of (at least two) known features of the spectrum, it is possible to infer the cosmological recession velocity of a galaxy (neglecting its peculiar motions inside its cluster); hence, given the cosmology, the distance is obtained on the basis of the Hubble law.

In large spectroscopic surveys, the redshift-inferred spatial distribution of galaxies, in clusters, is always affected by the neglection of the peculiar motions, besides the Hubble flow; the resulting observed distribution is subject to the so-called *redshift space distortions* effect. The precise evaluation of the parameter which describes such distortions, is still challenging, and it's a key step to understand the conflicting interactions between dark matter and cosmic expansion, described by the cosmic *growth function* of structures. An accurate evaluation of f is crucial for modern cosmology: in fact it will be able to discriminate between a Dark Energy dominated Universe scenario, and one in which a Modification of Einstein's laws of gravity can explain the present evidence of acceler-

ated expansion (Guzzo et al., 2008). The EUCLID satellite is forecast for 2017, with this precise aim (`http://www.euclid-ec.org/`).

2.2 Principal components spectroscopic classification

Spectroscopy provides the deeper insight into the physics of galaxies, but the high information content of the data set makes it difficult in general to compress and classify all the information contained in a galaxy spectrum in a compact and efficient way.

The main goal of my thesis was to develop a classification method, able to encompass the most important spectral features of a survey sample, with a fast and efficient statistical approach. Amongst many statistical methods, an important one that can be used to identify the essential information from complex multi-dimensional datasets is represented by Principal Component Analysis (PCA). This method is the basis of my classification routine, which I refined, further on, through the addition of another statistical tool, the Linear Discriminant Analysis (LDA) . Besides the classification goal, I also applied the PCA, and, separately, the LDA, to try to select a peculiar sample of galaxies, and the PCA itself has been applied to build a cleaning mask for the sky features of the spectra in a survey.

2.2.1 Principal Components Analysis

The Principal Component Analysis is a non-parametric way, to extract the majority of information from a noisy dataset, composed of objects which are not completely different one from another. The key characteristic of the PCA in this case is, in fact, the ability to describe a large sample, through a reduced number of components, which is guaranteed by the fact that the objects in the sample share many common features (e.g. different measurements of the same quantity, a collection of objects in a catalogue, etc...). This holds true for a sample of galaxy spectra, that are generated by a common underlying physical mechanism, i.e. the radiative physics in the galaxies.

PCA finds the linear transformation that changes the frame of reference from the observed or natural one, to a frame of reference that highlights the structure and correlations in the data. This is done through a rotation of the parameter space, such that the axes are aligned along the directions of maximum variance of the data. This transformation may be found by diagonalizing the data correlation (or covariance) matrix, whose eigenvectors effectively represent the axes of the new coordinate system. In the specific case of galaxy spectra, I started from a correlation matrix (the mean has not been subtracted to the data, as in the case of the covariance matrix), such that the first eigenvector represents the mean of the spectra, and the other eigenvectors residuals form this mean. This way the information content of the first eigenvectors is maximized.

The basis of the principal components one obtains will be made up by orthogonal (i.e. uncorrelated) vectors or *eigenvectors* which are linear combinations of the original variables. The PCA has the advantage to describe a set of measurements exploiting dimensions of the problem which are uncorrelated, and that can be easily ordered by decreasing importance. This allows us to retain just a (small) subset of components, describing the data using a basis of only a few eigenvectors.

My goal is to reduce the complexity of a sample of spectra by expressing them through just a handful of the principal components. In particular, one may write an observed spectrum as a data vector containing N fluxes f_λ, where λ indexes the N wavelength bins. My sample contains M spectra, and one can write the sample correlation between

wavelength bins as a matrix,

$$C_{\lambda_1,\lambda_2} = \frac{1}{M-1}\sum_{i=1}^{M} f^i_{\lambda_1} f^i_{\lambda_2}, \tag{2.1}$$

where i indexes the spectra in the sample and λ_1 and λ_2 index wavelength bins. The correlation matrix can be decomposed into a set of orthonormal eigenvectors, or *eigenspectra* e^i_λ and eigenvalues Λ_i,

$$C_{\lambda_1,\lambda_2} = \sum_{i=1}^{M} e^i_{\lambda_1} \Lambda_i e^i_{\lambda_2}. \tag{2.2}$$

The eigenspectra are ordered with decreasing eigenvalue such that the most common features within the spectra are contained in the first few eigenspectra.

The eigenspectra form an orthogonal basis or *eigensystem* and any spectral energy distribution, f_λ, can be expressed as a sum of the M eigenspectra with linear coefficients a_i:

$$f_\lambda = \sum_{i=1}^{M} a_i e^i_\lambda. \tag{2.3}$$

Since the higher eigenspectra carry little statistical information about the spectra, I may truncate the sum to use only the first $K \ll M$ components. I refer to this as the reconstructed spectrum $\hat{f}_\lambda$,

$$\hat{f}_\lambda = \sum_{i=1}^{K} a_i e^i_\lambda, \tag{2.4}$$

The correlation matrix, as defined in (2.1), will have dimension given by the number of wavelength bins (2486x2486 in my particular case, as will be clearer later). In the literature, it is also common to define the correlation matrix such that the dimension is the number of spectra (Connolly et al., 1995). This is clearly inefficient when the number of spectra is greater than the number of wavelength bins.

An additional result obtainable by the PCA projection of eq. (2.4) is a measure of the signal-to-noise ratio for each spectrum, as

$$\frac{S}{N}(f_\lambda) = \sqrt{\sum_\lambda \left(\frac{\hat{f}_\lambda}{\overline{n}_\lambda}\right)^2} \tag{2.5}$$

where $\overline{n}_\lambda$ is the normalized noise spectrum, relative to the spectrum f_λ. Given the noise spectrum n_λ, the normalized noise spectrum is given by $\overline{n}_\lambda = n_\lambda / \sqrt{\sum f^2_\lambda}$.

2.2.2 Linear Discriminant Analysis

Linear Discriminant Analysis is, together with PCA, a widely used technique for dimensionality reduction and data classification. The primary difference between the two approaches is that, while the latter is useful for feature classifications, the former is suitable to do data classification: PCA changes the shape and location of the data to rotate them in a space where the overall variance is maximized; LDA instead just draws a separation region, which "points" in the direction that maximizes the difference between classes, thus providing better class separability. To perform that, anyway, LDA needs to base itself on a training set, composed by data that have been already roughly separated

into distinct classes, by a different method. Relying on this preliminary class distinction, LDA finds its own best separator between the different classes, supplementing the fiducial separation.
The mathematical process requires the computation of the mean μ of the data **a**, within the training classes; in the simplest case, which is the one of interest for me, I deal with 2 classes, thus two means, namely μ_1 and μ_2. The covariance matrices C_1 and C_2 of the two classes are then computed in the usual way. The vector w, which represents the direction of maximum variance between the two classes, is then given by

$$w = C^{-1} \cdot \Delta \tag{2.6}$$

where $C = C_1 + C_2$ and $\Delta = \mu_1 - \mu_2$. The data can be projected on the separator w

$$p = w \cdot \mathbf{a}^T \tag{2.7}$$

and put into a histogram, where the boundary line l between the two classes, that hopefully are separated into two distinct histograms, is set by

$$l = w \cdot \frac{\mu_1 + \mu_2}{2}. \tag{2.8}$$

CHAPTER 3

The Data

3.1 Extragalactic spectra from the V.I.P.E.R.S. survey

The PCA classification developed in this thesis has been applied to the spectra of the ongoing VIMOS Public Extragalactic Redshift Survey (VIPERS) (Guzzo et al., 2014), of which we analyze here the first public data release (PDR-1) (Garilli et al., 2013).

VIPERS has been designed to collect $\sim 10^5$ redshifts to the same depth of VVDS-Wide and zCOSMOS (Lilly et al., 2009) (i_{AB} <22.5), but over a significantly larger volume and with high sampling (40%). The general aim of the VIPERS project is to build a sample of the global galaxy population that matches in several respects those available locally (z <0.2) from the 2dFGRS (Colless et al., 2001) and SDSS projects, thus allowing combined evolutionary studies of both clustering and galaxy physical properties, in a comparable statistical footing. Building upon the experience and results of previous VIMOS surveys, VIPERS arguably provides the most detailed and representative picture to date of the whole galaxy population and its large-scale structures, when the Universe was about half of its current age.

The survey is providing the community with an unprecedented spectroscopic database at 0.5< z <1.5, covering a total area of $\sim$ 24 deg^2 within the CFHTLS-Wide W1 and W4 fields, including extensive information on galaxy physical properties. The latter is made possible by combining the spectral information with the CFHTLS five-band magnitudes on which the survey is based (Goranova et al., 2009), plus additional ancillary data in the UV and infrared bands, enabling to derive SED information and automatic galaxy/AGN/stellar classification (Marchetti et al. 2013, Davidzon et al. 2013) . The set of data used in this paper has been made public with the VIPERS Public Data Release 1 (PDR-1). Several aspects of the survey construction and the data are also discussed in detail in Guzzo et al. (2014): it is conceived with primary the goal of studying Redshift Space Distortions (RSD), but it provides raw material for studies of large scale structures and galaxy evolution. VIPERS was designed to maximize the number of galaxies observed at z> 0.5, while the contamination by stars reaches about 30% in some fields. The desired redshift range was determined through a color-color selection in the (*r-i*) vs (*u-g*) plane,

$$(r-i) > 0.5(u-g) \ \text{OR} \ (\text{r}-\text{i}) > 0.7, \tag{3.1}$$

which removes galaxies at z< 0.5 and produces a sample complete at the 98% level for z> 0.6 (Fig. 3.1). The magnitude limit is set as 17.5$\leq i_{AB} \leq$22.5, after correcting for galactic extinction. The resulting redshift distribution is then the one of Fig. 3.2.

The Target Sampling Rate (TSR) for the VIPERS measurements, defined as the ratio of the targeted galaxies over the potential targets is around 40% (Figs. 3.3, 3.5), while the Spectroscopic Success Rate, i.e. the ratio of the reliably measured redshifts over the

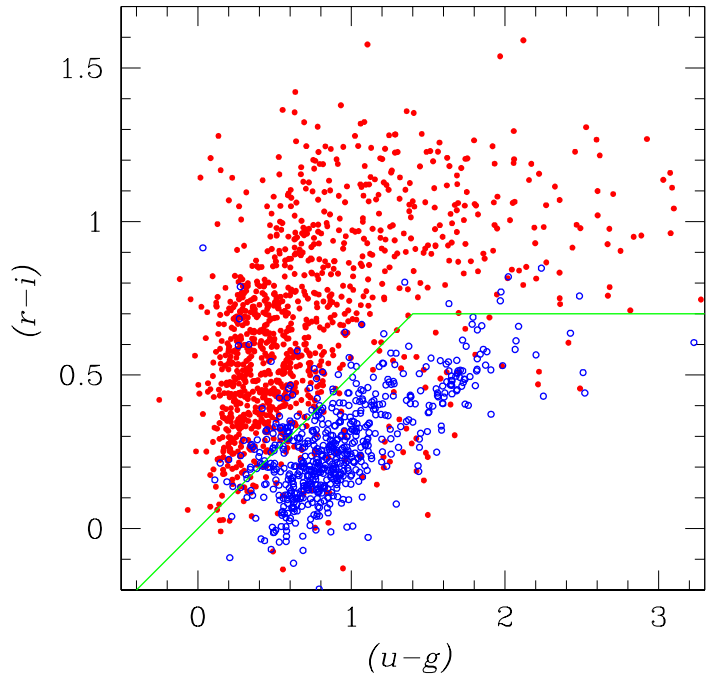


Figure 3.1: Colour selection of the VIPERS galaxies with i_{AB} <22.5, in the $(r-i)$ vs $(u-g)$ plane. The red filled circles represent the z >0.5 objects, while the blue empty one are the z <0.5 ones. The green line is the cut expressed by eq.3.1 (Guzzo et al., 2014).

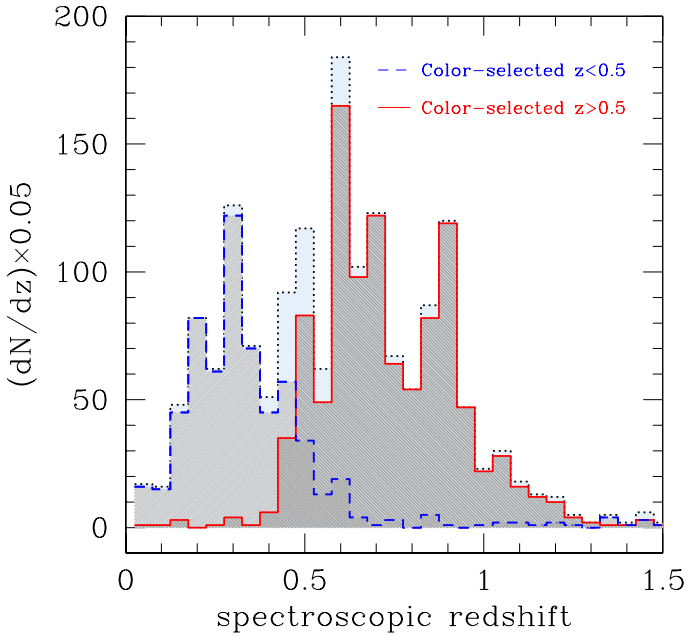


Figure 3.2: VIPERS Spectroscopic redshift distribution after the selection of eq.3.1 (Guzzo et al., 2014).

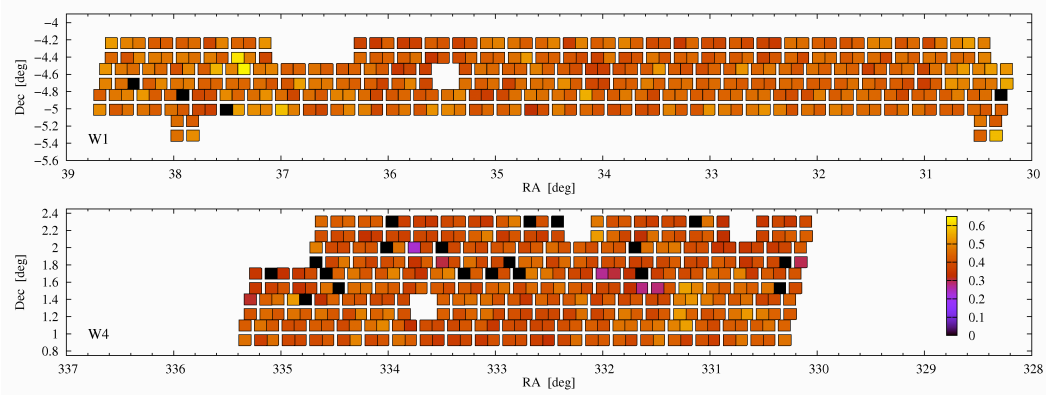


Figure 3.3: Representation of the pointings on the sky, for both fields W1 and W4. The colors indicate the Target Sampling Rate (TSR) as illustrated by the color gradient in the bottom plot. The black quadrants correspond to the loss of the data, due to a failure in the insertion of the mask Guzzo et al. (2014).

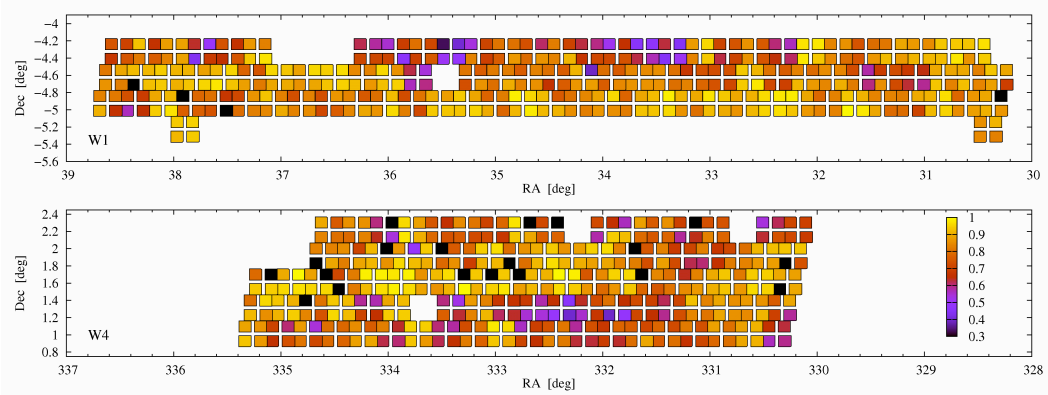


Figure 3.4: Representation of the poitings on the sky, for both fields W1 and W4. The colors indicate the Spectroscopic Success Rate (SSR) as illustrated by the color gradient in the bottom plot. The black quadrants correspond to the loss of the data, due to a failure in the insertion of the mask (Guzzo et al., 2014).

number of target galaxies, is larger than 80% for the majority of observed quadrants (Figs. 3.4, 3.5).

I anticipate here that the incompleteness of the data may slightly affect my PCA classification, in this direction: since some galaxy types (in particular emission line galaxies) are more easily assigned a redshift than other galaxy types, in the final active-passive or blue-red classification, this may have a non-negligible impact.

The data of the first VIPERS Public Data Release (PDR-1) are available at `http://vipers.inaf.it` ($\sim$ 55.000 redshifts and related informations).
Redshifts and quality flags are measured with the PANDORA EZ (Easy Z) package (Garilli et al., 2010): VIPERS is the first VIMOS redshift survey for which the data reduction is fully automated. The redshift and flags are assigned by the PANDORA pipeline, and have been checked and re-fined, for every spectrum, by members of the VIPERS team, ensuring the reliability of the assignments. The quality flag indicates the confidence of the redshift measurement in a similar manner as used in the VVDS (Le Fèvre et al., 2005) and zCosmos catalogues (Lilly et al., 2009). The flag takes the form $\pm XY.Z$. Negative values are reserved for spurious, undetected or unidentified serendipitous sources. The first digit X indicates the class of object: it is blank for normal galaxies; 1 for broad-line AGNs, and 2 for untargeted sources serendipitously measured. The second digit Y indicates the confidence of the redshift measurement. Secure redshift measurements with nearly 95% confidence are assigned $Y = 4$. Measurements with 90 % confidence limit are assigned Y=3. Y=2 measurements have been shown to correspond to a confidence limit of about 80%. Y=1 sources are highly uncertain at the 50% confidence level, and Y=0 is given when a redshift could not be assigned. For this reason, these two classes are not considered in the present analysis, to guarantee a clean and reliable sample. Finally, Y=9 is given to redshift measurements that are based upon only a single emission line feature. The flag also has a decimal part that indicates the agreement between the photometric redshift estimate and the spectroscopic redshift, but I do not use it here. Throughout this work I will apply the word "flag" the Y digit (or to the XY digits, in case of AGNs).

The total number of VIPERS spectra available for this first study before any quality cut is 57204.

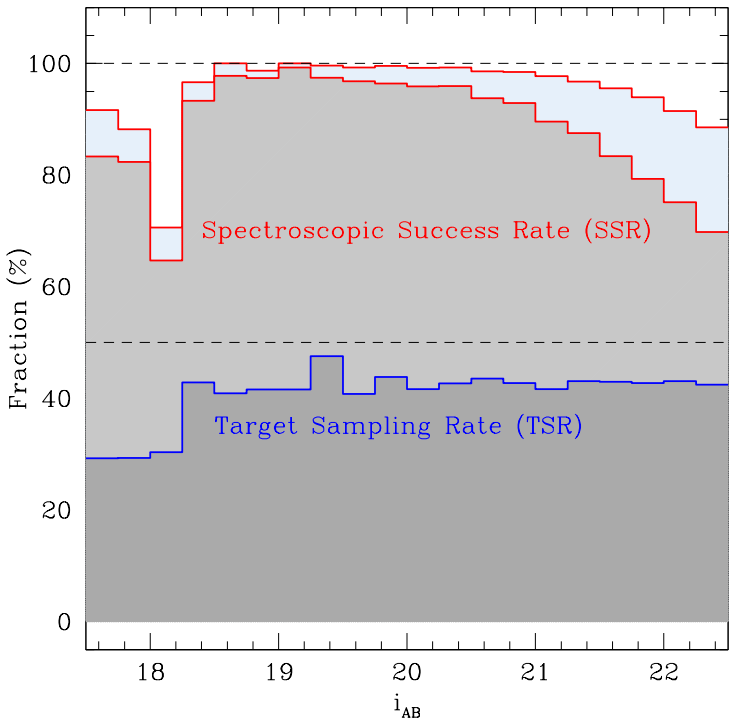


Figure 3.5: Target Sampling Rate (TSR, lower dark grey histogram) and Spectroscopic Success Rate (SSR, two upper light grey histograms), as a function of galaxy magnitudes. The TSR is independent of galaxy magnitudes, indicating that there is no bias in terms of apparent luminosity in the process of assigning galaxy targets to slits. About efficiency in measuring redshifts, the two top histograms correspond to the SSR when all measured redshifts (flag ≥ 1) are considered and when reliable redshifts (flag ≥ 2) are used.(Guzzo et al., 2014).

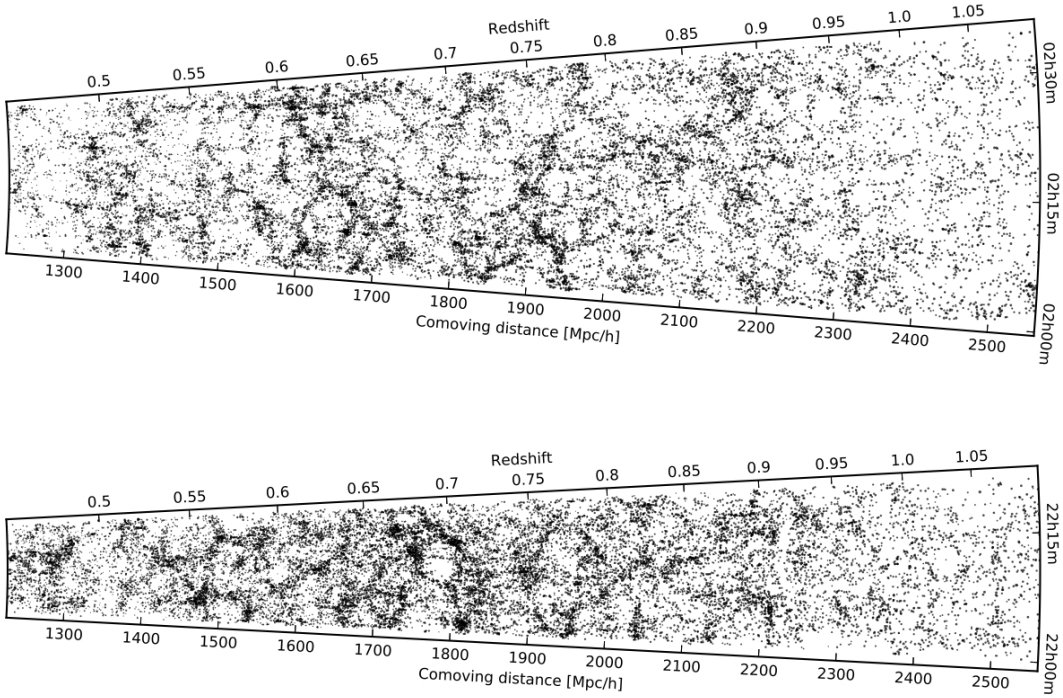


Figure 3.6: Large scale structure distribution of VIPERS PDR-1 catalogue in the W1 (top) and W4 (bottom) fields, where the position of each galaxy is projected along the declination direction (Guzzo et al., 2014).

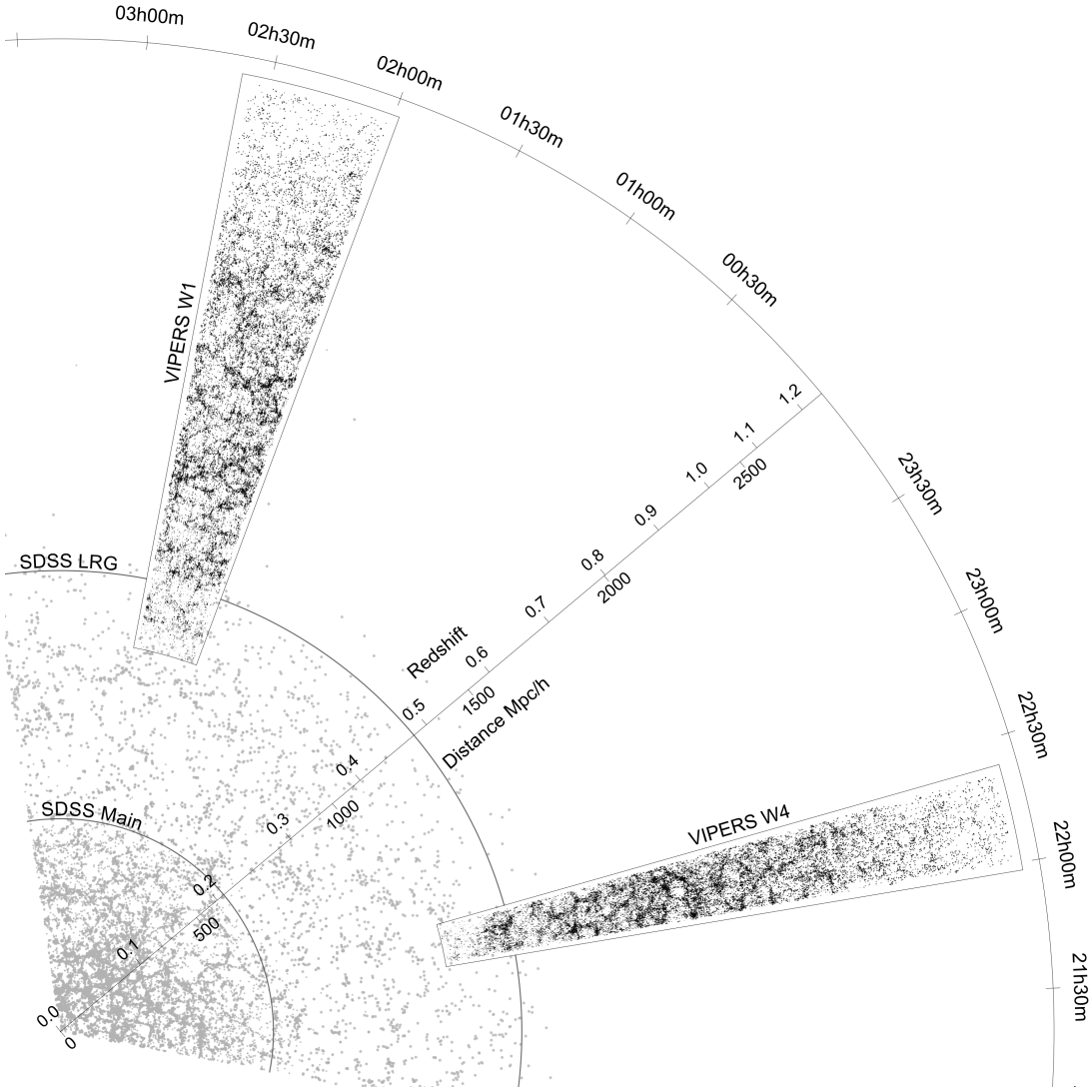


Figure 3.7: Comparison of the two VIPERS fields with the SDSS main sample and the SDSS LRG sample (Guzzo et al., 2014).

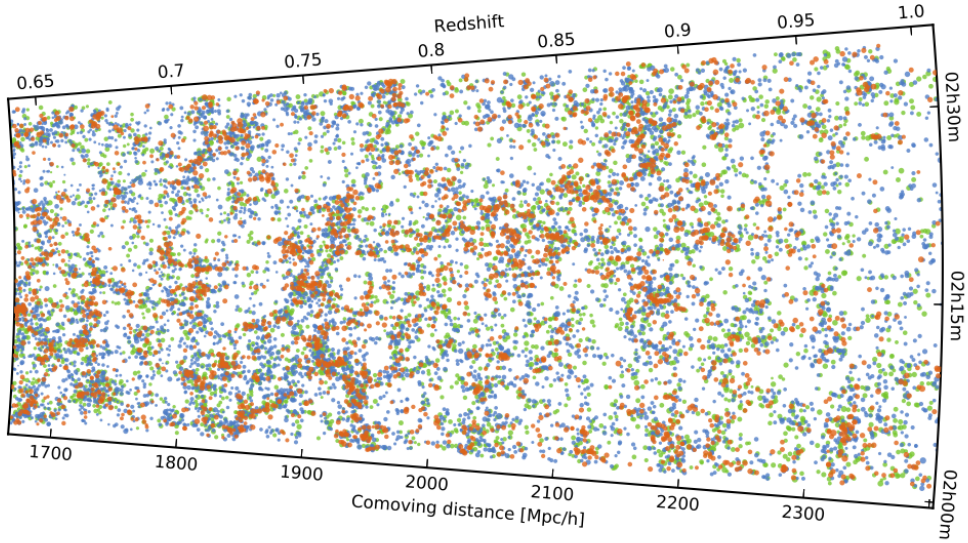


Figure 3.8: Zoom into the cone diagram of the W1 field. The reddish points mark early-type galaxies, the green intermediate ones, and the blue the late-types. This plot evidences the tendency of early-types to distribute along the main structures of the underlying matter distribution, contrarily to the bluer objects, which prefer the lower density regions (Guzzo et al., 2014).

The most impressive result from the PDR-1 sample is given by the maps of the 3D galaxy distribution, in the range $0.5< z <1.2$, in Fig. 3.6: this represents an unprecedented combination, in size and sampling, of the galaxy population of the Universe at its half-of-present size. Fig. 3.7 shows a comparison of VIPERS with the SDSS main and Luminous Red Galaxy (LRG) sample: while SDSS LRG sample is an excellent statistical probe on the largest scales, VIPERS is much more efficient in unveiling the details of the underlying nonlinear structure. Fig. 3.8 finally, reveals the power of VIPERS measurements, in correlating galaxy properties with the surrounding large-scale structures, in particular the clustering of galaxies as a function of galaxy properties: the redder points, marking early-type galaxies according to their $U-B$ rest-frame colors, lie along the "pillars" of structure, while the blue-green points, representing star-forming or intermediate type galaxies, occupy low-density regions. This proves that the colour-density relation for galaxies is already standing at these redshifts (Cucciati et al., 2006).

On the PDR-1 sample, I applied a further selection, which excludes low-quality spectra as defined above, but includes sources classified explicitly as broad-line AGN and secondary sources observed by chance. I note that there is no harm in including peculiar spectra as AGN in the overall PCA. Being rare cases ($\sim$5% of the total in VIPERS, between Narrow and Broad line objects), these have no effect on the evaluation of the Principal Components characterizing the main galaxy sample (see Chapter 4). At the same time, as we will discuss in Chapter 6 and 7, it is possible to identify AGN-like spectra by the PCA as "outliers" or through an LDA analysis, among the more standard galaxy spectra. This may lead also to detection of more AGN-like spectra, which do not appear explicitly classified as such.

3.2 Data Manipulation

The spectra are observed over a fixed wavelength range. Thus, they must be shifted and mapped to a common rest frame wavelength scale, to match all the common features, that will be recognized by the PCA. I have defined a rest-frame wavelength scale, ranging in $3500\text{Å} < \lambda < 5500\text{Å}$, to get the maximum coverage of signal in all redshift bins. The redshift range is $0.4 < z < 1.0$, which encompasses a large fraction of the redshift range of the survey, excluding the very far and very near objects. The final sample, after these cuts, includes 42,036 spectra ($\sim$ 73% of the total in PDR-1). The wavelength binning we chose to adopt in this work increases logarithmically, and in such a way that the last interval in the reddest region has a width of 1Å. This results in a total number of bins of 2486. This wavelength scale ensures that every VIPERS spectrum is oversampled in the rest frame. All the spectra are shifted by a factor of $(1+\mathrm{z})^{-1}$, to bring them to rest-frame, and resampled with a linear interpolation on to the previously defined rest-frame grid (this resampling does not preserve the flux density, but since the interpolation grid as a much finer scale than the observed one, the effect on the analysis will be negligible). The variance is given, for each spectrum, by the square of the relative VIPERS noise spectrum, and it is processed in the same fashion as the spectra.

Obviously, as in Marchetti et al. 2013, resampling a spectrum on to the rest-frame grid can leave gaps at the start or end of the scale, depending on the redshift (Fig. 3.9). And again, additionally, noise spikes due mainly to sky fringing may have been manually edited, producing a gap in the corresponding wavelength range (again Fig. 3.9). Nevertheless, the iterative algorithm used here (Marchetti et al., 2013) repairs the spectra before finding the principal components.

An important consideration before moving to the real analysis is how to normalize each spectrum. The apparent flux of any source introduces an arbitrary scaling factor, that

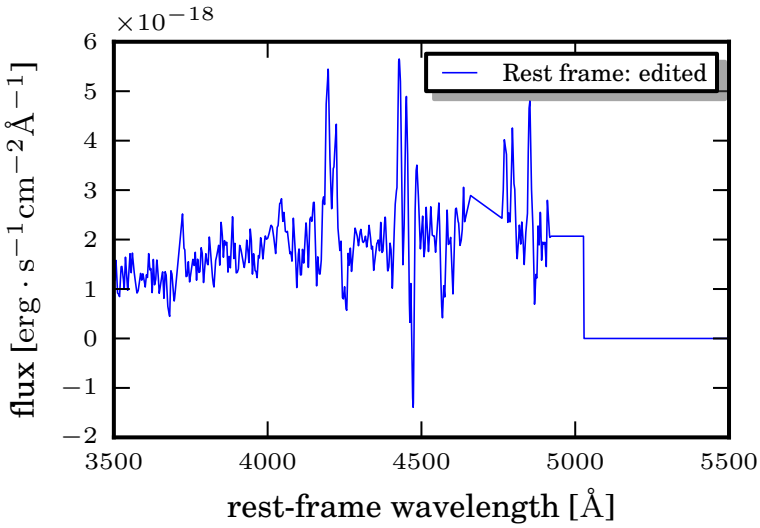


Figure 3.9: Vipers spectrum after the rest-frame moving, displaying an evident gap in the larger wavelength region, and a couple of manual linear interpolations in place of large noise spikes.

should be normalized out to build a homogeneous sample to feed the PCA algorithm. Amongst many possible normalizations, we choose to normalize each spectrum by a scalar-product normalization, such that for a spectrum f_λ, the normalized spectrum becomes

$$\overline{f}_\lambda = f_\lambda / \sqrt{\sum f_\lambda^2}. \tag{3.2}$$

The choice is dictated by the fact that normalizing by scalar product offers advantages for our classification over other possible normalizations (Connolly et al., 1995): a normalization based on morphology would rely on a model distribution of morphological types in given sample, and may lead to the accidental suppression of a common galaxy type within the first principal components of the sample; a normalization by the integrated flux will give similar results as one done by scalar product, in terms of principal components, but this second one (that we prefer) produces unit vectors representing the spectra, and unit principal components. This means that the coefficients of the decomposition of each SED on the principal components lie on the surface of an N-dimensional hypersphere (if we consider N principal components), and thus can be parametrized by using only N-1 parameters (see §4.1).

CHAPTER 4

Principal Component Analysis of VIPERS data

4.1 Repairing the spectra

As hinted at the end of Chapter 3 , a spectrum can be corrupted by instrumental artefacts, as well as poorly subtracted sky features, or contamination of light from a nearby object. As a matter of fact, VIMOS has its own specific features: it is possible that a slit in the mask got contaminated by a zero-order image from the adjacent slit above, or that residuals may persist after the subtraction of sky lines. In some cases, such artefacts have been removed from the spectra by the reduction pipepline, or manually, and have been replaced by linear interpolations, creating what I call "gaps" in the spectra, i.e. regions where flux data was lost (Fig. 4.1). Fig. 3.9, in the previous Chapter, illustrates the same spectrum of Fig. 4.1, with the region that has been removed, around 4700-4800Å in rest-frame, and replaced with an unrealistic inclined straight line, besides another interpolation before a hole in the red region of the spectrum, due to the transport of the spectrum to rest-frame. These modifications must be properly taken into account when applying a PCA decomposition, to avoid treating some bad features, gaps, poor interpolations or noise artefacts as physical peculiarities, that will influence the shape of the eigenspectra, and hence the whole analysis.

To account for that, I assigned a weight to each spectral bin, according to the usual definition:

$$w_{f_\lambda} = \frac{1}{n_\lambda^2}. \tag{4.1}$$

Here I used the square of the statistical VIPERS noise n (properly brought to rest-frame and resampled on the common binning) as the variance, as already specified in Chapter 3. With the definition above, each wavelength bin of each single spectrum has its own weight.

This weight is set to 0 whenever there is a 0-flux gap or a gap due to manual editing: the latter have been located by simple comparison of the observed spectra to the edited ones in the VIPERS database. This weight mask is essential to derive accurate eigenspectra from data containing gaps. In fact, with a naive application of PCA to these "gappy" spectra, it is no longer possible to construct a set of orthogonal eigenspectra (Connolly & Szalay, 1999), and the PCA would likely lose some fetures or treat the gaps as real properties of the spectra, if many gaps happend to gather in the same wavelength region. We have therefore developed an algorithm to simultaneously repair those gaps in the spectrum and thus compute orthogonal eigenspectra.

At the start of the repairing routine, the gaps in the spectra are indeed simply replaced by "horizontal" linear interpolations. Although, for gaps at the start or end of

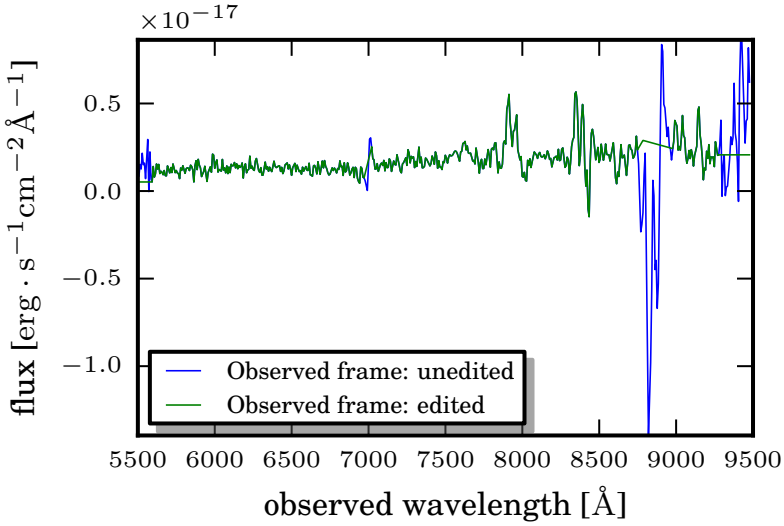


Figure 4.1: An observed frame VIPERS spectrum, presenting a huge unsubtracted sky spike around 8800Å (blue), and the edited version of the same spectrum, with linear interpolations in place of the noise spikes (green).

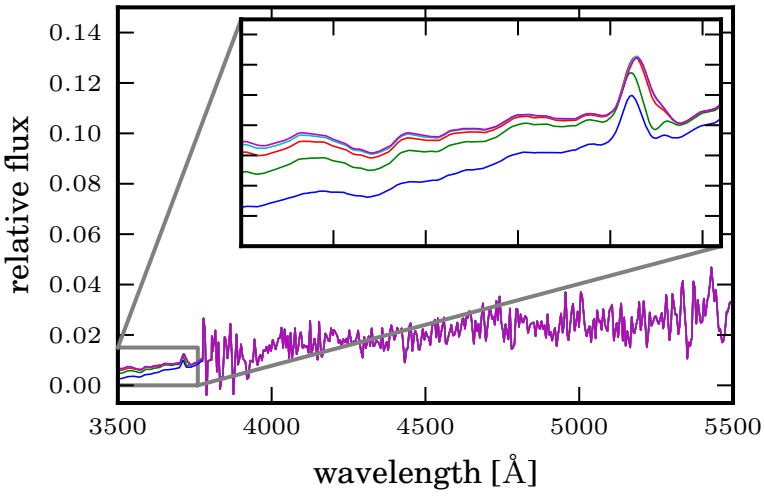


Figure 4.2: A VIPERS spectrum presenting a gap on the blue side, due to rest-frame shifting. The missing data is reconstructed through an iterative routine. The first five steps (zoomed in the box) go from the first (bottom line) to the fifth iteration (top line).

the spectrum, I find that it is sufficient to simply leave the flux to 0. I then proceed in an iterative manner.

First, I create a data matrix containing all the rest-frame spectra in the rows

$$\mathbf{M} = \begin{bmatrix} f_{1\lambda_1} & \cdots & f_{1\lambda_n} \\ \vdots & \ddots & \vdots \\ f_{N\lambda_1} & \cdots & f_{N\lambda_n} \end{bmatrix} \tag{4.2}$$

then the correlation matrix (no mean subtraction) is constructed from the spectra

$$C_{\lambda_1,\lambda_2} = \frac{1}{n-1} \sum_{i=1}^{n} f^i_{\lambda_1} f^i_{\lambda_2} \tag{4.3}$$

and the eigenspectra e_i are computed, from the diagonalization of the correlation matrix:

$$C_{\lambda_1,\lambda_2} = \sum_{i=1}^{n} e^i_{\lambda_1} \Lambda_i e^i_{\lambda_2}. \tag{4.4}$$

We keep only the 3 most significant eigenspectra to perform the following repairing steps. The choice of the number of eigenspectra, as discussed later in this Chapter, and justified by the tests I performed on the routine (§4.3), is dictated by the need to be able to describe all the spectra in the sample, while avoiding the noise, which, as I will show, is reflected by the eigenspectra from the fourth on.

Once determined the number of principal components, I compute the set of eigencoefficients, $\{a_i\}$, for each spectrum, f_λ, i.e., the projection of every single spectrum on the 3-eigenspectra basis, by means of a least squares minimization routine. In this case the objective function to be minimized is given by,

$$\chi^2 = \sum_{\lambda} w_\lambda (f^{(i)}_\lambda - \sum_j a_j e_{j\lambda})^2. \tag{4.5}$$

where $f^{(i)}$ is the single spectrum data vector on the i^{th} iteration, $e_{j\lambda}$ is the set of eigenspectra and w_λ is the weight vector. Thus, at the end of the process, I am left with a set of 3 principal eigencoefficients for every spectrum. Those coefficients could be computed, in principle, through a mathematical projection, but I will show later that this would produce some unphysical reconstructions. To avoid this, I will introduce a penalty term in the coefficients determination process, optimizing the minimization through a non linear solver, the Levenberg-Marquardt algorithm implemented in the Python Scientifical Library (SCIPY)[1].

4.2 A new PCA approach

After the application of the PCA machinery described above, I found that, in some cases, the best-fitting coefficients, used to combine the 3 principal components, did not result into physical spectra. For example, it may happen that the continuum of the repaired spectrum goes negative in some region, or strong emission lines could be inverted with respect to their expected appearance.

[1] `www.scipy.org`

These poor results are usually found for very noisy spectra, or for spectra that are more than 50% masked (i.e. for which the weight vector w is composed by more than 50% by zeros): in fact, when many spectra have been masked in the same range of wavelengths, the PCA process becomes unable to find the information to repair those gaps. In our VIPERS sample, there are 57 spectra that are missing more than 50% of the wavelegth coverage, while instead the average gap fraction for the sample is $\sim 10\%$.

The other possibility, that outcomes in an unphysical reconstruction, is that some peculiar piece of information needed to recover a spectrum is not reflected within the chosen eigenspectra (see §2.3).

These problems, in the majority of the interested cases, cause the PCA to fail to reproduce simultaneously the continuum and the line features of these spectra. This likely leads to the inversion of some lines; indeed, the pixels representing the spectral continuum have more weight than those in the lines, being present in larger fraction, and the PCA routine reproduces the former as accurately as possible, but at the expense of the line features: in facts, the routine combines the eigenspectra to return a repaired spectrum, which is as much fitted to the global shape of the observed one as possible; and of course, being the global shape much more related to the continua than to the line features, the PCA may combine the eigenspectra with some negative coefficient, such a way that the resulting continuum is respected, but the resulting line features come out negative or shorter than expected.

To avoid these degenerate solutions, we introduced a check within the wavelength range of the line features that mostly suffer from this problem in our routine: [OII], Hβ and [OIII]. Whenever the least-square repairing routine finds an inverted line (i.e. a negative line in my case) as a solution for the fitting problem, I add an exponential penalty term to the χ^2 in the minimization routine:

$$\chi^2 = \chi^2 + c * \sum_l e^{(D_l - D_0)/D_0} \tag{4.6}$$

where c =2486 is the number of bins in a spectrum, D_l is the difference between the continuum and the line peak for each line l, and D_0 =0.005 is the threshold above which the penalty is applied. The value of D_0 has been chosen such a way to impede the PCA to reverse emission lines, whilst avoiding this penalty to be applied by small real dips within the elected wavelengths, for example in red galaxy spectra. In this way, whenever the PCA finds a negative solution for a real emission line, during the phase of repairing, the χ^2 gets raised and the routine is therefore forced to find a set of eigencoefficients corresponding to a more physically realistic reparation. The specific choice of this shape for the penalty has been the result of a number of tests using different functions, given the freedom allowed by the problem.

After finding the best-fitting coefficients, $\{\hat{a_i}\}$, I find a global repairing of the spectrum as,

$$y_\lambda = \sum_j \hat{a}_j e_{j\lambda}. \tag{4.7}$$

Then I replace the gaps (and only the gaps) in the original spectrum with portions of the projection. In Fig. 4.2 we show an example of different stages of repairing. At each iteration the spectra are renormalized by their scalar products (the normalization changes at every iteration, because the gaps are updated on every loop). The routine progresses as shown in the diagram of Fig (4.3). When the repairing is complete (see $4.3), the eigenspectra that I obtain are the ones representing the fully repaired sample. The first 4 eigenspectra are depicted in Fig. 4.8. Although we make use of the first 3 only,

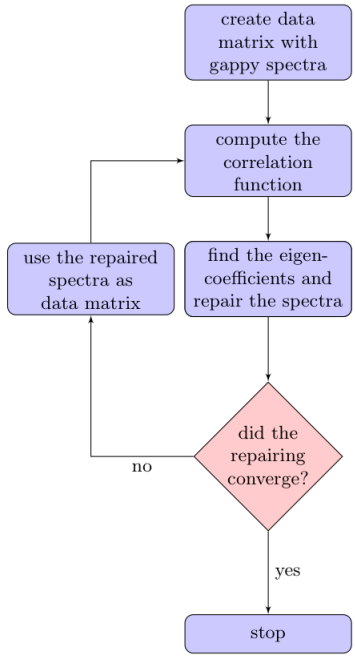


Figure 4.3: Flow chart of the PCA repairing process.

I plotted the fourth to show how it already displays the presence of noise, as a bump in the red wavelength part. At this stage I can project each spectrum on to the eigenbasis according to the set of 3 eigencoefficients a_i.

The convergence of the routine is safely reached, for each of the spectra, within the twentieth iteration of the process: after this, any further refinement of the value of the eigencoefficients for the repairing does not change the repairing significantly, as shown in §4.3.

4.3 Testing the PCA routine

To test my routine I created a synthetic sample of galaxy spectra. The spectra were generated using two sets of templates: a subset of the Bruzual&Charlot (B-C hereafter) (Bruzual & Charlot, 2003) model spectra (which do not contain emission lines), to obtain realistic early-type galaxies, and the 12 Kinney-Calzetti templates (K-C hereafter) (Kinney et al., 1996; Calzetti et al., 1994) (the plots of these model spectra can be found in Kinney et al. (1996)), covering from pure bulges to starburst galaxies, to give a total of 45 template spectra. I computed the first five eigenspectra of these templates to define an orthogonal basis spanning all the range of galaxy types.

Then I constructed mock spectra, that are similar to the templates, by generating Gaussian distributed numbers as eigencoefficients. This Gaussian distribution is centered on the first 5 eigencoefficients of the starting template set, with variance given by the relative eigenvalues. This way I generated 450 mock spectra around each template giving a total sample of 20,250 spectra, that reduces to about 16,000 once spectra presenting unphysical features (i.e. inverted emission lines) are removed.

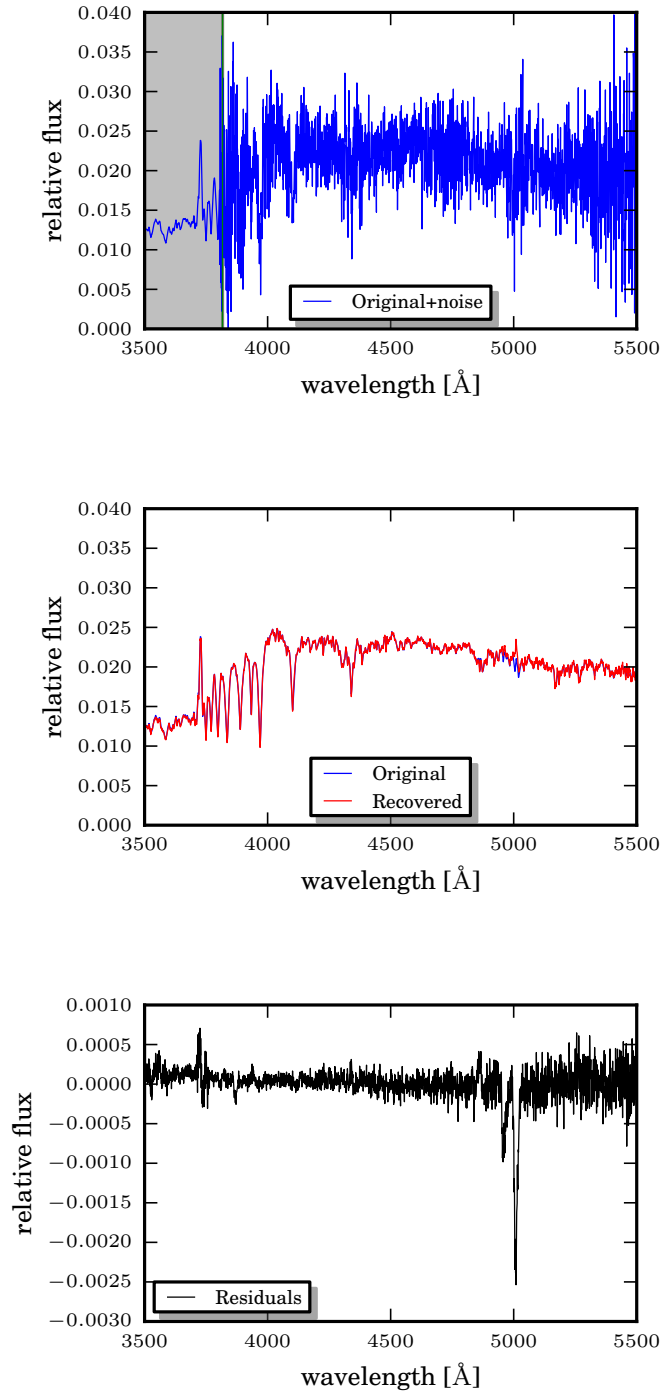


Figure 4.4: Top: a synthetic spectrum with synthetic noise added. The shaded region would be masked and reconstructed. **Middle**: qualitative comparison between the original spectrum before the noise has been added (blue) and its reconstruction through the PCA routine (red). **Bottom**: residuals between the mock and its reconstruction. The possible differences between the intensities of the real and the recovered emission lines are acceptable for our classification system, since it is more sensitive to the continua of the spectra than to the line features.

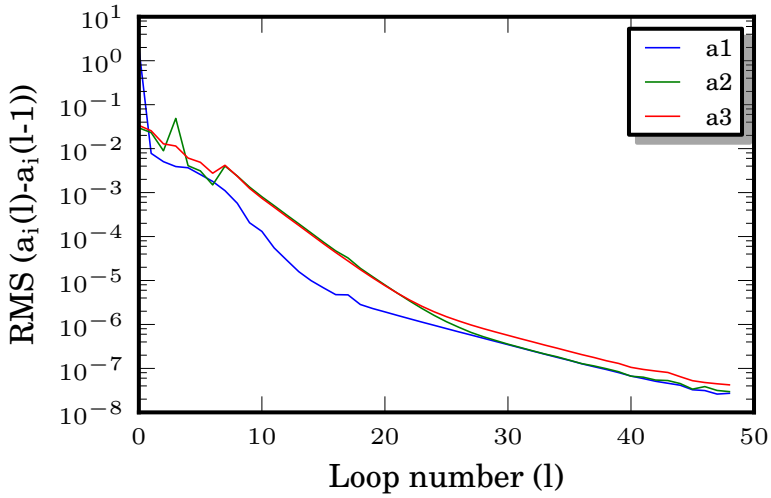


Figure 4.5: The root mean square difference between the eigencoefficients and themselves at the previous iteration, for the repairing of the synthetic spectra. The RMS difference steadily decreases on subsequent iterations.

I next degrade the spectra with synthetic noise to simulate the VIPERS data. Each synthetic spectrum is assigned the same data variance and weight mask of a randomly selected VIPERS galaxy. The synthetic noise spectra are generated from a Gaussian realization, with the associated VIPERS variance, as illustrated in the top panel of Fig. 4.4, and the mask is applied to reproduce the synthetic gaps. In this way, I produce an artificial data set, of which the expected shape is known, that can be used to test the fidelity of the reconstruction procedure.

I apply the PCA repairing routine with three eigenspectra, since 3 is the number I reasonably expect to be sufficient to describe the majority of the sample. Then I project the spectra on them, to clean from noise and be able to compare the recovered spectra to the noise-free synthetic ones.

Apart from slight differences in the intensity of the emission lines (as anticipated in §4.2) the reconstruction is qualitatively good, even where the region to be repaired was a line feature (Fig.4.4: middle-bottom, Fig. 4.5 for a more quantitative check) . The fit can be improved by adding more components to the PCA, but, as was anticipated, and will be discussed later, the 4th eigenspectrum is already affected by noise for the VIPERS sample, and the reconstruction obtained with three is sufficient for the classification system.

The PCA routine has been run on the synthetic spectra for a large number of iterations, that I arbitrarily chose to be a large number, 50. By looking at the root mean square difference between the eigencoefficients at each iteration (Fig. 4.5) I see that the routine is converging: in particular, the differences between the eigencoefficients become steadily smaller. The effects of this on the repairing is actually negligible after five iterations, so I halt the code when the difference between the eigencoefficients at consecutive loops is $\leq 10^{-3}$: although the threshold can be set in more rigorous ways (Yip et al., 2004), I

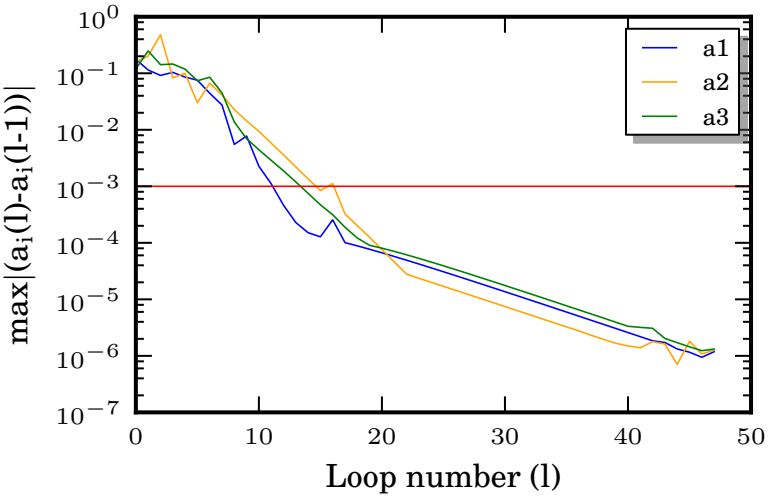


Figure 4.6: Difference between the coefficent and itself at the previous iteration (for the 3 coefficients) for the last objects reaching the convergence for at least one of the 3 coefficients.

chose this value empirically, as the threshold beneath which any further refinement has a negligible effect on the repairing.

I also found that the repairing for every single spectrum has surely reached the 10^{-3} difference threshold within the 12th iteration for a_1, at the 17th for a_2 and at the 14th for a_3 (as illustrated in Fig.4.6), so in this case 17 iterations are enough to repair and recover the original spectra for the synthetic spectra. To be on the safe side, we decide to take 20 iterations.

4.4 PCA decomposing the sample

Here I'll show the results of the application of the PCA routine to the VIPERS sample. As anticipated in sections §4.2 and §4.3, I must decide on a stopping point for the repairing routine and the number of eigenspectra to use.

As suggested by the tests on mock spectra, I halt the repairing procedure after 20 iterations. I may also estimate the relative error in the coefficients after each iteration, by measuring the root mean square difference between the value at iteration i and iteration 20. Fig. 4.7 shows that this error is oscillating at the level of 10^{-4} alredy by the 10^{th} iteration.

We use three eigenspectra in the repairing procedure to reconstruct the spectra inside the gaps. This number should be chosen to be large enough such that the repairing can reproduce the signal without adding spurious noise, although the results are not strongly dependent on the exact number used.

As said, after the convergence of the repairing process, I obtain the complete eigenspectra for the VIPERS sample. The first four eigenspectra ordered by significance are shown in Fig. 4.8. The first three VIPERS eigenspectra, as quantified later in this section, contain the large majority of information on the sample, particularly the first one, which

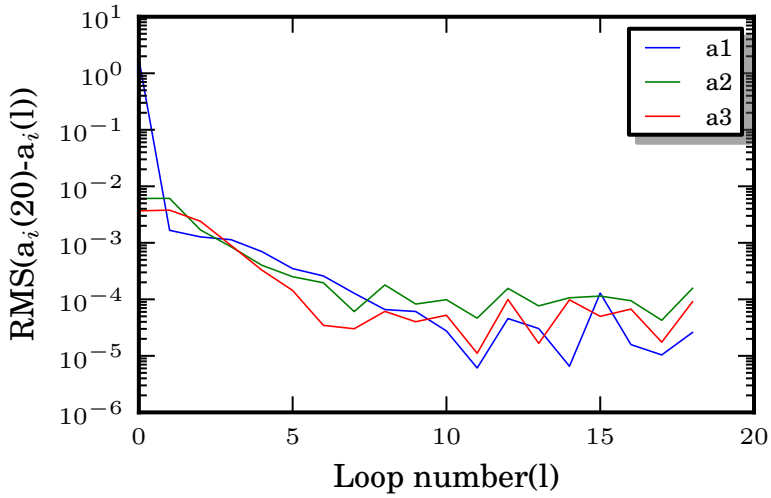


Figure 4.7: The RMS error on coefficients for VIPERS spectra. Plotted is the root mean square difference of the coefficients of the decomposition after 20 iterations, and themselves at the i^{th} iteration. For a particular spectrum the difference actually starts oscillating around 0 with decreasing amplitude after the 5-10th iteration on average.

mirrors the average of all the spectra, while the others represent the residuals from the mean. In particular, the shape of the continuum of the first eigenspectrum is comparable to the one of an early-type galaxy, while it contains also emission lines typical of a star-forming galaxy. The second one instead can be associated to a late-type spectrum, while the third can be thought of as an intermediate galaxy SED.

The fourth one, at $\lambda < 4500$Å, adds information about the intensity of the [OII] emission line and the continuum resembles the one of a blue galaxy, but redward of 4700 Å it shows an unphysical bump that is not expected in a galaxy continuum. We attribute this to the fact that, redwards of $\lambda_{obs} > 8000$Å , VIPERS spectra are affected by systematic effects arising from the coupled effect of detector fringing and strong sky emission lines (Guzzo et al., 2014); in fact, if I project the spectra onto 4 eigenspectra, for the VIPERS spectra for which this fringing problem was not already fixed (pre-refurbishment spectra), the value of the 4th coefficient is on average higher (in module) than for post-refurbishment spectra (Fig. 4.9).

For low signal-to-noise objects the repairing of this region is probably more affected by systematic uncertainties that can heavily influence the PCA reconstruction. Thus, to effectively repair the spectra without spurious features, I use only the first three eigenspectra.

The physical shape of the first 3 eigenspectra is offered by the fact that VIPERS spectra are low resolution (R=210): this leads to the incapability to resolve some doublets, like [OII], but also guarantees that the eigenspectra are not forced to catch the line width variabilities (except for broad line AGNs), resulting into P-cygni like profiles (Beals, 1953).

To justify quantitatively the choice of using the first 3 eigenspectra, I compute a sim-

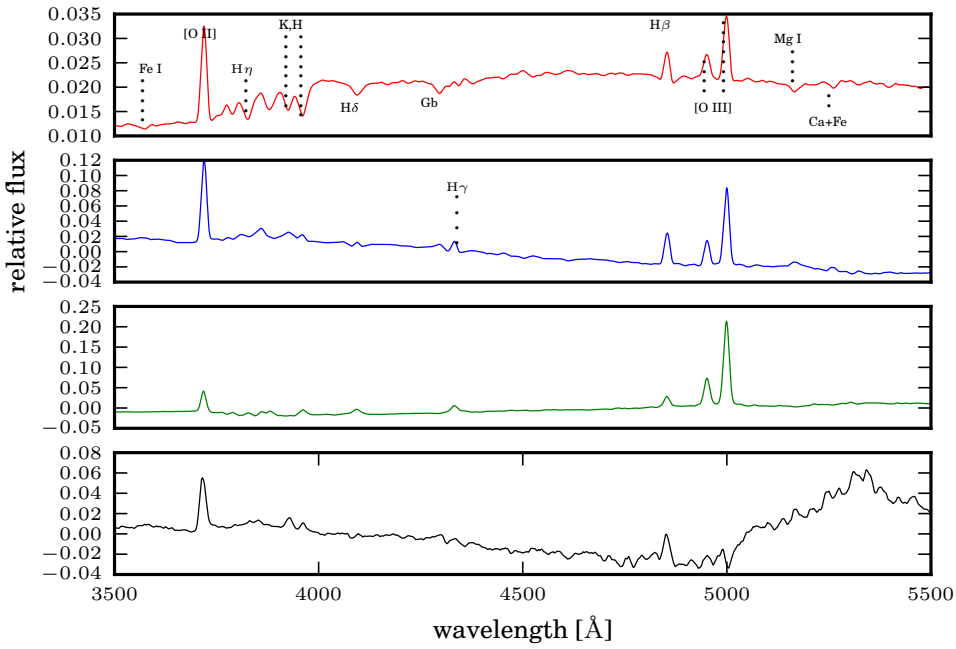


Figure 4.8: The first four VIPERS eigenspectra computed after repairing. From top to bottom the power is decreasing (the first eigenspectrum is at the top, the fourth at the bottom). The first eigenspectrum mirrors the average of all the spectra, while the second and the third are residuals form the average. Some of the most common spectral features present in the eigenspectra are highlighted in the first eigenspectrum. Systematic effects in the spectra begin to be visible in the fourth spectrum at $\lambda > 5000$Å.

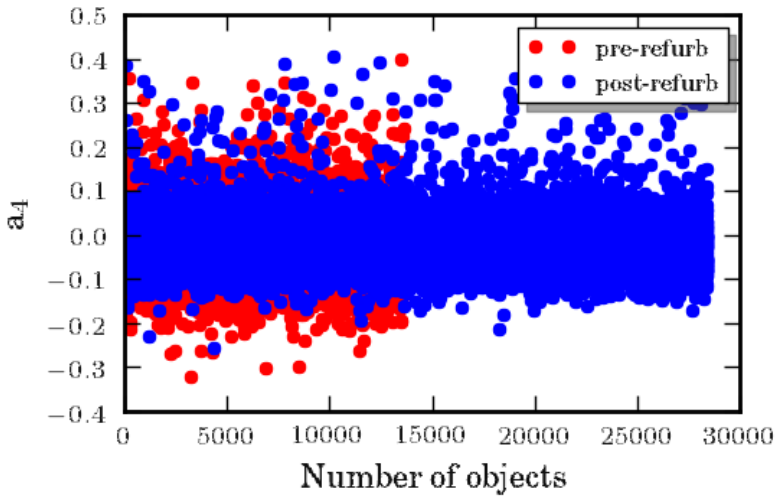


Figure 4.9: Values of the coefficient of the 4th eigenspectrum, in a 4 eigenspectra decomposition, for pre-refurbishment spectra (affected by strong fringing effect redwards of 4700 Å) (red dots) and post-refurbishment ones (finging fixed) (blue dots). For the pre-refurbishment objects, the contribution of the 4th eigenspectrum is in general more important, as expected, since the 4th eigenspectrum continuum is affected by noise at the same wavelengths of pre-refurbishment spectra.

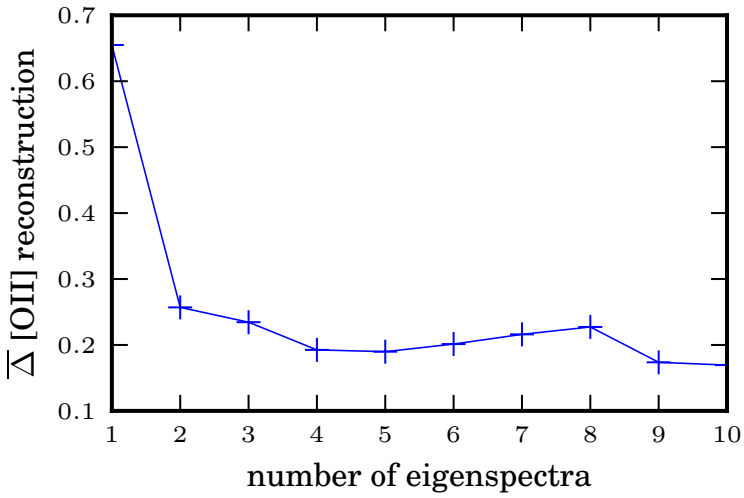


Figure 4.10: Average difference between observed and recostructed [OII] fluxes (divided by observed fluxes) as a function of number of eigenspectra. The best line reconstructions seem to be obtained beyond 9 eigenspectra.

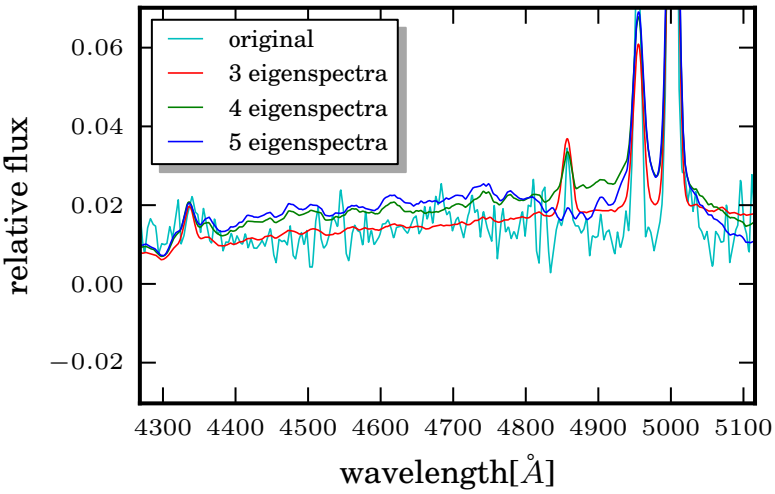


Figure 4.11: Projection of an example spectrum over 3, 4 and 5 eigenspectra. The continuum projection positions progressively too high w.r.t. the observed continuum, as the number of eigenspectra increases. Furthermore, for 5 eigenspectra (blue line) the reconstruction also lacks of the Hβ line.

ple estimate of the power enclosed in each eigenspectrum

$$P(e_i) = \frac{\Lambda_i}{\sum_{i=1}^{tot} \Lambda_i}, \tag{4.8}$$

where Λ_i are the eigenvalues of the correlation matrix. I find that the first three eigenspectra hold $\sim$ 90.6% of the total power; the first contains $\sim$ 87.3%, the second $\sim$ 2.5%, the third $\sim$0.7%, and from the fourth on the power content starts to decrease rapidly with respect to the first three, see Table 4.1. The variance in each component is a measure of the information content and I can conclude that three eigenspectra are enough to describe the sample in a statistical sense. However, I will show that this measure of information does not translate directly to the physical information contained in spectral features, as anticipated in §4.1. For example, I found that the slope of the continuum is well described by just a few eigenspectra, but this is not true for the line features (at least for the more intense ones) (Fig.4.10). The information on the lines in some cases is contained into higher-order components, that I neglect to avoid the noise, even though I recognize that this information is essential for understanding the physical properties of galaxies. But, as shown in the example Fig. 4.11, alredy adding the 4th eigenspectrum to the reconstruction, the resulting continuum starts to detach from the observed one; and since the shape of the continuum is the major contibution to the classification (see later in Chapter 5), I want it to be as adherent as possible to the observations. Furthermore, due to the twisting of the continuum, the shortest emission lines may happen to be canceled in the reconstruction.

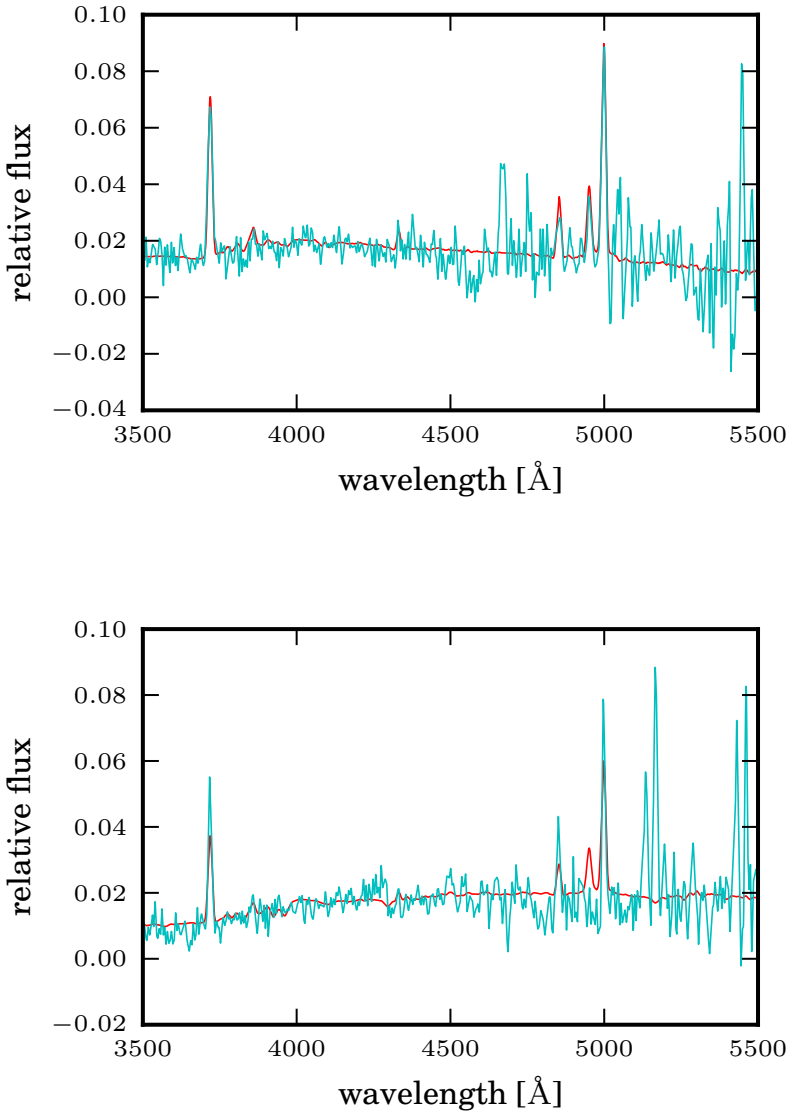


Figure 4.12: Two repaired and cleaned VIPERS spectra (red) superposed to themselves after the only repairing process (cyan). Our projection method is statistically able to recover the realistic emission and absorption features together with the slope of the continuum. As shown in the figure, in some cases the intensity of the line features is not fully realistically recovered. This is a consequence of the combination of "cleaning", operated by the description of the spectra through the first three eigenspectra, which do not reflect the noise of the sample, and least-square fitting with introduction of penalty terms in the regions of the lines.

Power of the first three eigenspectra	$\sim 90.56\%$
First eigenspectrum	$\sim 87.30\%$
Second eigenspectrum	$\sim 2.54\%$
Third eigenspectrum	$\sim 0.71\%$
Fourth eigenspectrum	$\sim 0.17\%$

Table 4.1: The power contained in the first four eigenspectra.

4.5 Cleaning the spectra from noise

After the repairing process, by projecting the VIPERS spectra on to the basis of three final eigenspectra I can achieve the goal of cleaning the spectra from noise, as illustrated in Fig. 4.12. This result is guaranteed by the fact that the first three eigenspectra are affected very little by noise. The same simplification offered by the PCA in using only three components makes it impossible, though, in my specific case, to naively apply Eq. (2.4) to recover properly VIPERS spectra. In fact, as for the repairing process, the projection on to only a few components is not guaranteed to reproduce spectral features matching the data. And again, as for the repairing, the projection can invert lines or add lines not present in the data. These errors arise because additional components are needed to recover all the lines accurately. I find that about 5% of spectra show unphysical line feaures once projected on to 3 components only. The situation, as said, could be improved by adding more components to the projection; however, this will re-introduce noise and artefacts, again degrading spectral features.

I can arrive at a compromise by assigning greater importance to the physical recovering of emission lines. This is precisely what was done in §4.1 where penalty terms were added in the least-squares minimization procedure to find the best-fitting, but physical repairing. I adopt this routine again in the final step to project each spectrum. The safeguard of the physicality of spectra is constrained imposing that the continuum is positive and the [OII], Hβ and [OIII] lines are not inverted. By comparison of the equivalent width of the [OII]-[OIII]-Hβ lines in the repaired and projected spectra to the same features in the original spectrum, I find that the lines, on average, are recovered with a precision of $\sim 80\%$ (Fig.4.10 for the case of [OII]). This is in agreement with the results found by Yip et al. (2004) for the majority of SDSS spectra in their analysis with 3 eigenspectra. For the reconstruction of the problematic emission line spectra only, they chose instead to use 10 eigenspectra, obtaining an error on the recovering of the lines of order 15-25%. Finally, the final quality of the repairing in my analysis, after the penalty has been applied, doesn't show any clear correlation to the portion of gaps in a spectrum, even if larger gaps easily increase the possibility of unphysical reconstructions at first step.

CHAPTER 5

Isolating populations of galaxies

5.1 PCA spectral classification

The Principal Component Analysis on spectra in a sample produces a set of 3 eigenspectra (Fig.4.8), of which every spectrum can be expressed as a linear combination, according to a set of eigencoefficients. These eigen-coefficients, namely a_1, a_2 and a_3, form an optimal basis in which to classify the spectra. In particular the a_1 coefficient, being related to the first eigenspctrum, is an indicator of the "redness" of the continuum, while the a_2, being related to the second eigenspectrum, is an indicator of the "blueness". So these two parameters together contain important information on the shape of the continuum, but also on the presence of line features, containing both those eigenspectra pretty strong lines. The a_3 eigencoefficient is instead related to the third eigenspectrum, which shows a fairly flat (though very slightly "bluish") continuum, and evident emission lines, so, for fixed continua shapes, it may be a stronger and more continuum-disentangled indicator of the intensity of line features.

To further reduce the parameter space of the eigencoefficients to a non-degenerate basis, and express them in a more convenient form, we compute the related Karhunen-Loève angles (ϕ-θ hearafter) (Connolly et al., 1995; Karhunen, 1947; Loève, 1948), so defined:

$$\phi = \tan^{-1}\left(\frac{a_2}{a_1}\right) \tag{5.1}$$

$$\theta = \cos^{-1} a_3 \tag{5.2}$$

These two angles, ϕ and θ, fully parametrize the three dimensional space, because, owing to the normalisation constraint, the coefficients fall on the surface of a 3D sphere. A more detailed discussion of the respective roles of ϕ and θ is presented, later on in the Chapter, in §5.3.

5.1.1 Comparison to template spectra

I am interested in exploiting the ϕ-θ parameters in order to classify the spectra on the basis of their position on a $\theta - \phi$ scatter plot.

To pin down the location of the different galaxy types on the $\phi - \theta$ plane, I will take advantage of the same group of B-C model spectra from which I picked the templates used to test the repairing routine (keeping also the blue galaxy representatives, although these are not fully realistic because of the lack of emission lines). I then project them on the three VIPERS eigenspectra and obtain their relative ϕ-θ angles. The distribution of the ϕ-θ parameters for the VIPERS sample, together with the one for the B-C models, is shown in Fig. 5.1. The VIPERS points are coloured according to colour scale displayed

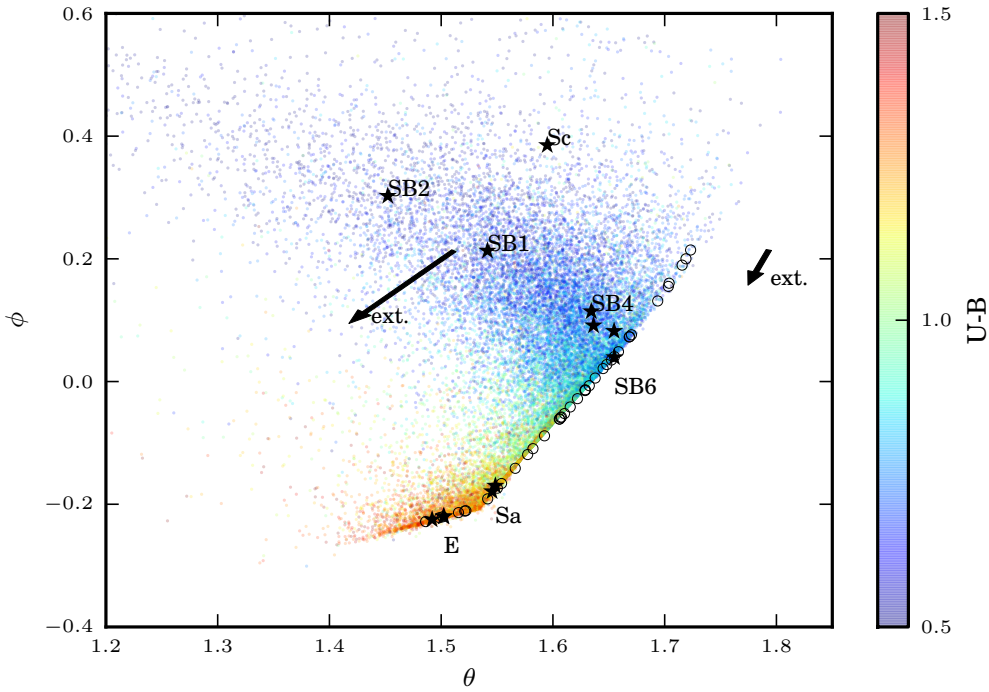


Figure 5.1: The ϕ versus θ plot, for VIPERS repaired and cleaned galaxies, with the position of Bruzual-Charlot and Kinney-Calzetti model galaxies overplotted. The colour gradient of the points from red to blue through green represents the $U - B$ rest frame color of each galaxy in the sample. The sequence of circle markers represents the B-C models ranging from the reddest (early-type) to the bluest (late-type) continuum slopes (see Fig. /reffig:move). The Kinney-Calzetti templates (star markers) are labelled with galaxy type. The early type galaxies are positioned with the early-type B-C templates, while the starburst templates are found in the middle (see Fig. /reffig:sequence for an idea of how mean spectra look like for starburst galaxies). The sharp edges in the distribution on the right hand side arise from constraints applied in the PCA reconstruction. Finally, the arrows show the effects of dust extinction for the two sets of models, with $A(V)$=1 mag and R_V=3.52.

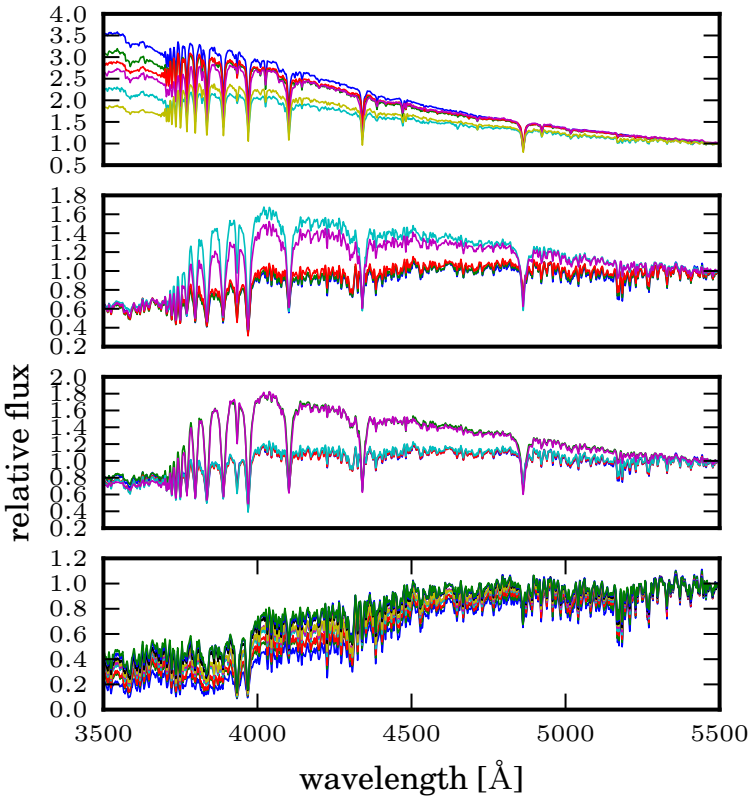


Figure 5.2: The B-C spectra corresponding to the circles in Fig. 5.1: red templates (bottom) lie in the low-ϕ region, with intermediate templates instead occupying the range $-0.2 < \phi < 0$ (middle boxes), and bluer ones lying at the top of the ϕ-θ plot.

at the right side of the figure, in agreement with their rest frame (U-B) color, and with what is suggested by the positions occupied by the B-C points on the same plot. The earliest-type models lay on the bottom, and they position upper and upper in ϕ as they become bluer, while the reason of their peripherical position on the plot depends mainly by their complete lack of emission lines, and will be clarified later.

Thus, analyzing the ϕ-θ plot, I can state that the redder galaxies lie towards negative values of ϕ and quite small values of θ, while, as ϕ and θ increase, the galaxies become bluer (Fig. 5.2), as suggested by the $U - B$ rest-frame color of VIPERS galaxies. Since an increase in ϕ is equivalent to an increase in a_2, this means that the bluer galaxies are represented by larger values of a_2 (and viceversa for the redder ones). This was expected, since the shape of the second eigenspectrum is the one that most resembles the spectrum of a blue galaxy. I do not consider now the first eigencoefficient a_1, because, being related to the first eigenspectrum, which is the average of all the spectra, it is not a significant discriminator by itself. Let us remark again, though, that I am basing this interpretation on a set of model spectra that do not present emission lines, although they do trace the continuum of blue galaxies in some cases. So they give a general idea of the arrangement of different spectral types on the ϕ-θ plot, but they are not apparently able to span the full distribution.

To get still more quantitative information on how galaxies spread on the ϕ-θ plane, I performed the same comparison with a model, by using the Kinney-Calzetti templates (Fig. 5.1). These are the same I used in §4.3 to build the synthetic spectra for the test, together with the B-C red-intermediate spectra.

From Fig. 5.1 it's clear that the K-C templates provide confirmation that the earliest type galaxies are at the bottom of the ϕ-θ plot, as suggested by the bulge and elliptical K-C templates. Additionally, the K-C-Sa and K-C-Sb spiral galaxies fall near to the region of intermediate B-C models, consistent with them presenting a certain level of star formation. The starburst galaxies, instead, follow a branch which is nearly orthogonal to the trend followed by red and intermediate galaxies. Finally, the K-C-Sc template occupies the highest position in ϕ in the plot, due to the steepness of its continuum, and it is more shifted towards lower values of θ with respects to B-C models: this fact, together with the appearance of a starburst branch winding up towards smaller values of the θ parameter, suggests that the intensity of emission lines gets stronger, at least at fixed values of ϕ, for smaller value of θ. This hypothesis will take more shape in section §5.2. So I can state that the two ϕ-θ parameters are related to the age and to the star-formation-rate in a rather complex way: an age sequence can be observed moving along the direction of the ridge of normal galaxies, marked by the B-C models trend at the right edge of the ϕ-θ plot, while an instantaneous star formation sequence can be observed on the perpendicular direction.

The peculiar shape of the ϕ-θ cloud obtained for the spectra, presenting sharp boundaries in the bottom and right region of the plot, deserves some further insight.

The sharp boundaries are a direct consequence of the application of the least-square penalty terms, introduced in the projection of the sample over the eigenspectra basis, together with the limits imposed by our two-components parametrization. These two boundaries impose limits on forbidden regions beyond which the reconstructions would be unphysical, with negative continua or inverted emission lines, due to the possible lack of information of the chosen components, if the penalty was not applied. Consequently, spectra with no emission lines are found right at these edges of the cloud of points, as demonstrated by the same position of the B-C models.

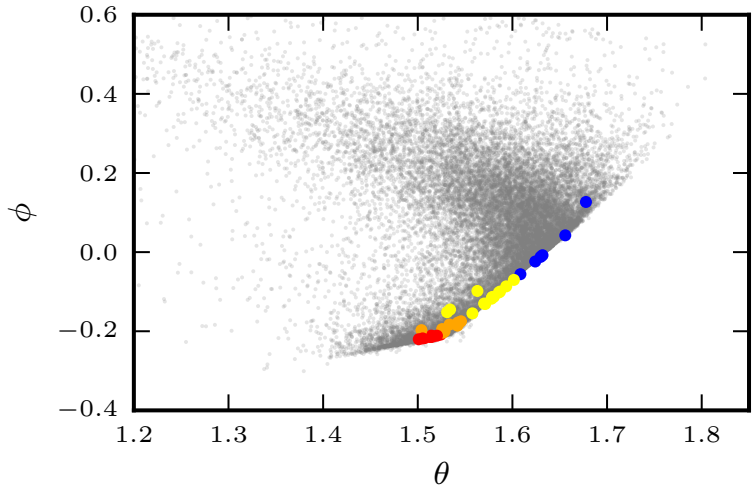


Figure 5.3: The set of 38 SDSS templates by Dobos et al. (2012) as projected on the VIPERS eigenspectra. The templates roughly follow the evolutionary track marked by the right edge of the ϕ-θ plot, apart from 3 templates that present stronger emission lines in the red part.

5.1.2 Comparison to SDSS LRGs

I compare here the distribution of VIPERS galaxies to SDSS galaxies on the ϕ-θ plot. To this purpose, I used a set of 38 SDSS templates computed through a PCA projection by Dobos et al. (2012). The templates were first re-binned on the same wavelength scale of VIPERS data, and normalized through their scalar product. They were then simply projected onto the VIPERS first 3 eigenspectra with the same routine discussed earlier.

The SDSS templates fall in the region at the right edge of the plot, following the same track found for the other datasets. In particular, the majority of them can be found near to the right sharp edge, because their PCA projection over the VIPERS first 3 eigenspectra was finding unphysical solutions for the line features and needed the χ^2 penalty to be applied. The colour gradient, from red to blue, gives a qualitative idea of the colour of the relative template (Fig. 5.3). Only a group of three spectra seem to detach from the main branch, positioning in a region of slightly smaller θ. The reason for that, as expected, is that those spectra present slightly stronger emission lines, mainly in the red part, than all the other SDSS templates. This is another tip that PCA proves actually much more sensitive to the slope of the spectra than to emission lines, in positioning the objects on the ϕ scale. In fact, although the blue SDSS templates present strong emission lines, their slope is flatter than many VIPERS blue galaxies, causing the templates to hardly produce related large numbers in ϕ.

5.1.3 Dust extinction

A natural question I can now ask about my classification regards the effects of dust extinction on the position in the ϕ-θ plot. Indeed, the precence of dust may represent an important interference in relating the observed color of a galaxy to its proper spectral

type, and the ϕ-θ plot, neglecting the comparison to the models, has been visually staggered, for the moment, on the basis of the galaxy colours. Since the dust attenuates the radiation originated by the source, it may produce a reddening in the observed overall color of a galaxy; and since blue light is much more affected by this effect than red one, I expect the reddening is expected to be more conspicuous for later-type galaxies, that may be taken for earlier-type objects by the simple recording of their color.

To determine how this effect may affect the ϕ-θ distribution, I applied an extinction law to the model templates. Since my purpose is only to check the direction to which extinction moves the galaxies in the ϕ-θ plot, I choose to apply the same simple Cardelli-Clayton-Mathis extinction laws (Cardelli, Clayton and Mathis, 1989) to all galaxy types, over the optical-near infrared wavelength range ($3000\text{Å} \leq \lambda \leq 9000\text{Å}$), which contains the rest frame range I am considering for my VIPERS data. The parameter $R_V[= A(V)/E(B-V)]$, with $A(V)$=1 mag, is set to 3.52. The resulting extinction effects on the B-C and Kinney-Calzetti models are then represented by the arrows shown in Fig. 5.1.

Once the B-C models have been corrected for dust-extinction, they all shift towards the bottom of the ϕ-θ plot (Fig. 5.1), in the same direction marked by the B-C curve. This is consistent with a reddening of the continuum. For the Kinney-Calzetti templates, and in particular for the starburst spectra, I find that dust extinction causes a larger shift within the ϕ-θ plot than for B-C spectra, probably due to the fact that young or starburst galaxies have a higher gas content, which is yet source of extinction; this explains also why the points in that region of the ϕ-θ plot display a broader distribution w.r.t. to the points n other regions: because of the higher gas content of the galaxies represented in that region, extinction causes larger shifts, both in the intensity of emission lines and in the slope of the continua.

5.2 The group-finding analysis

In this section I will better explore the diversity of the same VIPERS spectra represented on the ϕ-θ plot.

To do that, I choose to apply a k-means group-finding algorithm, a mathematical tool that is able to partition a space of objects, conveniently distributed in this space on the basis of their peculiar features, into maximally diverse classes (Ascasibar & Sánchez Almeida, 2011).

In general, one has n data points that have to be partitioned into data classes. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μ_i,i=1...k of the clusters that minimize the distance from the data points to the cluster. K-means clustering finds the set of points for which

$$\Sigma_{i=1}^{k} \Sigma_{\mathbf{x} \in c_i} d(\mathbf{x}, \mu_i) \tag{5.3}$$

is minimized, where c_i is the set of points beloning to the $i-th$ cluster and $d(\mathbf{x}, \mu_i)$ is the square of the Euclidean distance:

$$d(\mathbf{x}, \mu_i) = ||\mathbf{x} - \mu_i||^2 \tag{5.4}$$

This problem is not trivial, so the K-means algorithm only "hopes" to find the global minimum, possibly getting stuck in different solutions for the same data point.

The k-mean algorithm is used to solve the k-means clustering problem, and works according to the following diagram:

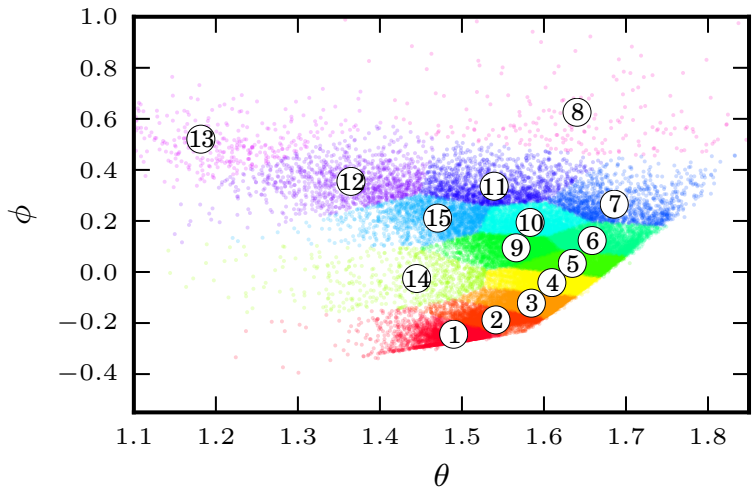
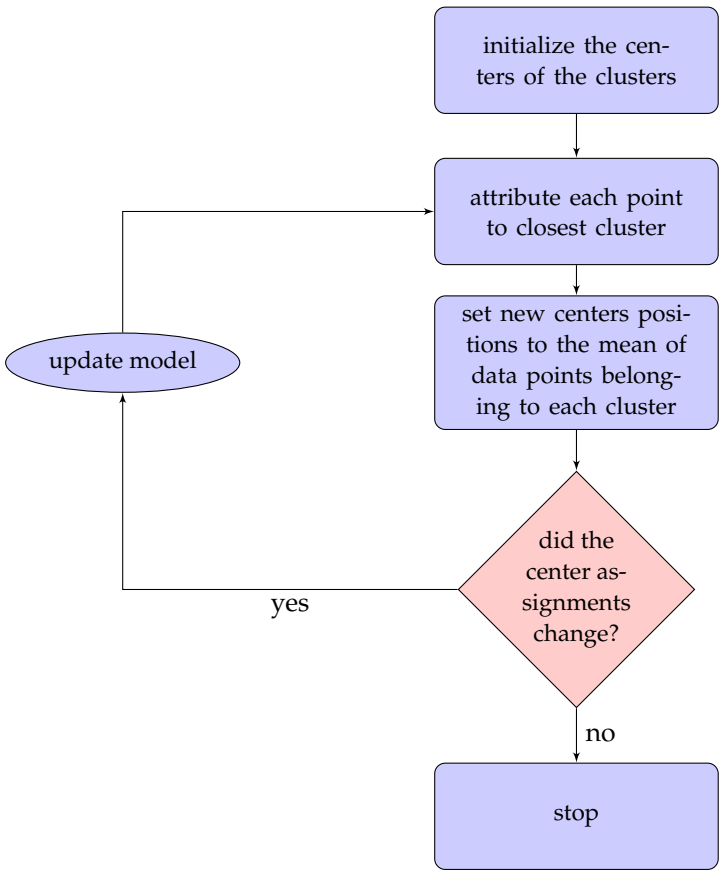


Figure 5.4: ϕ-θ plot of VIPERS repaired and cleaned galaxies, labelled with numbers 1-15, that represent the diversity of spectral types. The primary locus is traced by markers 1-8, and we find a secondary branch, marked 9-13. The mean spectrum at each marker is plotted in Fig. 5.5.



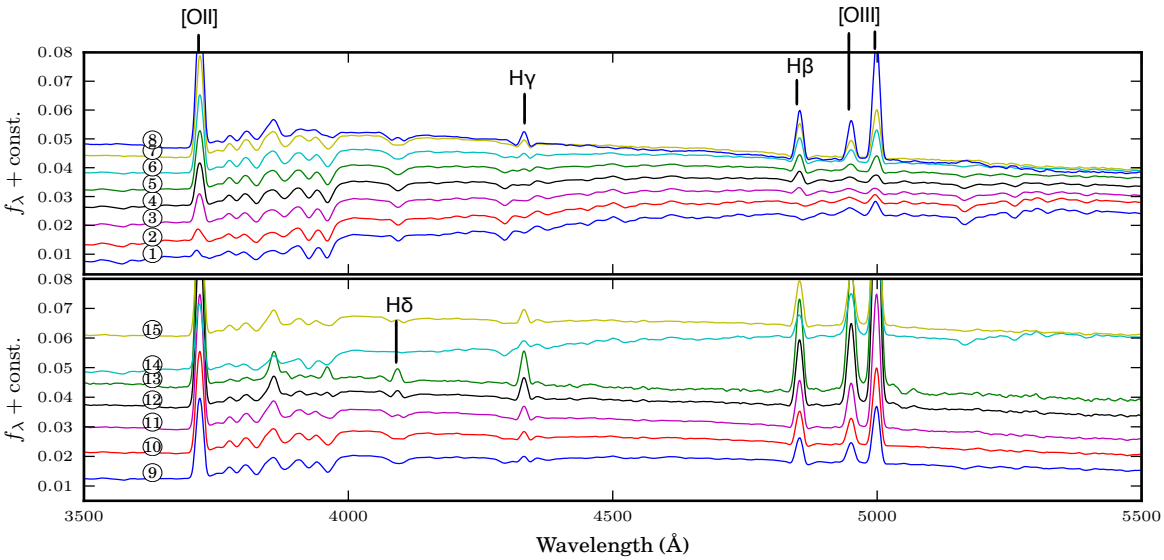


Figure 5.5: Representative average spectra obtained by grouping the VIPERS spectra through a group-finding algorithm into 15 classes in the (θ, ϕ) plane, as labelled in Fig. 5.4. We average the repaired and cleaned spectra (i.e. considering only the three principal components). In the top frame, we show that spectra 1-8 follow a sequence from early to late types, with the continuum becoming progressively bluer and with stronger [OII] emission. Note that the spectrum labelled as 1, i.e. the reddest one, still presents a hint of emission lines (although pure red spectra exist in the sample), since it is an average spectrum. In the bottom frame, spectra 9-13 represent starburst galaxies with flatter continua and strong emission lines. Mean spectra 14-15 effectively seem to pertain to none of the two branches, showing a mixture of blue and red galaxy properties.

The number of clusters should match the data. An incorrect choice of the number of clusters could invalidate the whole process. An empirical way to find the best number of clusters is to try K-means clustering with different number of clusters and measure the resulting sum of squares, or simply randomly partitioning the dataset.

The group-finding algorithm is suitable to be applied to the ϕ-θ distribution of points.

Through this method, the ϕ-θ plot is divided into an arbitrary number of groups, and galaxies are associated with any of these groups, on the basis of their distance, in the $\theta - \phi$ coordinates, from the center of every group, i.e.: each galaxy is associated to the nearest group center. It is necessary, as said, to specify the number of groups beforehand, and I chose 15, since, after some tests, this number appears to be sufficient to span all features visible by eye, without repetitions.

The positions of the classes I have identified are marked in Fig. 5.4. These points visually trace out essentially two branches, that can be thought of as the skeleton of the data cloud, and that immediately recall the two traces suggested by the model spectra. The first branch, marked by the numbers 1-8, shows a sequence very similar to what we can imagine as the prosecution of the B-C red and intermediate models discussed previously, encompassing, though, also the galaxy types 3-4-5-6, that before seemed already more likely pertaining to the starburst branch. In particular, the K-C Sc template would appear to lie between the 7 and 8 classes. A second branch, whose primary direction is marked by 9-13, including group 15, lies almost perpendicular and passes through the K-C starburst 1-2. Group 14 marks an intermediate position.

The mean spectrum that represents each class is plotted in Fig. 5.5. In particular, in the top panel of Fig. 5.5, it becomes clear that moving from 1 to 8 means an increase in the intensity of emission lines and a change in the slope of the continuum, from redder to bluer. In the bottom panel of Fig. 5.5, mean spectra from 9 to 13, pertaining to the perpendicular "starburst" branch, show only an increase in the intensity of emission lines, particularly evident also by looking carefully at the Hγ, Hδ and Hβ emission, while the slope of the continuum is substantially unchanged.

In general consecutive numbers here label very similar average spectra in almost all cases, apart from spectra 14-15, which do not resemble the prosecution of spectrum 13. Actually, because of its position on the plot, 15 is instead very similar to 11. Mean spectrum 14 instead, lying beyond the imaginary starburst branch in Fig. 5.4, actually doesn't follow the trend of that branch, but shows a redder continuum, in agreement with its ϕ position on the ϕ-θ plot, though presenting important emission lines. It looks more similar to mean spectra 3 and 7 respectively, but for the intensity of emission lines, since it exhibits stronger line features. The combination of red continua and strong emission lines shown by mean spectrum 14, makes it hardly includable in any of the two branches; it maybe perhaps associated to a post-starburst phase, or, more naively, to the bad reconstruction (by definition) of the non-regular spectra; the latter hypothesis, though, is not fully founded, since I don't find in this region any clustering of this kind of objects, which conversely spread out all over the ϕ-θ cloud (including the 14 region).

Deepening the insight into the two branches, the 1-8 branch can be associated to the integrated star formation of the spectra: the position of a galaxy on the branch is an indicator of how much star formation the object has undergone, from its formation to the time of the observation. Conversely, the starburst branch gives an indication on the instantaneous star formation: the nearer to the 1-8 branch, the lowest instantaneous star-formation, viceversa for the opposite extreme objects. Furthermore, moving leftwards, not only the intensity of the emission lines increases, but the ratio of H$\gamma/H\beta$ is mantained almost constant. Since this ratio is an indicator of the presence of attenuation, due to the presence of dust, which is typical of gaseous, star forming regions, it's clear

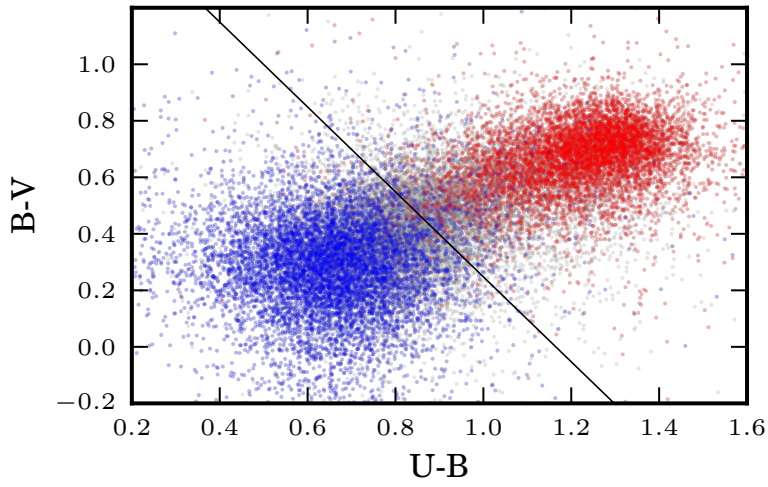


Figure 5.6: The rest frame $U - B$, $B - V$ colours of VIPERS galaxies. Red points have PCA parameter $\phi < -0.1$ and blue points have $\phi > 0.01$ (intermediate values of ϕ are coloured grey). The line dividing the two samples optimally separates $\phi > 0$ from $\phi < 0$ in colour space with a contamination of $\sim 13\%$.

that moving leftwards on the starburst branch, i.e. raising the intensity of the starburst, indeed means leaving the dust attenuation unchanged (of course, for the optical emission that it's still detectable). Thus even if there is more dust attenuation in the starburst branch than in the perpendicular one (Fig. 5.1), this attenuation seems almost independent from the intensity of the starburst within the starburst branch: this means that the position on the starburst branch is not dominated by extinction, but it's fully dependent by the amount of star formation.

This analysis reinforces the intuition that, while moving upwards in the ϕ direction in the ϕ-θ plot can be associated to a change in the slope and the intensity of the lines, moving from right to left in the θ direction also means a strengthening in the intensity of the emission lines.

The shape of the mean spectra for the different groups and the position of the same groups on the ϕ-θ plot reinforce the evidence that galaxies can be to split into two nearly orthogonal spectral sequences, of which one reflects the evolutive phases of a normal galaxy (though not being an evolutionary track), while the other describes the starburst phases. This suggests a route for building a physical classification of the spectra based on the ϕ-θ parameters.

5.3 Comparison to other classification methods

It's interesting to compare side by side the PCA classification against more familiar ones: one example is the classification based on rest-frame broad-band photometric colours. In Fig. 5.6 I plot the VIPERS rest-frame $U - B$ and $B - V$ for each galaxy (Bolzonella *et al.*, in prep, Fritz *et al.*, in prep.). I divide the sample into red and blue classes using the angle

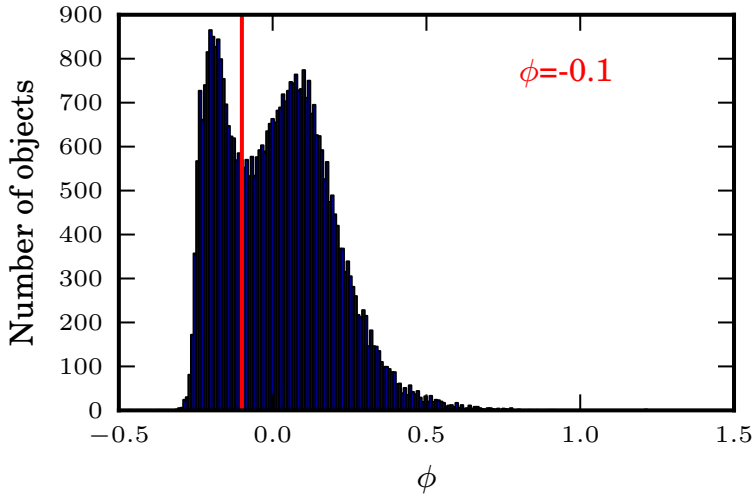


Figure 5.7: Histogram for the distribution of ϕ, with the ϕ=-0.1 threshold I chose to define the red sample.

ϕ. Based on the comparison to the model spectra and the discussion of the previous sections, a reasonable definition of the red class can be ϕ <-0.1 (Fig. 5.7), with the very blue galaxies confined at ϕ >0.01. In this way, I cleanly exclude intermediate types.

For comparison, I construct a red-blue classification using the $U - B$ and $B - V$ colours, that matches as well as possible the PCA selection. This is shown in Fig. 5.6, where the two classes defined through the ϕ-θ angle are plotted in blue and red and the intermediate types in grey. I clearly note that the PCA selection is correctly capturing the bimodal distribution. Conversely, I want to verify how a crude color-color selection performs, with respect to that based on the spectral information "compressed" into the PCA parameters. I therefore separate photometrically red and blue classes by tracing a line perpendicular to the axis connecting the centres of the two clouds, (Fig. 5.6). This axis is defined by computing, through the simplest two-dimensional PCA, the two eigenvectors of the distribution of points on the colour plane: the first eigenvector marks the principal direction of the data, while the second is orthogonal to the first one. Here the total number of eigenvectors is only two, since the correlation matrix of a two-dimensional distribution has dimension 2. The position of the line is set such that there is an approximately equal number of contaminating galaxies on the red and blue sides. With respect to the PCA classification, I find that: (1) in selecting red galaxies, the color-color selection has a $\sim 14\%$ contamination of spetroscopically blue galaxies and an $\sim 88\%$ completeness; (2) for photometrically blue galaxies, the contamination of objects that spectroscopically are classified as "red" is $\sim 12\%$ and the completeness is $\sim 86\%$.

I also perform a comparison between PCA and a time-evolving passive-active separation (Fig. 5.8), derived in Fritz et al. (2014), after converting our AB magnitudes into Vega magnitudes (Blanton et al., 2003), for which the relation holds.

The comparison of PCA versus this selection, gives the response: in selecting red galaxies, the PCA selection has a ∼20% contamination and a ∼86% completeness with

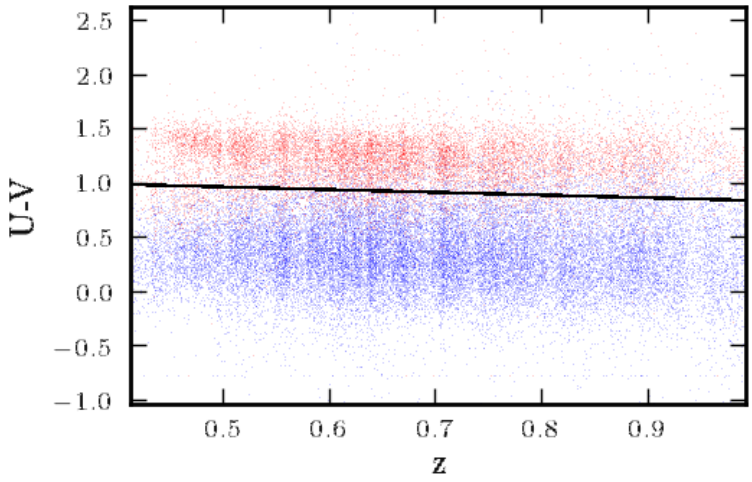


Figure 5.8: PCA based (red and blue points) and redshift-color separation of red and blue galaxies (black line)

respect to the photometric one; while for blue galaxies I obtain a ∼6% contamination with a ∼92% completeness.

It is encouraging that in these simple cases of classifying galaxies as red or blue, the two methods produce very similar results. The strength of the PCA approach though is that it encodes additional information about spectral features that is not available in the broad band photometry. Moreover, photometric filters have a sensitivity of order ∼1000Å, and they are not sensitive to details smaller than this wavelength range. PCA instead, even if it's "filtering" the entire spectrum to a two-parameters representation, is still sensitive to the large scales (shape of the continuum), and to the smaller ones, as the line features. Thus, with only two parameters, I mantain the capability to distinguish effects pertaining to differents scales in λ, as for example the instantaneous star formation, from the time integrated one (as shown in §sec:groupfinding).

A further step deepen the understanding of the ϕ-θ distribution, has beem to colour the ϕ-θ scatter plot with a a gradient based on the 4000Å break strength (Fig. 5.9), and one based the [OII] line equivalent width (whenever the measurements of these two quantities are meaningful) (Fig. 5.10), the values of which are present in the VIPERS database. In Fig. 5.9 I notice that the ϕ-θ coefficients separate the different intensities of the 4000Å break in near-to-horizontal lines, parallel to the θ axis. This means, as expected, that the θ parameter is the less sensitive to this feature, which, conversely, results almost totally gathered into the ϕ parameter. One has to keep in mind, though, that even if it contains the majority of information on the continuum slope, of course ϕ is also linked to the intensities of the line features. On the other hand, Fig. 5.10 reveals that the intensity of the [OII] line is marked in a more complicated way on the ϕ-θ plot, since the information on the line intensities is for sure recorded by θ (see Marchetti et al. 2013), but it is also generally linked, for a standard galaxy, to the shape of the continuum, which slope, as shown in Fig. 5.9, is more confined to the ϕ parameter. By looking at Fig.

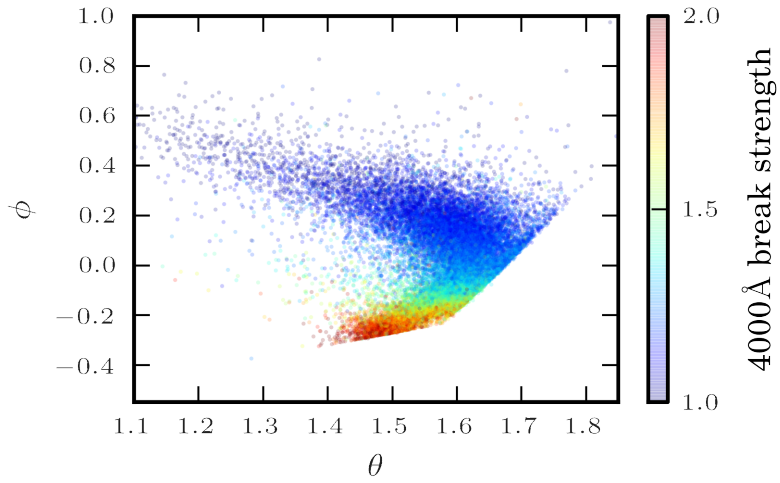


Figure 5.9: ϕ-θ plot with colour gradient based on the D4000 Å break intensity.

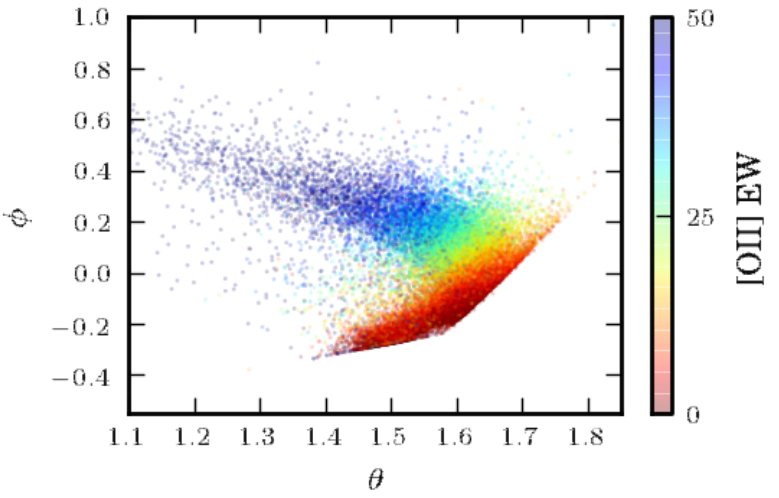


Figure 5.10: ϕ-θ plot with colour gradient based on the OII line equivalent width.

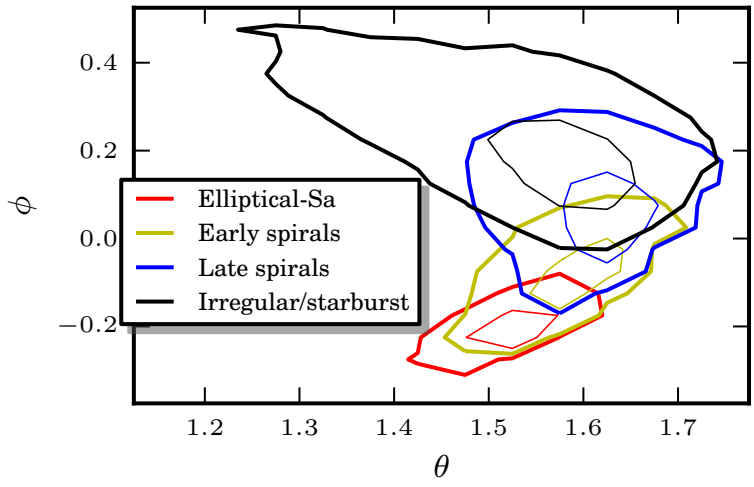


Figure 5.11: SED-type distribution contours on KL plot, at 95% and 50% levels.

5.5, and in particular at average spectra 3-14 and 6-15, the only thing I can state is that, again, at fixed continuum, smaller values of θ seem to represent stronger emission lines.

This was expected indeed, for what anticipated at the beginning of this Chapter: in facts the ϕ parameter comes from a combination of the two parameters related to the first two eigenspectra, presenting respectively a typical red continuum with emission lines, and a typical blue continuum with emission lines; this makes it sensitive to both the continua and the line intensities variations. On the other hand, θ is linked only to the third eigenspectrum, characterized by a rather flat slope of the continuum with emission lines, causing it to be pretty dull to continuum slope variations.

In the end, I also want to show how different Spectral Energy Distribution classes are displayed by the ϕ-θ plot. For this purpose I employ the SED template number assigned to each VIPERS spectrum. VIPERS data have in fact been assigned a number from 1 to 4 representing the template SED which fits best each of them. The four classes, encompassing many SED templates each, represent elliptical-Sa, early spirals, late spirals, irregular or starburst spectra.

The distribution of the SED types on the KL plot is showed in Fig. 5.11.

5.4 Galaxy evolution in the PCA parameters

Since the KL parameters, as stated earlier, contain informations on both the continua and the emission features of the spectra, I am interested in checking if they evolve with time, i.e. if ϕ and θ show an evolution comparable with what we expect, for both blue and red galaxies. In particular, I assume here that galaxies undergo a passive evolution, with blue galaxies becoming redder (also in terms of shape of the continuum) and exhibit lower emission lines, and red galaxies becoming redder and redder, as the time passes.

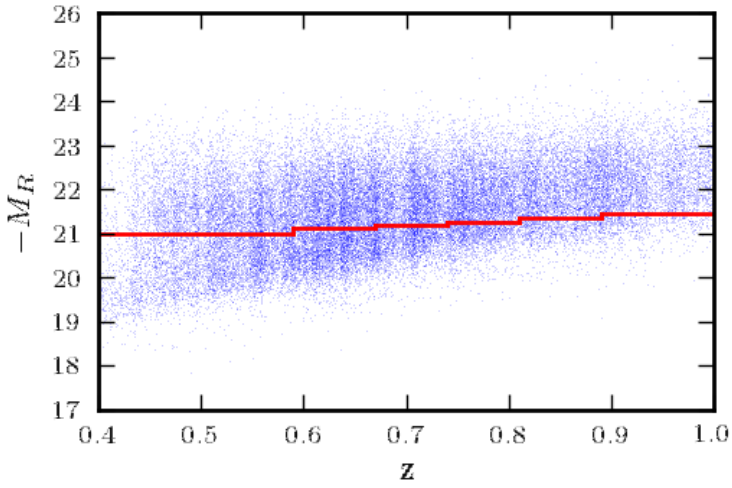


Figure 5.12: R band rest frame magnitude for VIPERS PDR-1. The red line marks the M_R luminosity cut.

As a first stage, I apply a naive step luminosity cut in the absolute R-band magnitude M_R, to mimick an evolution as a function of redshift, to the spectra I consider for this analysis (5.12). It's important, though, to point out here that I found, indeed, that changing the cut to a simple horizontal one, would impact the analysis in a negligible way).
Then I divide the red galaxies from the blue ones, excluding the spectra that lay in the green transition zone, as shown at the beginning of §5.3. I already explained that this is a spectral PCA-based subdivision that mimics a blue-red separation based on a color-color diagram.

At last, I divide the redshift space into redshift bins, characterized by approximately the same number of objects, and average the values of ϕ and θ relative to the red and the blue spectra within each bin (Figs. 5.13, 5.15, 5.14, 5.16). The errorbars, which would simply depict the statistical standard deviation from the average, are shorthest than or comparable to the size of the datapoints, thus I don't plot them in the figures.

The average values of ϕ in each redshift bin show an increasing (though very slightly) trend towards highest redshifts, which is clearer for the red sample (Fig. 5.15). For the blue one the position of the point in the range $0.6< z <0.7$ suggests a slight decrease, but the general trend is clearly increasing, even if, again, within a very short interval of values in θ. Since the ϕ parameter, as said, encloses the large majority of the information on the continua and part of the information on the line intensities, the trend in ϕ goes in the expected direction, at least for the red sample: at increasing z the continua are getting bluer and lines are getting stronger. This is in agreement with the prediction that at larger redshift there are more blue galaxies that in the nearer Universe, where the galaxies are older, thus more evolved and more red.
The interpretation of the behaviour of θ is instead, again, more complicated. The θ parameter shows a constant increase as a function of redshift only for the red sample, while

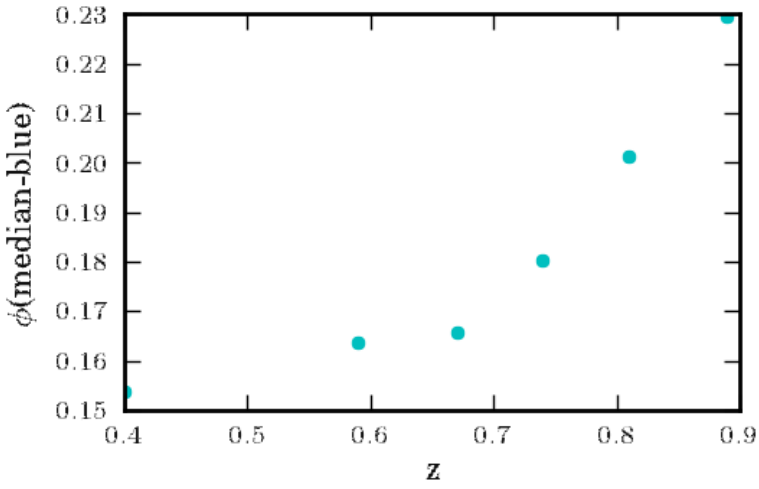


Figure 5.13: Average value of ϕ for the blue galaxies in different redshift bins

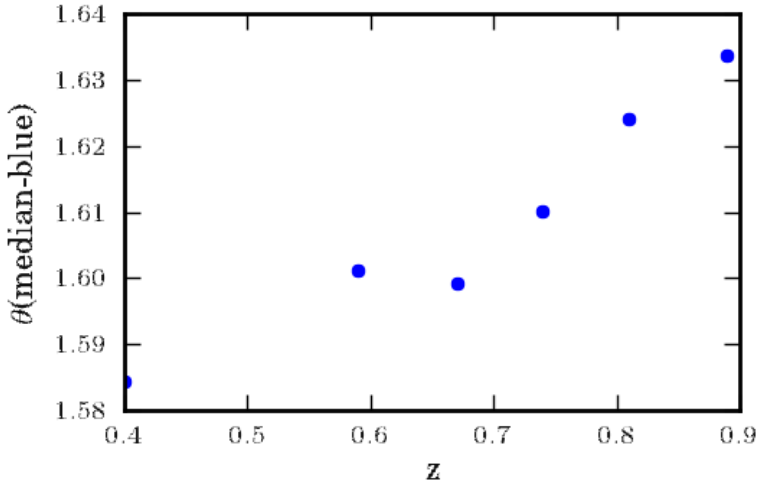


Figure 5.14: Average value of θ for the blue galaxies in different redshift bins

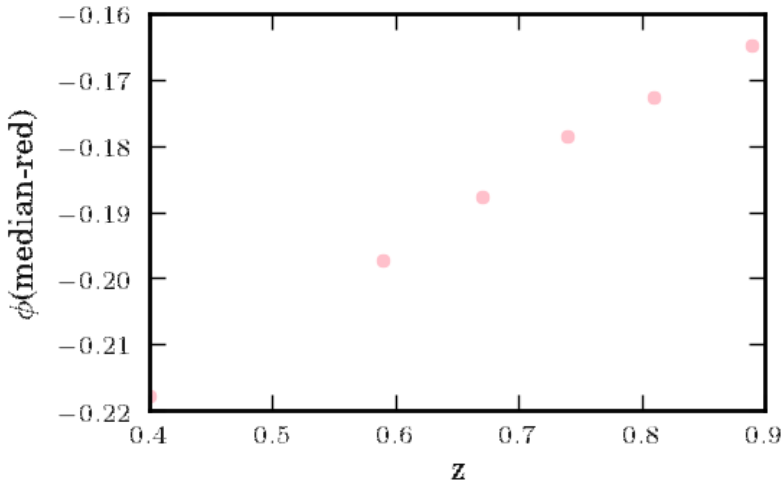


Figure 5.15: Average value of ϕ for the red galaxies in different redshift bins

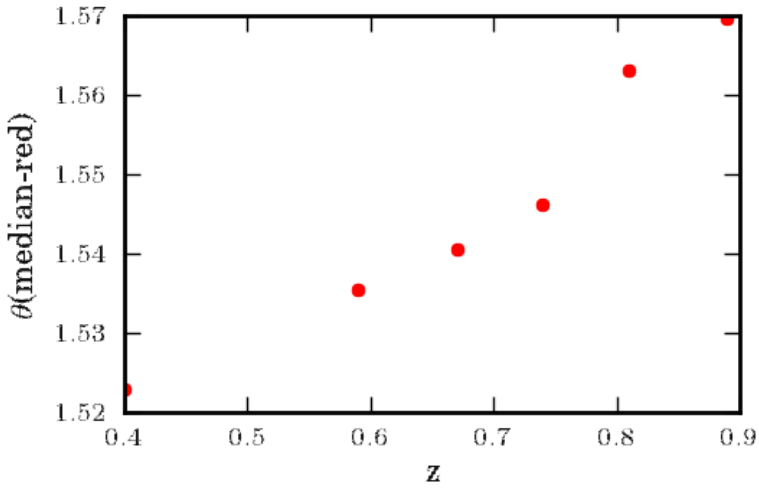


Figure 5.16: Average value of θ for the red galaxies in different redshift bins

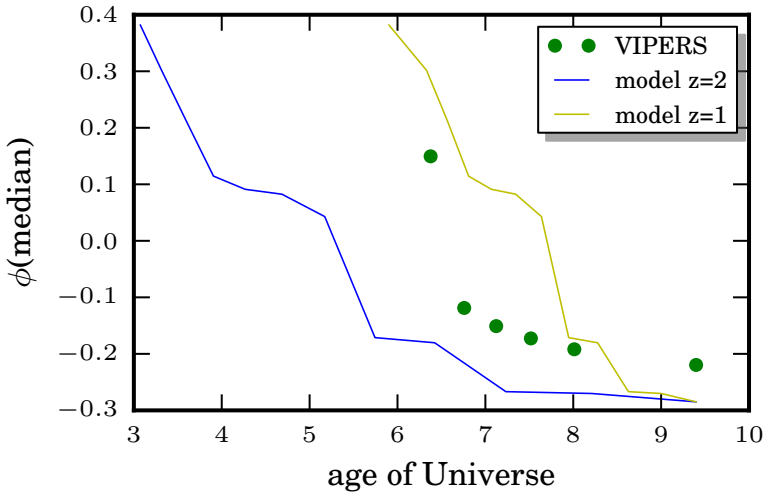


Figure 5.17: Median of the ϕ values as a function of age for VIPERS spectra (green dots), compared to the predicted evolution of a galaxy formed at z=1 (yellow line) and one formed at z=2 (blue line).

for the blue one it seems to be nearly constant in a narrow redshift range. Since θ has to do with the intensities of the line features, and in particular for highest θ the lines are getting weaker, this would suggest that the lines are weaker at larger redshifts, in contrast with the expected evolution and with the prediction of ϕ. But θ, as described in §5.3, cannot be interpreted as a pure indicator of the line intensities: in fact the small information on the continuum which is carried by θ, is in general more important than the one on the lines contained in θ itself, because of the pixel content of each. Viewed in this perspective, i.e. that θ is an indicator of the continuum more than of the line intensities (especially for the red galaxies, which contain very few emission lines), the increase in θ for increasing z represents an evolution of the continua in agreement with the expected galaxy evolution. For the blue sample, the effect of the opposite trends expected within θ for the emission lines and the continuum (increasing θ means weaker emission lines but bluer slope) is stronger, causing θ to exhibit a slightly oscllating beahviour in redshift, due to stronger presence of emission features than for the red sample. The line information contained in θ, instead, becomes dominant only for fixed slopes of the continuum, i.e. for fixed values of ϕ. Thus, as said, it is only within small stripes of ϕ values that smaller θs represent larger emission lines.
In any case, for both the samples, but surely more for the blue one, the trend of the ϕ-θ parameters as a function of redshift may have been very slightly smoothed by the repairing-cleaning process, that for some objects can recover a slightly shorter line feature, primarily for the χ^2 penalized spectra in the PCA process.

This trend obtained for the median of spectra in the ϕ-θ space, can be qualitatively compared with the evolution predicted by models, in particular with the one predicted by the B-C model. This comparison will only be qualitative, since it's hard to quantify it due to the reduced size of data uncertainties. To visually perform the comparison, I plotted the median values of ϕ and θ for the red and blue galaxies altogether, and

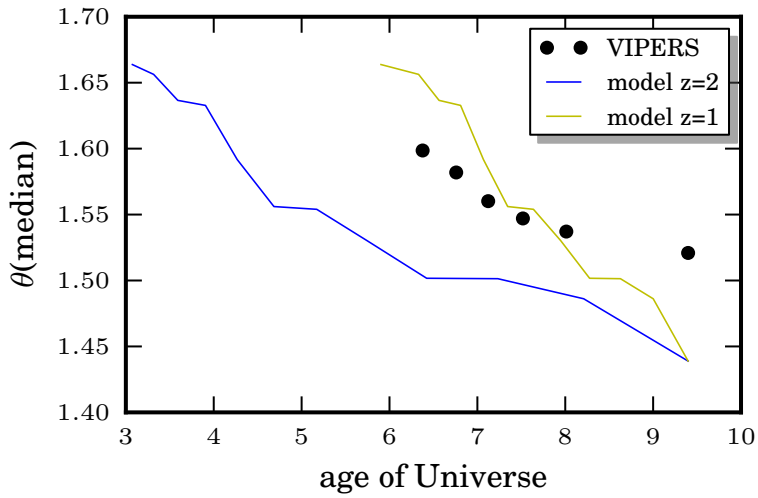


Figure 5.18: Median of the θ values as a function of redshift for VIPERS spectra (black dots), compared to the predicted evolution of a galaxy formed at z=1 (yellow line) and one formed at z=2 (blue line).

overplotted the predicted B-C evolution for a galaxy born at z=2 and one born at z=1 (Figs.5.17,5.18). The VIPERS points, being the median of ϕs and θs pertaining to galaxies of many different ages, lie approximately between the two models. At small redshifts, the ϕ of data suggest a median of galaxies which are less red than predicted by the models, which is expected, since the models do not contain emission lines, while the real data do, and models are not as blue as some spectra in the data, and this contributes in increasing the values ϕ.

The trend of θ for the data also seems to suggest, in agreement to ϕ, that real objects are bluer (or less red) than models (considering as prominent the link of θ to the continua).

For high redshifts both the data parameters seem to indicate an evolution in agreement with a population of younger galaxies, being nearest to the top line than to the bottom one. This behaviour is indeed a general trend for θ, while ϕ lies in the middle of the two models for intermediate redshifts. The trend for high redshift is also evidenced by a plot of the fraction of passive to active galaxies in the VIPERS sample (Fig. 5.19): according to the PCA-ϕ based blue-red separation, the population starts to be composed primarily of passive galaxies at z <0.8, while at z ∼1 (the extremity of my data distribution) it seems to be dominated by active galaxies.

To improve the resemblance of the data to the models, I also masked the emission lines in the data and performed the comparison again; unfotunately, the continua of the data are sometimes "recognized" by the PCA and repaired, so that the mask is patched with an emission line in the right place. Moreover, the data contain "bluer" galaxies (bluer shape) than the ones represented by the models. These two effects produce a similar distribution as Figs. 5.17 and 5.18 for the data points as a function of redshift/age, suggesting, again, that the major contribution to the value of the parameters is the shape

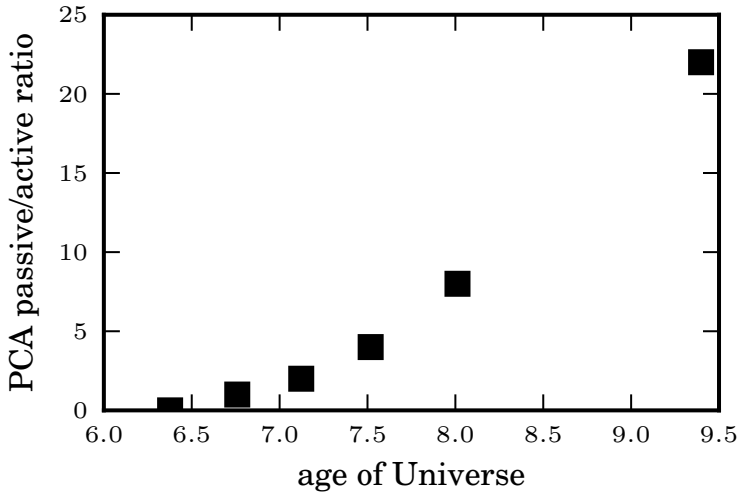


Figure 5.19: Fraction of passive to active VIPERS galaxies (according to PCA ϕ division) as a function of age.

of the continuum, more than the presence or absence of emission lines.

CHAPTER 6

Narrow Line AGN identification as PCA byproduct

6.1 PCA reconstruction of peculiar spectra

One of the limitations of the PCA reconstruction of spectra is that a spectral type that is represented by a few galaxies, only will be poorly (or even will be not) represented by the principal eigenspectra. Rare features will not be included in the main eigenspectra, but only in higher-order ones. This is for example the case of AGNs (as QSOs or Seyfert galaxies; it can be also the case of normal galaxies which have been assigned a wrong redshift). Their representation, in terms of the first three components only, will not be realistic. This will force them to resemble an intermediate, blue, or starburst galaxy. An example of this is shown in Fig. 6.1, where a broad-line AGN is reconstructed using only three eigenspectra. The continuum is approximately fitted, but the broad emission features do not have counterparts in the three basis vectors used, and are thus fitted with normal emission lines. Another interesting example, depicted in Fig. 6.2, is the projection of a NL AGN over the PCA eigenspectra: it noticeable that the reconstruction is more fitted to the observed spectrum for almost all the wavelengths, apart from the region of the [OIII] doublet and the [OII], where the reconstruction is not able to reproduce the strong intensity of the emission lines, typical of these kind of spectra. This effect sometimes arises also for normal starburst galaxies in my PCA reconstruction, but for the NL AGNs, this seems to be always the case. I have directly verified in my tests that AGN features start to emerge only when principal components up to orders $\gtrsim 50$ are included. This is due to the fact that AGNs are actually a minority in the VIPERS catalogue (they are expected to be $\sim 5\%$ of the total), so their peculiar features are treated as "noise" (i.e. uncommon features) by the PCA. For these reasons, as already hinted, the AGNs do not group as a separate population of outliers in the ϕ-θ plot computed with three or higher-order eigenspectra, but fall on the main locus in apparently random positions. For this reason PCA reconstruction of the AGN spectra will be better performed on an AGN-only sample, when a larger sample of AGNs only will be available. Nevertheless, the poor reconstruction of all the AGN-like galaxies can be exploited to pick them up authomatically in the sample, on the basis of the discrepancy of the reconstructions from the observations.

6.2 NL and BL AGNs characteristics

The AGNs, as already described in Chapter 1, are galaxies hosting an accreting black hole at their center. This is the common engine of their activity, but these objects present many diversifications, that have been tentatively assigned to a set of main groups and subgroups.

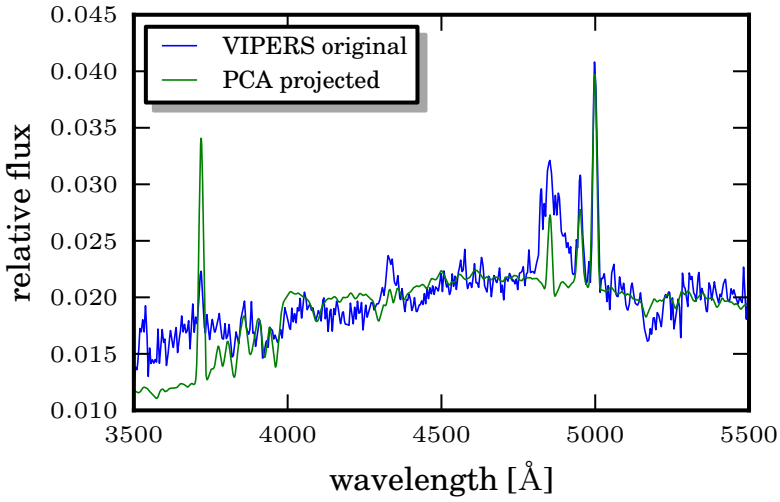


Figure 6.1: Example of a BL AGN in the VIPERS sample (blue) projected on to the PCA eigenspectra basis (green). The PCA reconstruction was not able to preserve the peculiarities of this rare spectrum, forcing it to resemble a typology of galaxy which is much more common within the VIPERS sample.

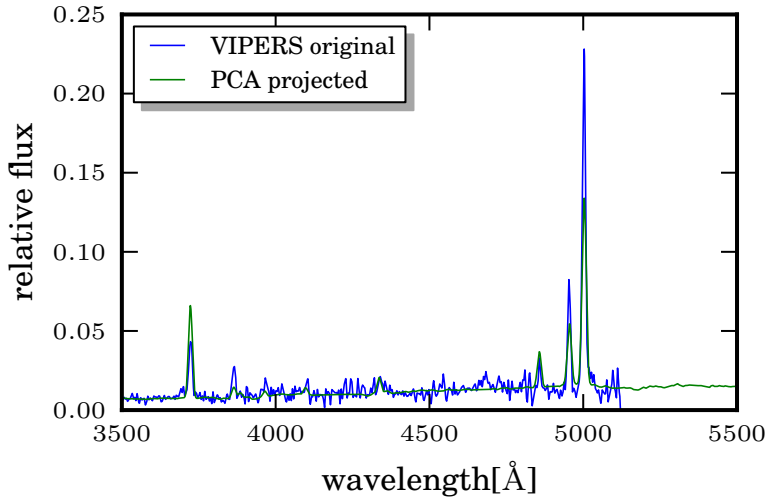


Figure 6.2: Example of a NL AGN in the VIPERS sample (blue) projected on to the PCA eigenspectra basis (green). The PCA reconstruction recovers pretty well the global characteristics of the spectrum, but for the intensity of the Hβ, [OIII], and slightly of the [OII] emission lines.

Seyfert 1s are normal spiral or elliptical galaxies characterized by a compact, star-like nucleus, and a nuclear emission line spectrum, characterized by broad (a few thousands km s^{-1}) permitted lines and narrow (a few hundreds km s^{-1}) high excitation lines. Seyfert 2s, instead, have a narrow emission line spectrum, similar to the one of Seyfert 1s, but they are lacking both the compact nucleus and the broad emission lines. Finally, Quasars (or QSOs) are simply high luminosity Seyfert 1 nuclei; their luminosity is so high that the host galaxy is difficult to detect. Seyferts and QSOs contain a compact nuclear continuum source ionizing a broad line region, surrounded by an optically thick torus of dust. Depending on the orientation of this torus with respect to the line of sight, the central object is seen or hidden; when it is hidden, we see only the narrow, extended emission line region typical of a Seyfert 2.

Many spiral galaxies contain a starburst region at their center, that would make their spectra similar to the one of a Seyfert 2 . But the spectra of narrow line Seyferts and starbursts are easily distinguishable by their main emission line ratios, on which the diagnostic, or BPT, diagrams are based (see §6.3). Nonetheless, diagnostic diagrams built with these emission line ratios revealed a third type of emission line spectra called Liners (Low Ionization Nuclear Emission line Region). Some of these objects are most probably low luminosity Active Galactic Nuclei (AGNs); some others must be related to a cooling flow phenomenon, occuring in clusters of galaxies, or are produced in the collision and merging of gas rich galaxies.

Seyferts and QSOs can further be divided into radio loud or radio quiet objects. Radio loud objects are always hosted by an alliptical E-type galaxy. Most radio galaxies have a double lobe structure; the high radio luminosity sources have edge-brightened lobes; they are called FR II radio sources (for Fanaroff-Riley type II). The low luminosity sources are called FR Is. FR II radio galaxies have the nuclear emission line spectrum of Seyferts; when they have broad emission lines they are called Broad Line Radio Galaxies (BLRGs); when they have the emission line spectrum of a Seyfert 2, they are called Narrow Line Radio Galaxies (NLRGs). All radio quasars have the samle morphology as FR II. FR Is have a weak low excitation emission line spectrum, very similar to Liners, or they have no detectable emission at all.

The lobes of radio galaxies (FR Is and FR IIs) are powered by a relativistic jet; when the angle between the jet axis and the line of sight is small, the jet is Doppler boosted by a large factor, and the whole spectrum (from radio to gamma-ray) is dominated by a compact, highly polarized, highly variable, superluminal, almost featureless continuum. These objects are called Blazars; they are divided into two subclasses: the Highly Polarized Quasars (HPQs) which show broad emission lines, and the BL Lacertae objects (BLLs) with no or weak broad emission lines. The parent population of the HPQs is made of the FR IIs, while the parent population of the BLLs is made of the FR Is.

As hinted, the most popular explanation for the AGN powerhouse involves accretion of gas onto a supermassive, perhaps spinning black hole (BH). Different regimes of accretion have been invoked to constitute the basis of a unified picture of AGNs. The predictions of the theory are that rotationally supported thin disks would form at lower accretion rates ($M<M_{Edd}$), while supercritical ($M\geq M_{Edd}$) accretion flows are expected to form thick disks supported by radiation pressure. A very subcritical flow may not be able to cool and, instead of forming a thin disk, it would puff up, giving rise to an ion torus supported by gas, rather than by radiation pressure. This is, very schematically, the generally accepted Unified Scheme of AGNs.

6.2.1 Seyfert II galaxies (NLAGNs)

Seyfert 2s are characterized by a spectrum having both strong high- and low- ionization emission lines. To explain both high and low ionization lines, the narrow-line region (NLR) must be composed of a mixture of dust-free, metal depleted clouds, with a radius-independent range in densities, distributed over a range of distances from the nucleus. To encompass the observed range of line intensities relative to Hβ, it is necessary to vary the spectral energy distribution incident on the clouds, by adding a varying contribution of a hot blackbody to a steep X-ray power-law. The width of the narrow emission lines correlates with the Hubble types in Seyferts, earlier types having broader lines. There is also a correlation between line width and nuclear stellar velocity dispersion, suggesting that gravitational motion plays an important role in the narrow line velocity field.

6.2.2 Seyfert I galaxies and QSOs (BLAGNs)

In addition to a narrow-line emission spectrum, Seyfert 1s have broad permitted lines (HI, HeI, He II and Fe II in the visible domain). Bright QSOs have relatively weak narrow lines. The relative amount of narrow-line emission decreases with luminosity: this suggests either that the ionizing flux to the narrow-line region is proportionately smaller in objects with a high-luminosity broad-line component, or else that there is proportionately less low-density gas in the higher luminosity Seyferts. The broad emission lines observed in AGNs have a FWHM which is typically in the range 5000-10000 km s^{-1} and show different kinds of profiles that usually are not Gaussian or even symmetrical. The half-width at zero intensity of Hβ can be extremely large, reaching 35 000 km s^{-1} in some cases. Hβ/Hα and He Iλ5876/Hβ ratios increase from the core to the wings of the lines, indicating that the broad-line region is not a thin spherical shell. Seyfert 1 BLRs have sizes of the order of a few light-days to light-months ($\sim$100 light-days). It is now widely believed that accretion of gas into a central supermassive BH lies at the heart of the phenomenon; the accretion flow takes the form of a geometrically thin disk which is the source of the X-ray, UV and optical continuum emission which ionizes circumnuclear gas in both the broad-line and narrow-line regions; the BLR is made of an assembly of small clouds, photoionized by the continuum emission of the disk.

6.3 NL AGNs separation methods

When interpreting the emission-line spectra, it is important to be able to distinguish emission produced by star-forming regions from AGNs. The conventional means for distinguishing between gas ionized by stars and nonthermal processes, are diagnostic line diagrams, also called BPT (Baldwin, Phillips and Terlevich, 1981) diagrams. They make use of reddening-corrected fluxes of the following lines: [OII], Hβ, [OIII](5007), [OI], [Hα], [NII], [SII]. The appropriate ratios of these lines can clearly separate extragalactic [HII] regions from AGNs.

AGN are characterized by the existence of a partially ionized zone. In this partly ionized region, ionized atoms as H and free electrons coexist with neutral atoms of other elements, as well as with ions having an ionization potential similar to that of H. The dominant forms of O, S, and N in the partly ionized zone are O^0, S^+ and N^0, while smaller fractions of N^+ and O^0 are also present. Hot free electrons produced in this region by X-ray photoionization will increase the strengths of lines produced by collisional excitation.

Important lines such [OI], [SII] , [NII] are of this type. Therefore, intensities of those lines are larger with respect to Hα in narrow-line AGNs than in H II region-like objects, because collisional excitation of these lines is more important in objects with extended partly ionized zones.

Since the ionization potential of O^0 matches the ionization potential of H very well, one should expect a large difference between the [OI]-Hα ratio of the H II region-like objects and that of narrow-line AGNs. The effect is also important for S[II]/Hα but the fact that S^+ can also exist within the H^+ zone of HII regions somewhat attenuates the difference between the two classes of objects.

Finally, O^{++} is produced predominantly by UV photons well inside the partly ionized zone, and close to the ionizing source. The relatively larger numbers of photons that can ionize O^+ to O^{++} in the power-law type spectra generally make OIII/Hβ larger in the AGNs than in all, but the highest HII region-like objects.

As recently pointed by Lamareille (2010), we can also use the "blue" emission lines (i.e. [OII], [OIII]λ5007 and Hβ) to perform the spectral classification for higher redshift galaxies (i.e. with no observable Hα and [NII]λ6584 "red" lines), but with a slightly lower accuracy.

For galaxies at intermediate redshift, the Mass-Excitation (MEx) (Juneau et al., 2011) diagnostic is able to identify AGNs in galaxies in the absence of near-infrared spectroscopy, necessary to use traditional nebular line diagrams at z>0.4. Combining [OIII]5007/Hβ and stellar mass this method successfully distinguishes between star formation and AGN emission.

6.4 NL AGNS PCA-based finding

Since, as explained before, the NL AGNs have a PCA reconstruction which traces well the majority of the spectrum, but for the Hβ and [OIII] emission lines, it is easier, in principle, to select these objects by comparing their projection to the observed spectrum. For the BL AGNs this would be harder, since the discrepancy between observed and projected spectrum spans over all the wavelengths, and a pick-up method for those objects would easlily collect many merely noisy spectra together with the BL AGNs.

I want thus to select NL AGNs, without the support of a training sample, but solely on the basis of the χ^2 of the difference of the observed and the projected spectrum. In particular I expect to have a small χ^2 (~ 1) regarding the continuum, and in particular the regions left and right of the Hβ-[OIII] doublet, and a high χ^2 (> 1) within the region of the Hβ-[OIII] doublet. I defined the χ^2 for the i-th spectrum as

$$\chi_i^2 = \frac{\Sigma_\lambda (P_{i,\lambda} - O_{i,\lambda})^2}{\Sigma_\lambda \sigma_{i,\lambda}^2}, \tag{6.1}$$

where P_i is the projected spectrum, O_i is the observed one, and σ is the relative statistical noise spectrum, as stored in the VIPERS database. I decided to pick up all the objects at different χ^2 thresholds (>1,2,3,4,5) in the region of Hβ-[OIII] doublet, but within a parent sample, selected to help the NL AGN finding.

Since I expect the projection of NL AGNs to fit well the observed spectra, outside the [OIII] region, the parent sample contains all the objects with χ^2 <2 in those wavelengths. Furthermore, I choose the objects with a good signal-to-noise ratio (S/N>5), where I

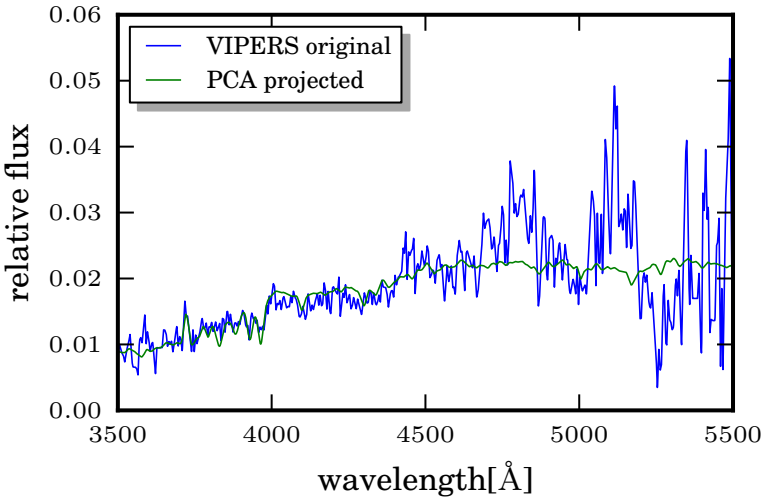


Figure 6.3: Example of an early-type spectrum, that can be mistaken for a NL AGN by the automatic NL finding routine, on the basis of the χ^2 on the Hβ and [OIII] emission lines.

defined the S/N as

$$(S/N)_i = \frac{\overline{O}_i}{\overline{\sigma}_i} \tag{6.2}$$

and decide to start avoiding the redder objects in the sample, selecting ϕ >-0.1; in facts, many spectra may be in general affected by strong noise spikes in the region of the χ^2 test we analyze for the NLAGNs (Fig. 6.3); for the redder spectra in particular, these noise spike would produce a very high χ^2 value in the test and would thus strongly contaminate the NL candidates' sample, appearing as NL candidates, even if there are no emission lines at all in the interested region. So I prefer to neglect early-type objects, even if a NLAGN my also lie in one of those, and such a candidate will be lost in this analysis.

Finally, I chose the flag interval 3≤flag≤10, to avoid too noisy or unsecure redshift objects (flag 0,1 and 2) and the already well known BL AGNs of the sample (flag 10-19).

6.4.1 Comparison to BPT diagrams

The results of the PCA based NL AGN identification have been compared to both the Lamareille and the Juneau criteria. The χ^2 threshold for the difference of the line intensities of [OII] and Hβ have been varied from 1. to 5. Figs.6.4 and 6.5 show the distribution of the PCA NL candidates (large purple dots) over the distribution of the VIPERS spectra for which a reliable measurement of the relevant emission lines is available (small green dots). The solid lines, in both figures, isolate the region of the NL AGNs (above the upper lines) from the one of the regular/starburst emission line galaxies. The purple dots in both figures refer to the χ^2=1 threshold.

A quantitative analysis of the completeness of the PCA NL sample vs. the Lamareille

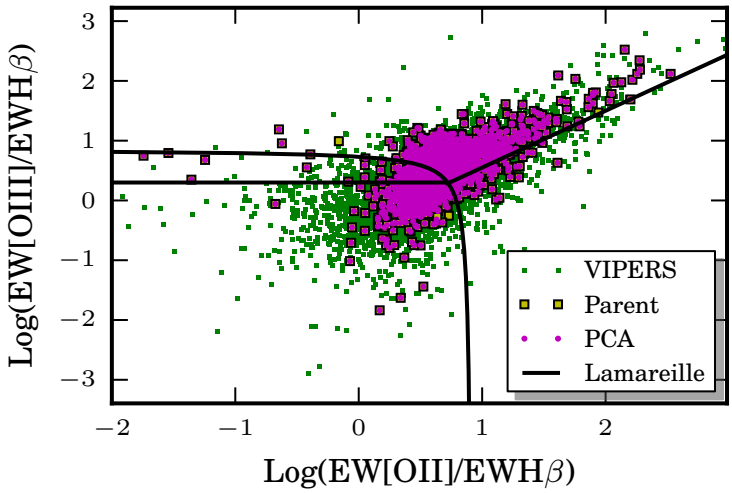


Figure 6.4: Comparison between Lamareille and PCA-based NL AGN separation (with threshold χ^2=1). The Lamareille criterion selects all points above the upper black lines.

and the selections have been performed. The results are shown in Figs. 6.6, 6.7: I notice that the completeness of our selection can be very high ($\sim$ 99%) for both the test samples, but we are left with an important contamination from regular galaxies, which is not acceptable for both comparisons, even if for the Lamareille criterion is at least beyond 50%. For the Juneau criterion the contamination of our selection method is almost 60% for $\sim$ 80% completeness, and the contamination is higher than 50% even for a completeness of $\sim$ 30%.

Even accounting for the fact that the Lamareille and Juneau test samples are not a highly secure sample of NLAGNs, the results suggest that this method is promising, but has to be refined: it is capable to find high purity samples of NLAGNs, but highly incomplete, or to find high completeness sample, but of course at the expenses of purity. A possible improvement, to be implemented in the near future, may be to use a more realistic noise indicator than the statistical noise spectra present in the VIPERS database, to compute the χ^2 of the difference in the regions of interest for this analysis. A more realistic indicator of this quantity has been computed, again on the basis of a PC analysis, as described in the Appendix A, and will be soon implemented in the NLAGN identification algorithm. A comparison of the candidates of this method to a highly secure set of VIPERS NLAGNs will be also performed in the next future.

6.5 Wrong redshift assignments

The machiner described to obtain a sample of candidate NL AGNs, can in principle be used to select, with ad hoc adjustments, also other peculiar objects, or to pick up rapidly the spectra whose redshift assignment is incorrect. Those spectra are expected to have a poor reconstruction, if compared with the observed shape, displaying a high global χ^2 value. Unfortunately, if the observed spectrum is very noisy, as it will almost always

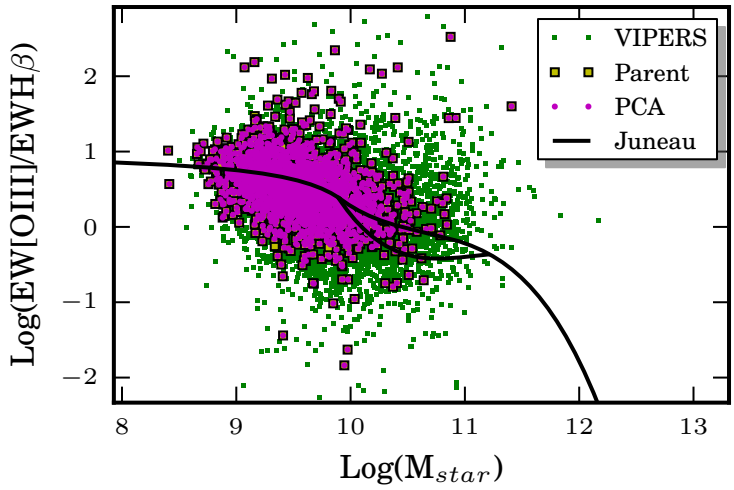


Figure 6.5: Comparison between Juneau and PCA-based NL AGN separation (with threshold χ^2=1). The Juneau criterion selects all points above the upper curved black line.

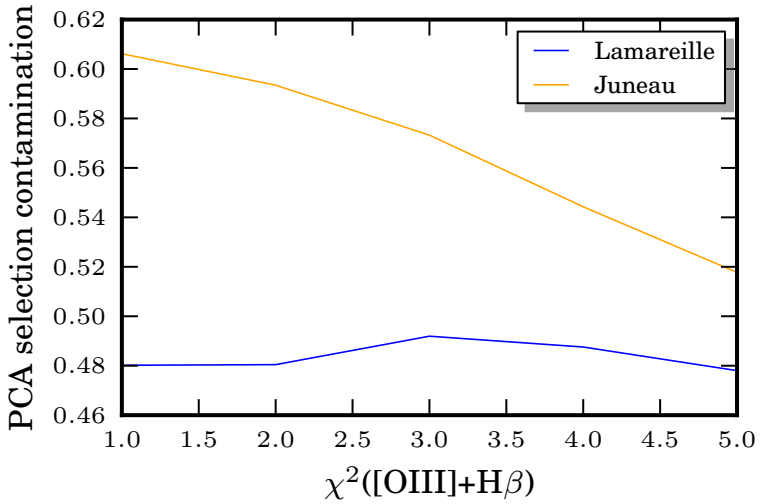


Figure 6.6: Contamination of PCA NL AGN selection with respect to Lamareille and Juneau ones, for different selecting χ^2 thresholds.

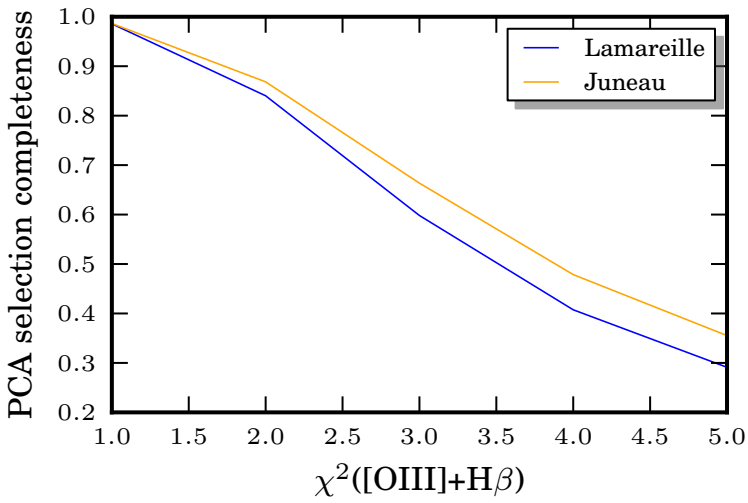


Figure 6.7: Completeness of PCA NL AGN selection with respect to Lamareille and Juneau ones, for different selecting χ^2 thresholds.

be the case for a wrong redshift assignment, the χ^2 will be comparable to the χ^2 of a noisy spectrum with right redshift assignment, frustrating the advantage of the bad reconstruction.

An alternative route is to try to link the wrong redshift spectra to a particular or particularly high value of an eigencoefficient. This have been tested by taking a good redshift spectrum and moving it to the wrong redshift. This causes the spectrum to show a highest spread of its 100 eigencoefficients, with respect to the right assigment, but not as high as to be detached "blindly" by the rest of the sample. I tested this over a dozen of spectra, obtaining the same result. So, I expect that a wrong redshift assignment will not show off an easily detectable peculiar kit of coefficients. Moreover, if there was a particularly high coefficient related to the wrong redshift spectrum, this would be different from one spectrum to the other, depending on the assigned redshift: this will also make it difficult to select those objects on the basis of their eigencoefficients.

Thus, the more promising route is still to select those objects on the basis of the χ^2 of their reconstruction, tuning carefully the wavelength range for the χ^2 test. This, together with improvements on the NL AGN selection, will be addressed in the next future.

CHAPTER 7

Linear Discriminant Analysis on PCA parameters

The data I will perform the Linear Discriminant Analysis on, are just the two coefficients ϕ and θ of the PCA decomposition obtained so far. The aim of this exercise is to separate active and passive galaxies, on the basis of spectroscopic data derived quantities, as ϕ and θ are.

The Linear Discriminant Analysis (or LDA) was originally developed by R. A. Fischer. LDA is a classification method that looks for a linear combination of variables (in a sample) that best separated two classes of objects. To capture the notion of separability, Fisher defined a function named "score", defined as:

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \tag{7.1}$$

where the βs are the coefficients of the linear model describing each sample

$$Z = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n, \tag{7.2}$$

C is the overall covariance matrix of the data

$$C = \frac{1}{n_1 + n_2}(n_1 C_1 + n_2 C_2) \tag{7.3}$$

and μ_1 and μ_2 are the mean vectors of the two classes, defined a priori in the sample on the basis of a fiducial/partial classification. Thus, the score function can also be written as

$$S(\beta) = \frac{\overline{Z}_1 - \overline{Z}_2}{\text{Variance of Z within groups}} \tag{7.4}$$

Given the score function, the goal is to estimate the linear coefficients that maximize it, and this can be obtained by solving the following equation:

$$\beta = C^{-1}(\mu_1 - \mu_2) \tag{7.5}$$

Then, each new point is projected onto the maximally separating direction, and classified as pertaining to the first class c_1 if

$$\beta^T \left(x - \left(\frac{\mu_1 + \mu_2}{2} \right) \right) > \log \frac{p(c_1)}{p(c_2)} \tag{7.6}$$

and viceversa for the second class c_2. p is the probability associated to each class.

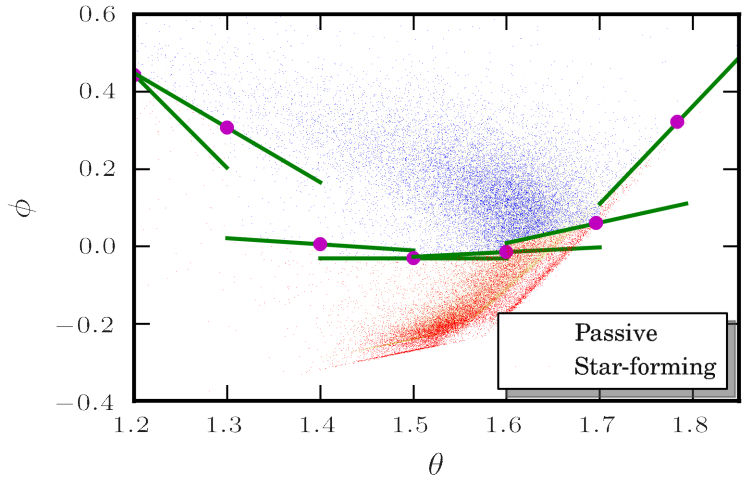


Figure 7.1: Starting step of LDA passive-active: the separation (green segments) is first performed into each single bin of θ, then the middle point of each segment (purple dots) is determined, and then they will be linearly interpolated

Since the LDA needs to lean on a pre-exhisting, even possibly partial subdivision into two classes, I decide to adopt two distinct training sets per time, each based on a different spectral feature.

First I use the intensity of the 4000Å break, and, separately, the equivalent width of [OII] line as a training set. The way these two quantities display on the ϕ-θ plot was already shown in Figs.5.9 and 5.10 respectively.

In particular, since a 4000Å break value ~ 1.5 is considered a reasonable boundary region between active and passive galaxies (Garilli et al., 2013) (see also Fig. 5.9), I choose two different training samples, to define an intermediate (though narrow) region. The first considers as starburst galaxies all the objects with 4000Å break$<$1.2, while the other uses 4000Å break$<$1.5 as a threshold. The objects that would fall within the two separators will thus be considered as transition objects.

I will apply the same idea to the [OII] equivalent width, although with a broader range of values. Since a good separator between active and passive galaxies can be represented by a value of 5 Å for the [OII] equivalent width (Mignoli et al., 2009), the two [OII] separators are fixed at 5 Å and 25Å (this value is large, to leave more space to the intermediate objects in the classification).

Let's keep in mind that not all the objects in the KL plot have a well defined or well measured 4000Å break and/or the [OII] equivalent width, and that the advantage of this LDA, applied to ϕ and θ, is just to be able to classify also those objects, for which a simple separation on the basis of the 4000Å break or [OII] would not be possible.

7.1 Active-passive galaxy separation

For both the boundaries of both the training sets I use, I first divide the KL plot into bins of 0.1 in θ, each overlapping by half to the previous and the following. Then I run

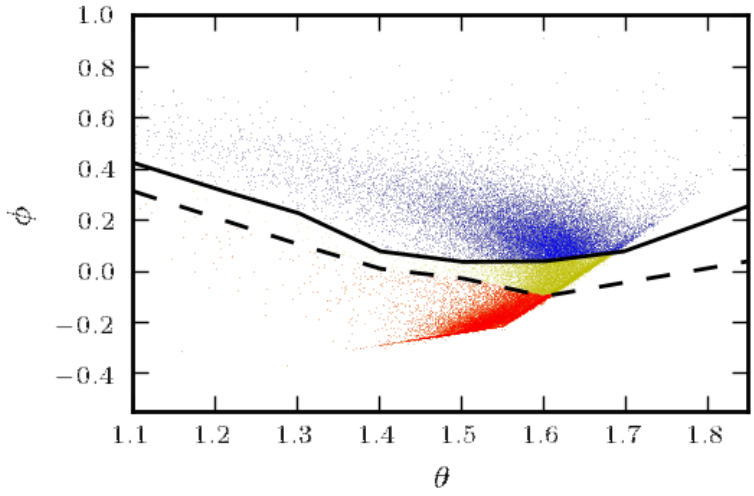


Figure 7.2: LDA active/passive separation: above the coninuous line, based on 4000Å break>1.2, the objects are active (blue dots), below the dashed line, based on 4000Å break>1.5, the objects are passive (red dots); the middle region contains transition objects (yellow dots).

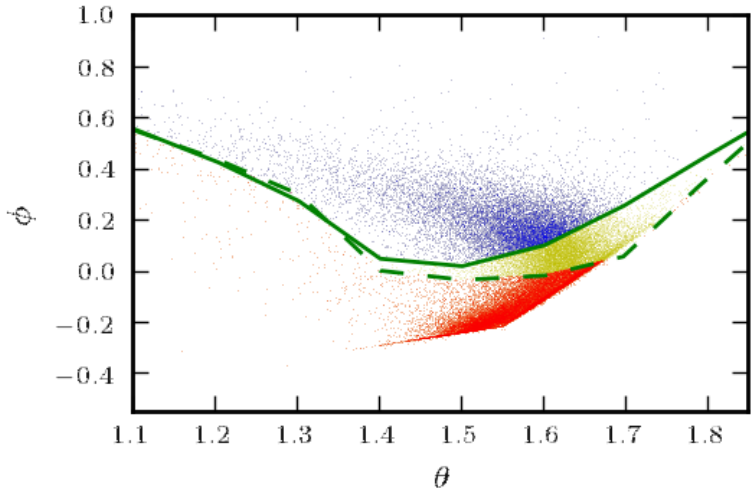


Figure 7.3: LDA active/passive separation: above the coninuous line, based on [OII]EW>25Å, the objects are active (blue dots), below the dashed line, based on [OII]EW<5Å, the objects are passive (red dots); the middle region contains transition objects (yellow dots).

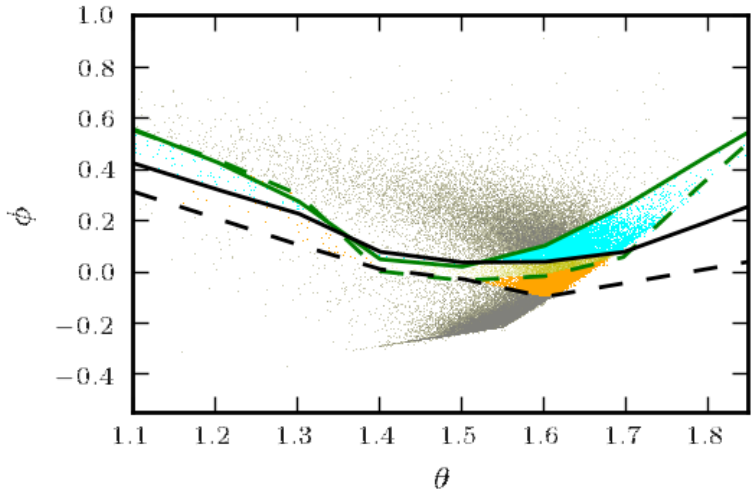


Figure 7.4: ϕ-θ LDA classification based on D4000 (black lines) matched with the one based on OII (green lines): the points in common to both the intermediate regions of the two methods are painted in yellow, the ones classified as intermediate by [OII] but as active by D4000 are cyan, the ones classified as intermediate for D4000 but as passive for [OII] are coloured orange.

the LDA over each bin for all the training samples, resulting into different segments, not necessarily crossing each other (see Fig. 7.1). At this point I calculate the average value for each of these segments, and join the middle points through a simple linear interpolation. The resulting division I obtain for the D4000 training set is shown in Fig. 7.2. The continuous line represents the LDA passive-active division based on 4000Å break>1.2, the dashed represents the one based on 4000Å break>1.5.

The classification based on the [OII] equivalent width is instead displayed by Fig. 7.3. Here again, the continuous line refers to a separation based on an equivalent width of 25Å, the dashed line to the one at 5Å. The fact that the 2 lines coincide in the left region of the plot, suggests that in that range of θ the ϕ-θ plot is not particularly sensitive to variations in the [OII] intensity. This is also evident form Fig. 5.10, where the points are all coloured bluish, and the effect may also be strengthened by the paucity of objects in the low ϕ-low θ region.

On the basis of this analysis, within each classifier, one could safely consider the objects below the dashed lines as passive galaxies, while the objects above the continuous lines are active. The middle yellow region represents the smooth passage between the active and the passive phase.

Since the [OII] equivalent width marks a separation between active and passive galaxies which is not completely overlapping to the one of D4000 (at least in the right region of the ϕ-θ plot), I finally overlap the two classifications, as shown in Fig. 7.4: this match defines a sort of "confusion" stripe within the ϕ-θ plot, approximately between the black dashed line and the green continuous line. Here the two LDA classifications agree in classifying the objects as transition ones in a smaller stripe (yellow dots), while the cyan points are considered active objects by D4000 and intermediate only by [OII];

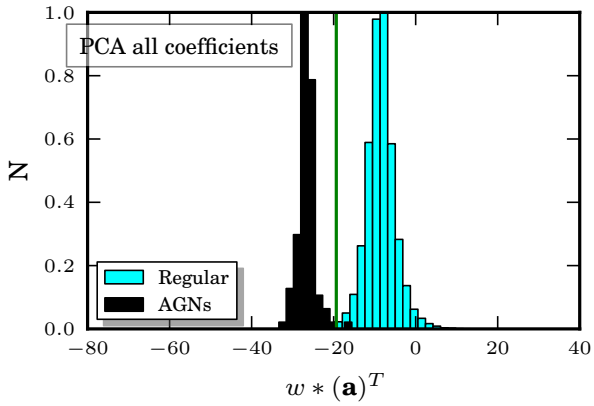


Figure 7.5: LDA normalized histograms for peculiar and regular VIPERS spectra, using all the PCA coefficients form PCA, and the AGN flags form VIPERS as a training set.

viceversa the orange points are intermediate for the [OII] but passive for the D4000. By this joint analysis I can conclude that all the points above the upper lines in the plot can be safely classified as passive, and all the points below the lower lines are passive. All the objects within these external lines can be considered intermediate or transition objects.

7.2 LDA AGN finding

The LDA can be used to separate distinct classes of objects in a sample; so in principle it can also distinguish between regular and peculiar spectra, even complementing a pre-exhisting identification, as may be te case for the AGN VIPERS catalogue. I will use the LDA to complement this catalogue with some more spectra, that for some reason may have escaped this formerly-performed AGN selection. I explained that PCA is not able to reconstruct the peculiar galaxies or AGN spectra, because the information contained into the first 3 eigenspectra encloses only galaxies from early type to regular-starburst, passing through late type. The AGN or peculiar spectra characteristics, because of the paucity of those spectra in our sample, are treated as noise and thus confined into higher order eigenspectra. Therefore, a LDA analysis of VIPERS spectra, using the first 3 coefficients of the PCA and the AGN VIPERS flags as a training set, may not be possibly helping in separating AGNs from regular galaxies.

On the other hand, I can still exploit the PCA results for this purpose. In fact, the first 3 coefficients do not contain information on peculiar spectra, but the whole set of PCA coefficients certainly would, since projecting every spectrum onto the whole set of eigenspectra shall result in a perfect recovery of the original spectrum with all its noise and peculiar features. So, for the purpose of identifying peculiar spectra, I take, as data, the entire set of 3 components' PCA repaired and not-yet-projected spectra, and project them on the full 2486 final eigenspectra basis. In this way I still have noise, since I projected over all the eigenspectra, but for the same reason I am contemporarily saving the majority of the peculiar spectra characteristics, whilst avoiding the presence of gaps (since I project the already repaired spectra on the already gaps-free eigenspectra).

For this analysis I used, as a training set, the sample of objects that have been flagged as AGNs in VIPERS: this sample contains mainly broad line AGNs (BL) and some narrow line AGN (NL).

The LDA analysis based on this training set can be represented by two well distinguished histograms (normalized to 1) (Fig. 7.5). The AGN spectra are expected to be at the left of the separation line, obtained through the equatoins of §2.2.2 while the regulars on the right. Nevertheless, when carefully examined, I find that an overlap of the regular (blue) histogram to the AGN (black) one does exist, and just this overlap should enclose the spectra that may have escaped the VIPERS AGN classification, at least on the left of the separation line l, precisely where we expect the AGNs to be.

Fig. 7.6[1] shows the stack of the spectra of this portion of regular galaxies' histogram, overlapping to AGN's one, as divided into 4 bins. Apart from the jaggies, their shape follow that of a BL AGN spectrum (Fig. 7.6). Nevertheless, when I look individually at the single spectra in the overlap region, I notice that the sample appears composed mainly by pretty noisy spectra, though with an AGN-like global resemblance, with the exception of some unequivocal (~ 20) broad-line AGN spectra (misflagged for some reason as regular galaxies by the pre-exhisting classification) and a couple of clean narrow line spectra, that are probably narrow line AGNs with particularly strong emission lines (these kind of objects, as already specified in the previous Chapter, share almost all the continuum characteristics with a regular galaxy, and are poorly represented in the training sample, so they are very unlikely to be segregated in this analysis, unless thay have extremely strong features). The predominance of the low signal-to-noise spectra to the AGN like ones in the overlap region, may explain the irregular shape of the lines in the stacked spectra. I specify that the effect of catching very noisy spectra in the AGN selection was indeed expected: in facts, the LDA separates the peculiar spectra not only on the basis of their similarity to the training sample objects, but also on the basis of their difference from the regular sample; thus some noisy spectra, for which very few characteristics of a regular spectrum are recognized, may have interpreted as peculiar objects.

Moreover, it's important to point out that the overlap of the histograms is mutual: the AGN histogram overlaps to the regular one also at the right of the separation line, meaning that the LDA separation of AGNs, operated through the PCA parameters of the repaired spectra, still needs some refinement, as explained below.

Concentrating on the AGN histogram (the black one), I know that, by construction, it should contain all and only the galaxies flagged as AGNs in the training set. Thus, first, it's interesting to see if there is a trend in the shape of the spectra, if one virtually moved along the abscissa of the histogram. So, dividing the AGN histogram into 3 large bins, and stacking all the spectra within each bin, I can notice that the average shape of the spectra shows continua and line intensities that seem to go from later to earlier type, moving from left to right (Fig. 7.7). As hinted above, I also see from Fig. 7.5 that one bin of the AGN's histogram lies at the right of the separation line, as if the LDA classified those objects as regular galaxies. This particular bin in facts contains the 2 AGN spectra at the bottom of Fig. 7.7: these 2 spectra present an important gap in their bluest region: this means that the PCA has attempted to repair them with the first 3 eigenspectra; the result of the repairing is not realistic though, since those 3 eigenspectra are not enough to describe an AGN spectrum, and the resulting weird global shape of the

[1]The spectra depicted in Figs.7.6, 7.7 and 7.8 are the observed ones, brought to rest-frame and binned to a common wavelength scale, not the repaired ones.

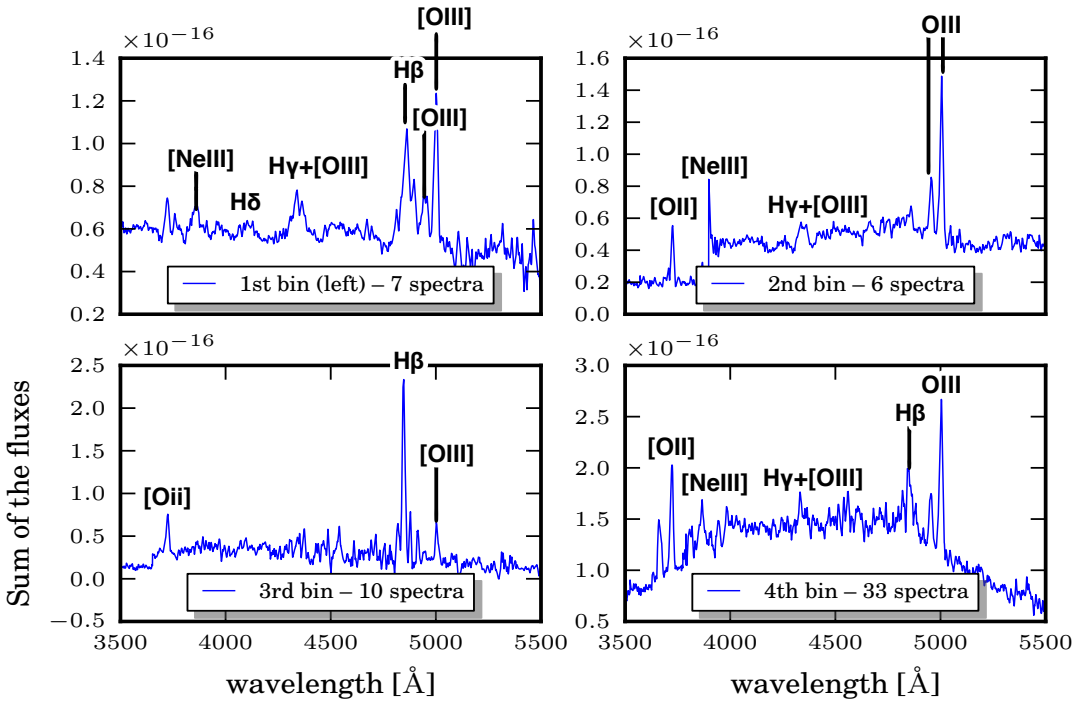


Figure 7.6: Stack of VIPERS spectra within 4 different bins of the LDA regular histogram, in the region that overlaps to the AGN histogram.

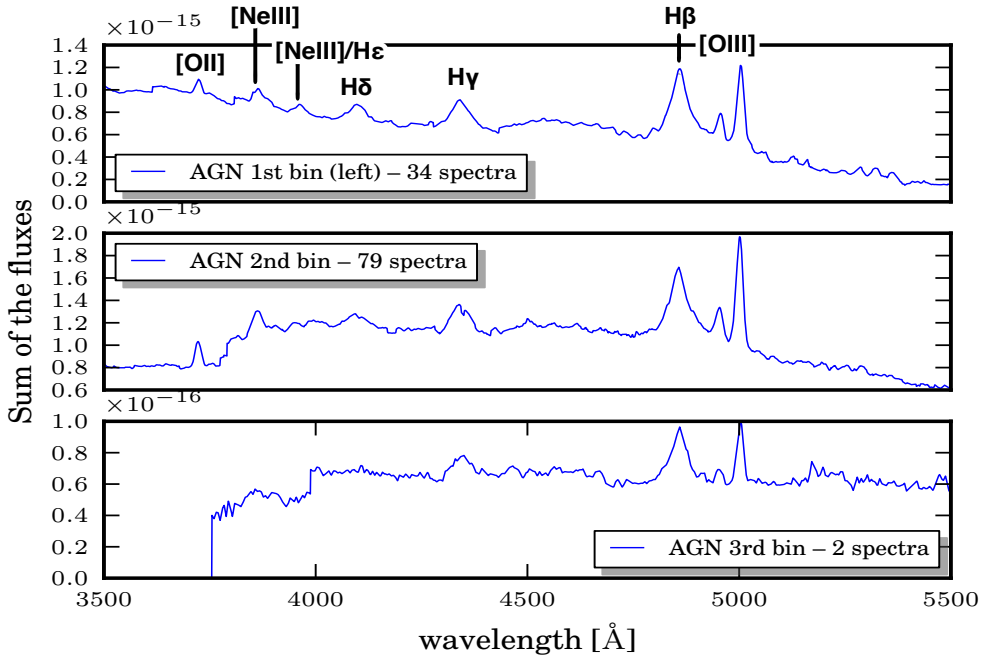


Figure 7.7: Stack of VIPERS spectra within 3 different bins of the LDA AGN histogram.

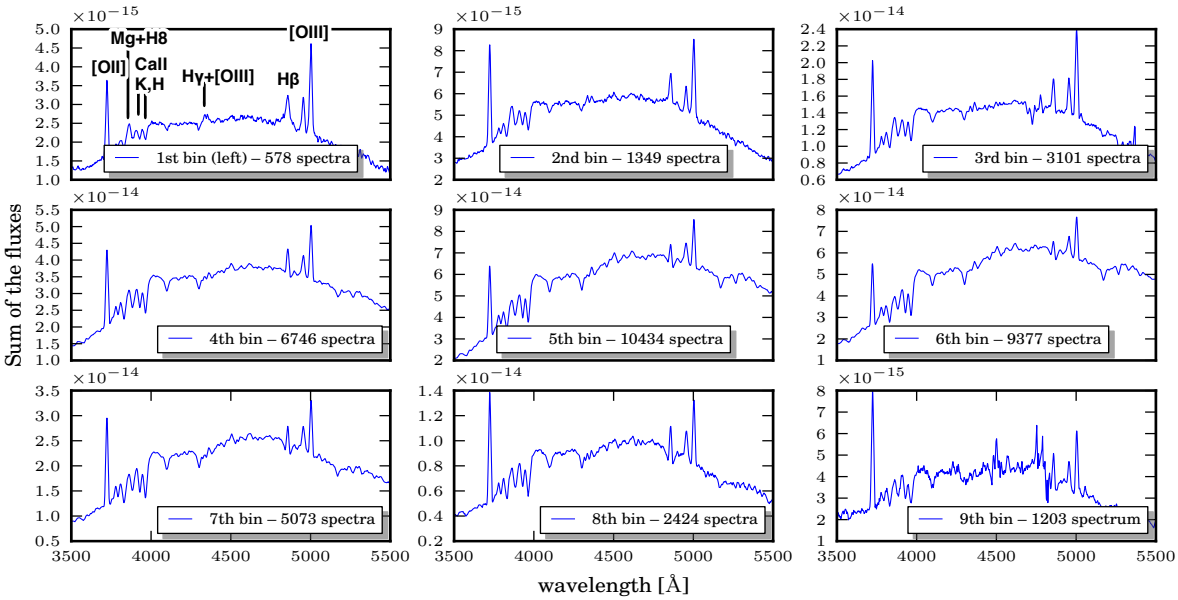


Figure 7.8: Stack of VIPERS spectra within 200 different bins in a 50000 bins regular galaxies LDA histogram.

repaired spectrum prevents the LDA to classify it as an AGN. The bin is thus inclosed by the regular "sector", in a region of the histogram that, as I will show, contains still some noisy spectra, whose shape is not securely distinguishable from a broad line AGN (see Fig. 7.7 and later discussion).

The regular galaxies histogram has in facts also been analyzed extensively; there is also a portion of it that overcomes the AGN histogram on the left side: once analyzed singly, these all appeared to be extremely noisy spectra.

For the other regular histogram spectra, that lay at the right of the separation line, I decided to analyze the objects contained in the more central and crowded region of the histogram, at 200 bins per time. The resulting stacks within each group show a weak increase in the intensity of the emission lines, but for the first two plots, which still resemble AGN spectra(Fig. 7.8), for the reason explained above: there may be a relation to the repairing of gaps using normal spectra, combined also to the presence of broad line AGNs in the regular histogram, that are nonetheless classified as regulars by the LDA probably on the basis of their repairings.

There are also other bins at the right of the denser part of the regular's histogram I considered so far: these contain again mainly quite noisy spectra, but with slightly clearer features than the ones at the extreme left of the AGN histogram.

If now I restrict our analysis to the high SNR spectra, i.e. only to spectra whose assigned flag is 3 or 4, except for all the spectra flagged from 10 to 19 (AGN flag), I can obtain an even better separation between peculiar and regular galaxies (Fig. 7.9): the separation line gets re-defined, and the AGN flagged spectra histogram now lies entirely on the left side of l, while there is still some superposition of the regular spectra histogram to the AGN one, containing the suspect AGN-like misflagged spectra. Restricting to the best flags, though, I lose some of the misflagged "regular" spectra that

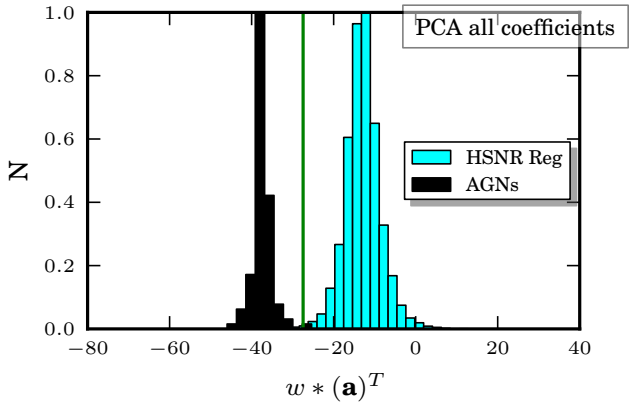


Figure 7.9: LDA separation between peculiar and regular galaxies, using the high signal to noise (flags 3 and 4) sample only.

were clearly AGN-like in the previous analysis, since in some cases they were originally flagged with a flag different from 3 and 4.

To better illustrate the performance of my binary classifier I slightly vary the threshold l, and plot the corresponding True Positive-False Positive rates as a function of the threshold, to produce the receiver operating characteristic (or ROC curve) of the classifier. I retain as True Positives all the bins of the AGN histogram falling at the left if the threshold l, and as False Positives all the regular histogram's bins falling at the left. The resulting ROC curve, presented in Fig.7.10, shows that the LDA classifier has a very high sensitivity, since the sensitivity ROC curve (red line) passes very near to the top-left region of the plot, meaning there is almost no false negatives and almost no false positives. The black dashed line instead depicts the so-called *line of no discrimination*, representing a completely random guess.

From this global analysis I can conclude that an LDA separation of AGN objects is certainly able to complement a pre-existing one, as proved for the VIPERS sample, by carefully checking the objects in the regular-spectra histogram that lay in the vicinities of the separation line, not only in the region overlapping to the AGN histogram, but also in that region pertaining to the regulars one, as proved by Fig. 7.6 and by the first two plots of Fig. 7.8. I can also state that for the regular's histogram, the presence of AGNs rarefies moving rightwards and the line intensities get stronger. Finally, in general, at the edges of the entire plot, the objects are more noisy; in particular ar the left edge of the AGN histogram the objects do not display clear features, while at the right edge of the regular's one, despite the high noise interference, the main spectral features, being stronger, are still recognizable.

I just remind here that the whole search of the AGN component of the sample, has been performed over the $0.4< z <1$. range: this is because that is the more convenient range for the PCA repairing and classification of VIPERS spectra, but it is also a region which is usually scarcely populated by AGNs themselves. So this kind of analysis could be surely improved in the future by extending the redshift range of the sample to be PCAnalyzed.

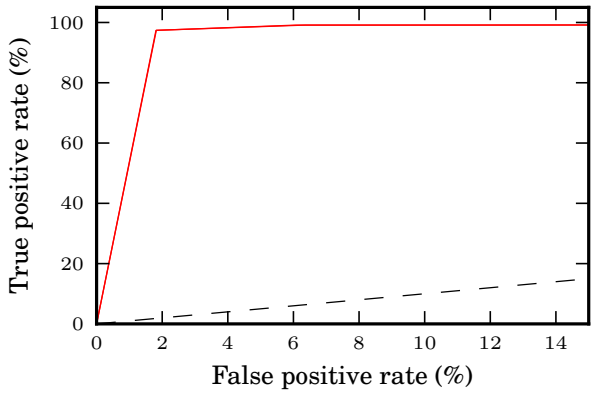


Figure 7.10: ROC curve for the LDA AGN-regular separator. The red line is the proper ROC sensitivity curve, while the black dashed line represents the response of a random guess.

I used this LDA peculiar separator to distinguish (mainly Broad Line) AGN objects from regular spectra. Indeed, if one could dispose of a secure (still not complete) sample of NL AGNs, to use as a trining set for the LDA, the LDA could be used to implement also a NL AGN separation, which is usually much more difficult to implement in a complete and secure way with other methods.

Aside to all this analysis, LDA could be also be developed to identify which eigenspectra are more important to describe the spectral continua, and which one are more relegated to the description of line features.

Conclusions and future directions

In this PhD thesis work I have developed an objective spectral classification system for the ongoing VIPERS survey using the Principal Component Analysis. Then I applied this to a subset consisting of 42036 galaxy spectra in the redshift range $0.4 < z < 1.0$. My implementation of a principal component analysis addresses the non-uniform characterstics of the dataset, that can impede the measurement and classification of spectral features; these include many effects: the variation of wavelength coverage in the rest frame has been taken into account by resampling all the spectra on a common wavelength grid, after bringing them to rest-frame. The noise properties and instrumental artefacts, together with the presence of gaps or manual interpolations, have been statistically corrected by applying the PCA to this rebinned rest-frame sample. The PCA I have developed here exploits an iterative algorithm that converges to a robust estimate of the eigenspectra templates (the first 3 eigenvectors of the PCA, or principal components): the novelty is represented by the impementation of a check that produces automatically realistic spectra, statistically repaired from the gaps and cleaned from noise.

From this statistical reconstruction of the sample, I have obtained a classification system, based upon three coefficients, a_1, a_2 and a_3, that are found by projecting the spectra on to the first three principal components. For the determination of the coefficients, for each spectrum, I have used a specific recipe to preserve the physicality of spectral lines such that both the continuum and line features are reconstructed accurately. The first three eigencoefficients thus provide a high-fidelity reconstruction of each spectrum, for a broad range of galaxy types.

The information enclosed in the three eigencoefficients can be compressed in the ϕ-θ angles representation: $\phi = \tan^{-1}(a_2/a_1)$ and $\theta = \cos^{-1} a_3$. This is a key step for my spectral classification: in a θ–ϕ plane, galaxies of different colour concentrate in different regions, according to the relative importance of the three eigenspectra. These same eigenspectra, at least in terms of the continuum, mirror the shape of realistic red, blue and intermediate galaxies.

To explore the physical meaning of the different positions on the θ–ϕ diagram, I have projected a set of Bruzual-Charlot model spectra on the same VIPERS eigenspectra and looked at their distribution on the same plot. I also added a set of 12 Kinney-Calzetti templates, as to verify the appearance of starburst galaxies over the same plane. An analysis with a group-finding algorithm, capable to divide space into maximally diverse classes, showed clear evidence of two different branches, following respectively the trend of the Bruzual-Charlot and Kinney-Calzetti models. The I have also dust extincted the models, to know in which direction the reddening for spectra moves the points in the ϕ-θ cloud.

Then to better explore the physical meaning of the different positions on the ϕ-θ diagram, I have also coloured it with a gradient based on the 4000 Å break and one based on the [OII] equivalent width, finding that the θ parameter, at least in the region of 4000 Å break > 1.3, is pretty insensitive to the continuum shape of the spectra, while, only at fixed values of ϕ, it is an indicator of the intensity of the emission lines, whereas in general it's also a very weak indicator of the continuum slope. ϕ instead is a strong indicator of the continuum shape, and it also contains some information on the line intensities.

A comparison of this classification method with two more common photometric selections showed that the PCA approach is comparable to a rest-frame color-color and also to a color-redshift plot in discriminating red from blue galaxies, whereas being more sensitive than photometry to intermediate spectral types, being based on spectra.

Finally, I have searched for evidence of evolution in the ϕ-θ parameters, and I have found that, for the red galaxies, mainly ϕ and slighty θ show weak trends in agreement that go in the direction of the expected evolution; for blue galaxies instead, due to a stronger competition of effects within each parameter (the lines and the continua), the evolution is smoothed and it's slightly oscillating near to $z \sim 0.7$. The trend of evolution of the global population, excluding intermediate type objects, is found in agreement to the model evolution of a population of galaxies formed at $z \sim 1 - 2$, but for the near redshifts, where the population appears slightly less "red" than the models. For the highest redshift avaiable for my data, instead, the evolution is well reflected by a model evolution of a young population, of which the VIPERS sample seems to be more largely composed at high redshifts.

I have subsequently applied a Linear Discriminant Analysis to the ϕ-θ coefficients, to define a quantitative boundary between early-type or passive galaxies, and late type or active ones, based on an interval of two 4000 Å break intensities of the spectra, joint with another LDA based on two [OII] equivalent widths. This LDA analysis brought to the definition of a separation region, which is composed of intermediate type objects.

Some peculiar spectra will not be well represented in the PCA eigenspectra, due to the rareness of their features in the sample. For instance, I have found that the eigenspectra do not fit AGN spectra well. However, in principle, interesting outlying spectra can be identified based upon poor χ^2 values for their PCA fit; on the basis of that I began studying the implementation of a new approach to separate NL AGNs, which looks promising but still needs to be refined and strengthened.

Then I applied the LDA to the VIPERS spectra, repaired with three eigencoefficients but projected over all of them (2486): I have done this to implement an AGN-regular spectra subdivision, based on the pre-existing VIPERS AGN flag system, which marks the AGN objects with peculiar flags. This LDA routine proved able to complement the original AGN identification with more peculiar objects.

Aside from all this analysis, I have also developed an automatic conservative masking of VIPERS noisy features, on the basis of an observed-frame PCA, able to assign a sky spectrum to each observed spectrum, and to mask and substitute the masked region with a realistic patch. The resulting masks are satisfactory, and the masking procedure, merged with the exhisting manual mask, it will be added to the reduction pipeline of VIPERS.

In future analysis also the LDA active-passive separation will be applied on the full VIPERS sample and refined, also testing a Quadratic Discriminant Analysis (QDA), which is able to handle and separate more than two classes per time.

I will also apply the LDA peculiar separator to a redshift range, which is expected to be more populated by AGNs, and also to find specific types of AGNs, such as the Narrow Line AGNs, provided I can use a clean and reliable sample of them, to use as a training set.

LDA will also be developed to define which eigencomponents are important to describe the continuum and which other to describe the line features.

A refinement of the PCA based NL AGN finder will instead be obtained by comparing the PCA candidates to a more secure set of NL than the Lamareille one, and by improving my same definition of χ^2 itself.

The χ^2 test separation, will also be implemented to the aim of finding automatically spectra that have been assigned a wrong redshift.

On the other hand, the same whole principal component analysis can be extended to include additional photometric parameters, whose information content is not already encompassed by the spectra, like photometric measurements in wavelength bands not included in the PCA spectral range, dust content, morphology or mass, provided I find an efficient way to weight them properly with respect to the spectral information itself. This will be useful to produce a "classification" cube or hypercube, and see which parameters are correlated, at least for the majority of galaxies, and to eventually find outliers into this global classification.

Such a spectral based PCA classification can also be applied to the studi of Large Scale Structures, to cleanly separate different populations of galaxies, and to implement a multi-population redshift distortions analysis, which should produce a lower error in the determination of the growth function of structures. Also, in the field of gaaly evolution, it can be developed to select a peculiar population of galaxies, such a the Luminous Red Galaxies in VIPERS: this way one could have a check of the expected passive evolution of those objects, by looking for the progenitors of SDSS LRGs into the VIPERS LRG sample.

Appendix A

Automatic PCA-based spectral cleaning

The PCA on the spectra, presented so far, is aimed at giving a statistical repairing and cleaning of observed spectra in a survey; for this reason, as explained, it is necessary that all the spectra have been brought to rest frame and on a common wavelength grid: this way all the features in the different spectra can be matched, producing spectral-shaped eigenspectra, which can be used to give realistic reconstructions.

Of course in every survey the objects are observed at different redshifts, which would cause the galactic features to mix up in a classical PCA reconstruction, if not brought to rest frame. On the other hand, if the galaxy features would get lost in an observed frame sample, the sky features of a given survey, clearly all pertaining to the Earth atmosphere, would sum up in an observed-frame PCA, whilst the spectral ones would be confined to high order eigenspectra, and interpreted as "noise".

Since, for a survey, one is interested in retaining the majority of measurements, trying to avoid the sky artefacts, in this technical appendix I will present a PCA based technique I developed, to estimate the shape and intensity of the sky signal within each spectrum in a survey, and to properly mask it.

This is a totally different approach from the PCA presented in the previous Chapters. In facts, the previous rest-frame PCA has been used to repair the spectra whenever they were "manually" edited, or where they presented zero-flux regions due to the rest-frame moving process. Then they were projected over the rest-frame eigenspectra to provide a statistical cleaning of the entire spectrum from all the noise features.

Here instead, I am taking the unedited spectra at their observed frame, and using the PCA machinery to edit them in a more conservative way on the basis of the sky signal, obtained from the spectra themselves; a similar approach has already been tested and applied to the SDSS data by Wild & Hewett (2005), providing a dramatic improvement in the quality of spectra, originally affected by strong OH sky emission lines.

Furthermore, I also want to replace the PCA-masked regions I will obtain with this analysis, with something better than a strainght line, whenever is possible.
Again, I will apply this routine to the VIPERS complete spectral sample, excepted for stars and AGN: the basis technique developed here will be suitable to any spectral survey, but the refinements I will apply to it, mainly in relation to the zero-orders masking, are strictly related to the VIPERS sample.

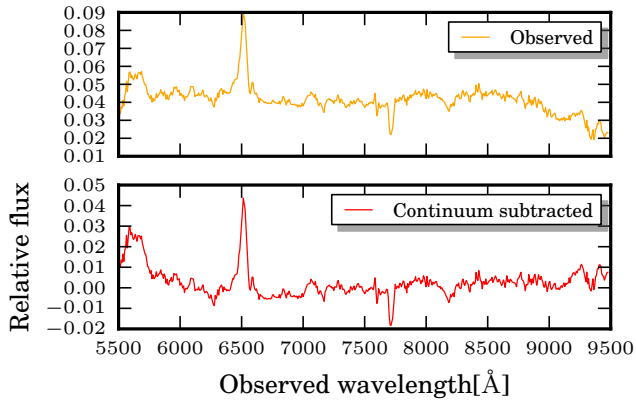


Figure 7.11: An example of continuum subtraction from an observed spectrum.

The sky eigenspectra

The first step of this analysis is to obtain the sky eigenspectra, i.e. the principal eigenspectra of the observed frame sample. To do this I take all the galaxy spectra available in the sample, with the only exclusion of Flag 0 objects (no redshift assignment) and of Flag 10-19 objects (AGNs); the mask I will obtain will be applicable to every other spectrum in the survey.

In the observed-frame PCA, the spectra continua and features will not add coherently; however, I will still find a smooth signal representing the superposition of all spectra, with offsets drawn from the redshift distribution. This signal is statistically due to the sky. Thus, to isolate it, I subtract the continuum of each spectrum before computing the eigenspectra (Fig. 7.11). To estimate the continuum I apply a Gaussian smoothing to the spectra, with standard deviation 50 pixels (i.e. $50 \times 7.1 = 355$Å). Although more precise continuum subtraction schemes could be implemented, this step does not seem to limit the usefulness of the eigenspectra later on.

Then I construct the data matrix and compute the corresponding correlation matrix and eigenspectra in the usual way (but in observed frame), without least-square routines or penalty terms, since I don't need to repair anything at this stage fo the work.
Since I know that the VIPERS spectra presented some problems of fringing for the reddest regions of the spectra, and that this problem has been corrected after a given date, I decided to divide the sample into pre- and post-refurbishment data, and to have a look at the different eigenspectra.
The number of eigenspectra to be retained, to describe the sky without including spectral features, has been found to be of 3, as in the case of rest-frame PCA for spectra.

In facts, from Figs 7.12, 7.13 and 7.14, and from different tests, projecting the spectra on the 3 terns of eigenspectra, it is clear that the post-refurbishment-only eigenspectra are the ones which contain the majority of sky features. For the other two terns, to be able to catch a comparable number of sky features in the spectra, I would need more than 3 eigenspectra, risking to include some spectral features.
Once determined the sky eigenspectra, I project every spectrum onto them, to get the statistical sky spectrum associated to each galaxy spectrum.

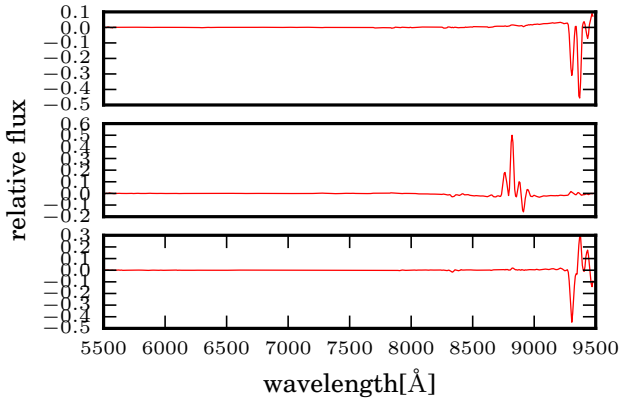


Figure 7.12: PCA first 3 eigenspectra for the observed frame pre-refurbishment data.

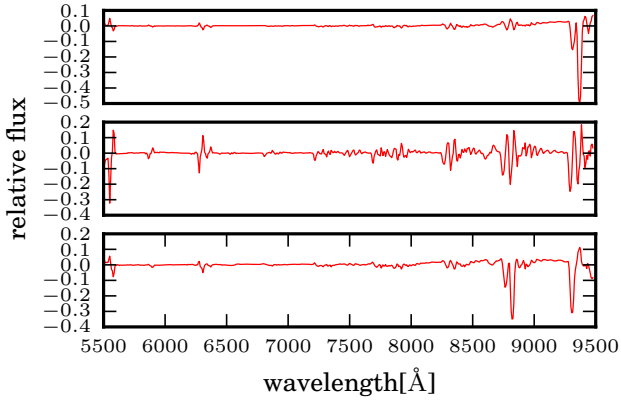


Figure 7.13: PCA first 3 eigenspectra for the observed frame post-refurbishment data.

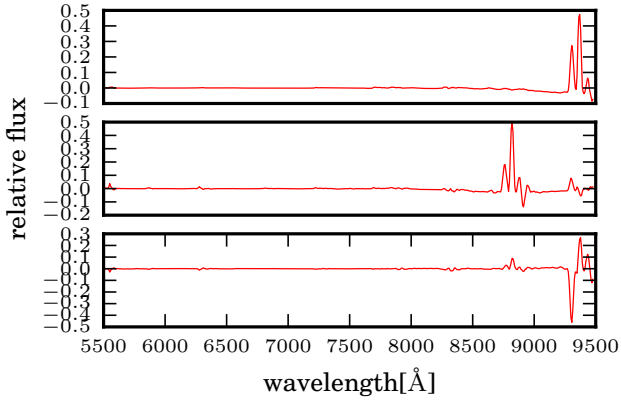


Figure 7.14: PCA first 3 eigenspectra for the entire set of observed frame data.

Automatic masking the spectra

The aim of the observed-frame PCA is to define a mask for each spectrum, in correspondence to the more intense and common noise, fringing and sky residual features.

Identifying the common sky features

Each spectrum is projected onto the most significant three eigenspectra. Since these eigenspectra are templates for common artefacts in the spectra, I obtain an estimate of the residuals of contamination in each spectra. I take the liberty to call these projections simply *sky spectra* (even if they represent sky residuals after a previous sky subtraction).

On the basis of the sky spectra I have to decide which features should be masked. The most natural approach is to define the mask in correspondence to the spikes of each sky spectrum. I first define the standard deviation as a function of each continuum-subtracted spectrum. This is used to define the threshold. The contamination is worse at $\lambda > 7500$Å, I thus adopt a different threshold in the blue and red wavelength ranges. Pixels in the spectrum are flagged when the difference is larger than the threshold.

This machinery is able to identify almost every important residual sky feature in the spectra, with some exceptions. In fact, there are some features, in particular the one around 6300Å, which, due to residual fringing, sometimes are placed at slightly shifted positions from one spectrum to another. For these objects the projection on the three sky eigenspectra, which displays those features at a given wavelength, returns a sky spectrum with a negligible spike in that region, preventing the automatic masking to detect it. For this reason I introduced a complementary masking procedure, based on the spikes of the corresponding VIPERS noise spectrum. Again, I define the mask in correspondence to the spikes and the dips of each noise spectrum. Also in this case, since the constamination is worse at $\lambda > 7500$Å, I use a different threshold in the two parts of the noise spectra.

Finding the zero-order spikes

The observed-frame PCA mask approach is able to identify the strong residual sky features in every spectrum. But of course, the sky is not the only source of noise which can corrupt the signal. For a fraction of VIPERS spectra there is also contamination from a zero-order image of a bright object (typically a star), corresponding to the spectrum lying on the VIMOS frame directly above the one we are considering. This appears in the form of a large spike at random wavelength positions, and cannot clearly be captured by the sky eigenspectra. Fortunately, the solution to this problems comes by itself, since the device introduced to mask the sky fringing residuals is also automatically able to catch the 0-orders, which are once more clearly reflected in the noise spectra.

Real features safeguard

I need to make sure that the mask positions do not coincide with or were very close to the expected position of known emission lines (3728, 4861, 4959, 5007, 6562Åfor regula galaxies, more for AGNs). For this reason a safety window, to the sides of the center of any expected emission line, has been assigned a 4 pixels length in total (28Å). In the case where a 0-order candidate fell into the window, the mask was not applied and that portion of spectrum was left in its original shape. This has been done to preserve as much as possible the measurements of the spectral emission features.

Mask repairing

Beyond defining a mask for all the spectra within a survey, I also want to provide with the most likely shape I would expect the spectra to have, in the regions of the mask, if there were no noise. A horizontal straight line, at the level of the rest of the spectra, is usually a good indicator of the intensity of the continuum. But I am interested in giving a more realistic aspect to the spectra, after the mask has been applied. For this reason I approach two different methods: a subtraction of the sky specrum to each spectrum, in the region(s) of the mask, or the replacement with a rest-frame PCA reconstruction.

For both the cases, since I used continuum subtracted spectra, I have first to add the continua back.

Sky residuals subtraction

The first attempt to give a representative, for the portion of masked spectrum, was to apply a noise subtraction to the spectra. All the noise spectra have been subtracted to the corresponding galaxy spectrum, only in the regions of the mask. This gives a good patch in the majority of the cases, but sometimes the "cleaning" is not realistic, or it goes in the direction of further degrading the spectrum. In fact, the noise spectra represent the more probable noise that should have contaminated the spectrum, on the basis of the 3 sky eigenspectra, and of the projection of the single spectrum on them; this procedure surely gives a reliable indication of the regions where there are important sky features, but the precise entity of those features is not always well recovered. For this reason it is preferable to adopt another approach to fill in the mask regions.

PCA reconstruction

Since, from the work done on the rest frame spectra, I dispose of the PCA reconstructed spectra, another possibility to fill in the mask is using the PCA reconstruction. Of course,

since the rest-frame PCA has been done on a restricted sample of objects in redshift, the reconstruction is not available for every spectrum. Furthermore, the wavelength range for each rest-frame spectrum is quite limited (3500-5500 Å), to lower the appearence of rest-frame-moving gaps in the spectra. So, also for any observed-frame spectrum which has its counterpart in the rest-frame PCA, not all the wavelengths are available.

For this reason, I decided to run a new rest-frame PCA, aimed at the mask substitution. This PCA has been run over all the spectra for which the mask has been created, and for a wavelength range able to contain all the spectral features, for spectra from z=0.4 to z=1.4. The more extreme redshifts have been excluded, because including them would have meant to enlarge significantly the wavelength range for a very reduced number of spectra, which, bacause of their paucity, would have not been well recovered in any case. Of course, running this PCA on the same rest-frame grid, chosen for the previous Chapters one, would have been very challenging in terms of computational resources. In fact, to have an optimal sampling of the spectra on the rest frame grid, I had defined a very dense logarithmic binning (see Chapter3).

For the current problem instead, once obtained the reconstructions, I will interpolate back to the observed frame grid. Thus, in this case, the high accuracy of the rest-frame sampling is not really crucial: for this reason I adopted a linear pixel separation which is more rarefied than in all the previous work, and in particular it's approximately the same as the observed spectra (7Å). This way I can perform a rest-frame PCA over all the spectra without computational impediment, and use the resulting reconstructions to fill in the mask regions. A little caveat to keep in mind is that, in the edge regions of some spectra, the reconstruction I obtain is not always a perfect indicator of the expected continuum; in fact, to be able to reconstruct all the spectra, I had to choose a large rest-frame redshift range, and there are few spectra at those redshifts that can provide information at that wavelengths.

So, for the objects at redshift lower than 0.4 or higher than 1.4, I decided to fill in the mask with a horizontal line, at the same level of the average value of the spectrum. This average has been computed within a short wavelength range near to the interested masked region. Indeed, in some particular circumstances, e.g. when the spectrum is very steep, this technique produces fake step bumps in the cleaned spectrum; thus, for those objects, I simply connected the 2 points at the left and right edge of the mask (this last approach may instead produce fake shark fin spikes in the spectra with a flatter continuum). For Flag 0 and AGN objects, then, the reconstruction within the mask must be taken with a grain of salt, since the PCA reconstruction for AGNs is twisted and for the Flag 0 is impossible, missing their redshifts (for the latter, we used a simple straight line).

The plots shown in the pages of Appendix B give the comparisons between the PCA-only built mask and reconstruction (top and second row), and a reliable manual editing of the spectra (third row). The difference between automatic and manual mask is plotted in the mfourth row, while in the bottom row the merged mask and relative cleaning is depicted.

In Appendix B I show examples spanning the range in quality flags and redshift. Eight examples were selected randomly for flags 0-19. All examples are from the VIPERS pointing W4P017 which is pre-refurbishment, but the mask is completely equivalent for post-refurbishment spectra.

As expected, in the majority of cases the PCA-mask agrees with the manual editing. In some cases, the PCA masks in regions where the eye would not, because there are some spikes in the noise spectrum, not evident in the galaxy spectrum. In some cases

some obvious spikes are not masked, because of their vicinity to emission features, or because they are not reflected in the noise eigenspectra, or they do not overcome the second-order threshold. But in general the authomatic masking is quite reliable. The mask substitution instead is more realistic than in the case of manual editing, but for some sporadic case, near to the edges of the spectrum.

Automatic cleaning vs. manual editing

The automatic masking and cleaning procedure presented here is generally helpful and conservative but not unfailing. Spurious and rare features that can be easily captured by the human eye may escape the PCA machinery. On the other hand, the manual cleaning can miss some residuals or, conversely, can mask too much, causing the loss of a large amount of spectral information, or even replace the masked region with some unphysical patching. For these reasons, and disposing of many careful manual editings for the spectra, I decided to merge the automatic mask with the manual one for every spectrum, and to produce spectra that are cleaned within the resulting merging of the two masks. In particular, to prevent the final mask to include too many consecutive pixels, I imposed that every manual mask composed by more than 70 consecutive masked pixels, was not merged to, but simply replaced with the automatic mask in that region.

The resulting cleaned spectra will be released along with the observed ones, and the automatic cleaning machinery will be added to the VIPERS reduction pipeline.

Appendix B

In this Appendix I present some PCA-mask examples from epoch 1 (pre-refurbishment) spectra: there are 8 examples for each of the flags.

In each plot, the top plot shows the original spectrum (solid black line) in the observed frame. The fit with the sky eigenspectrum is shown by the cyan line. The placement of the mask is shown by the step function between 0 and 1×10^{-17}. The second plot shows the spectrum after automatic cleaning. Vertical dashed yellow lines give the location of four common emission lines (3728, 4861, 4959, 5007, 6562Å). The third plot shows the result of manual cleaning by expert members of the VIPERS team. The fourth displays the difference between the manual and the PCA mask. The bottom plot shows the result of the merging of human and PCA mask with relative fillings.

Figure 7.15: Eight example spectra with flag 9

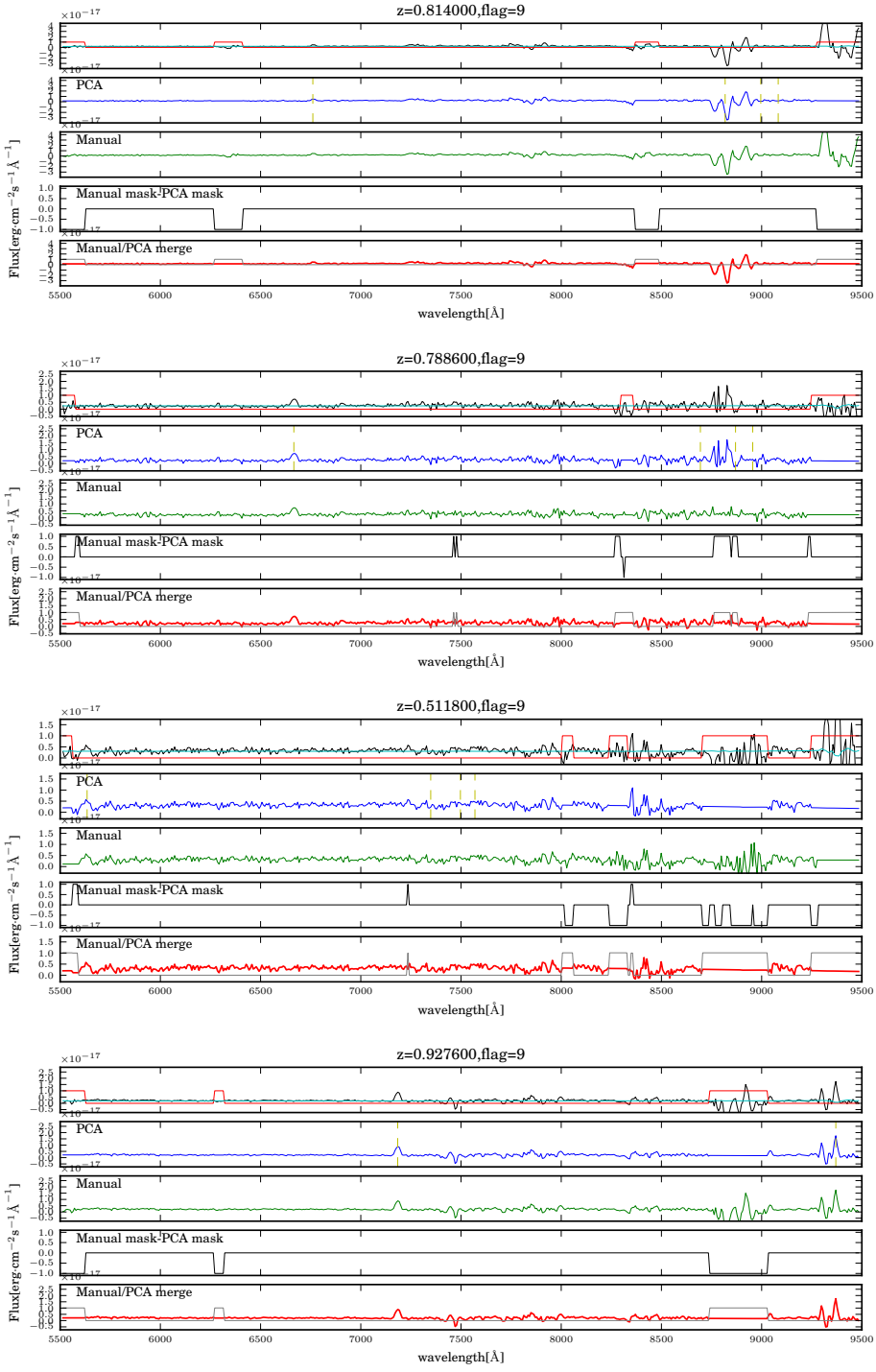
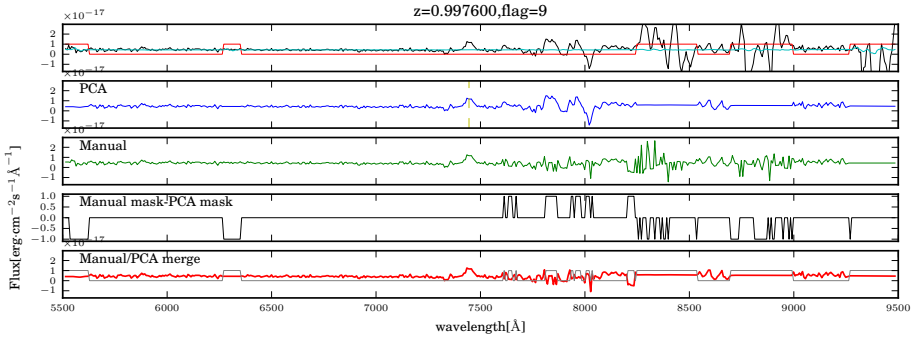
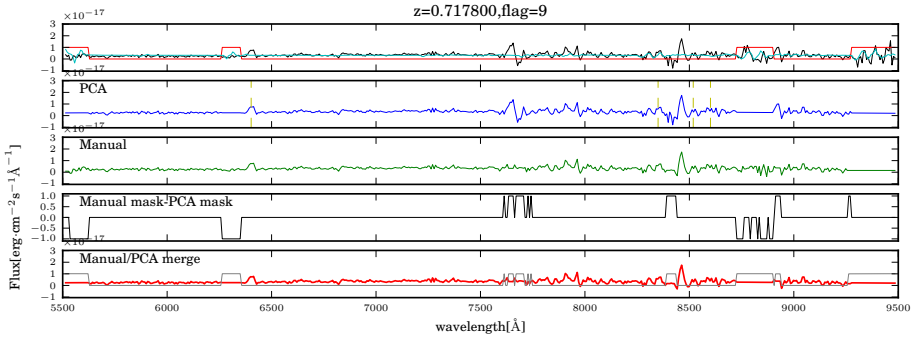
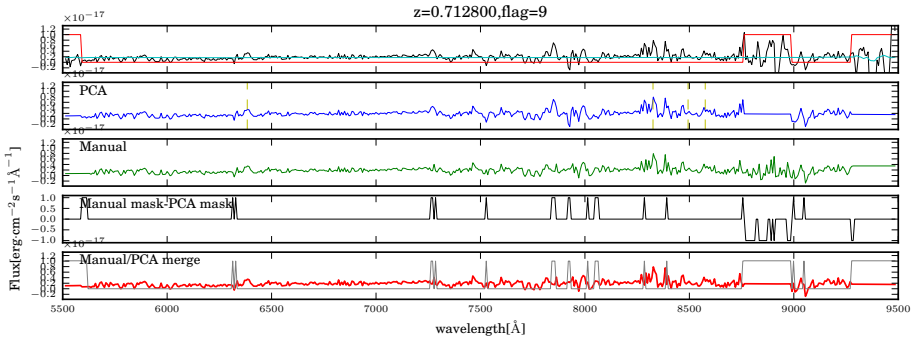


Figure 7.16: Continued... Eight example spectra with flag 9







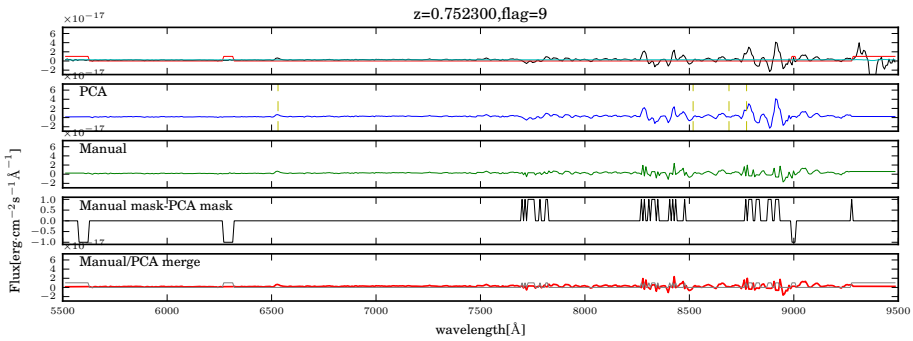


Figure 7.17: Eight example spectra with flag 4

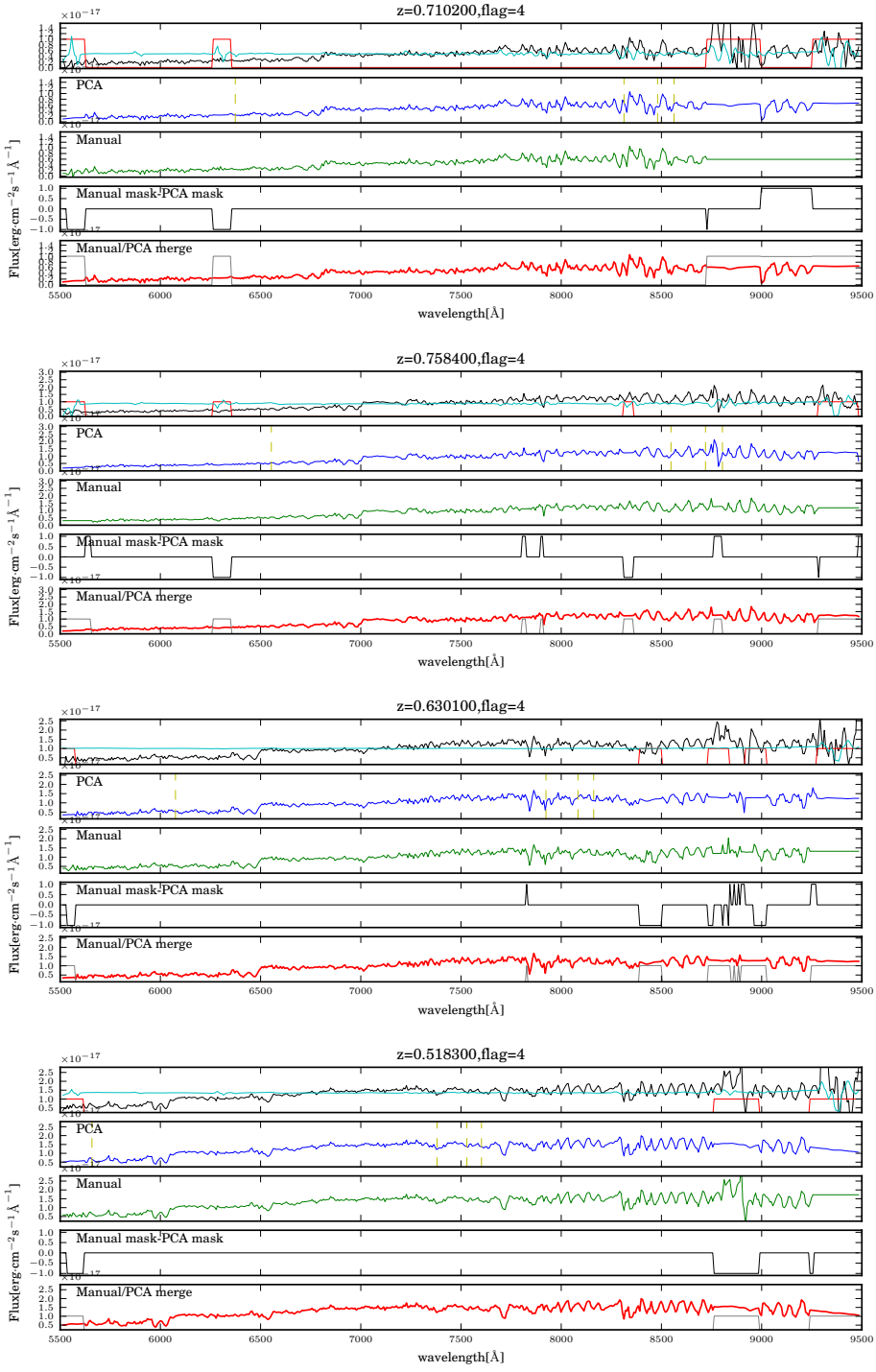
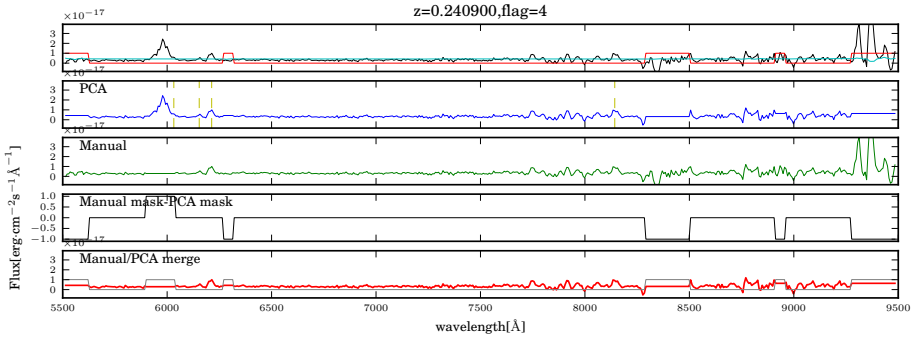
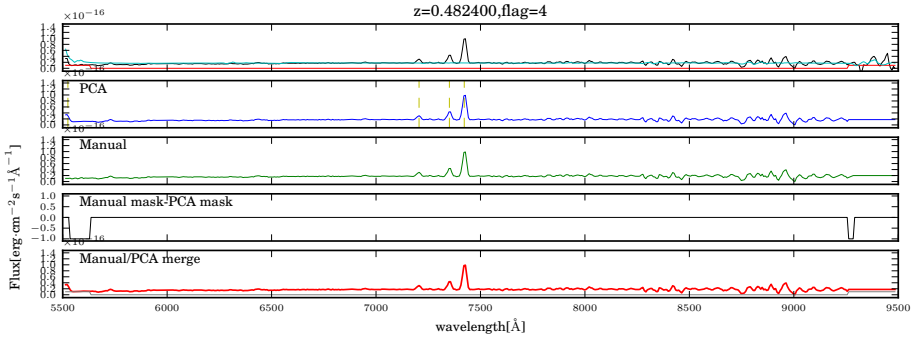
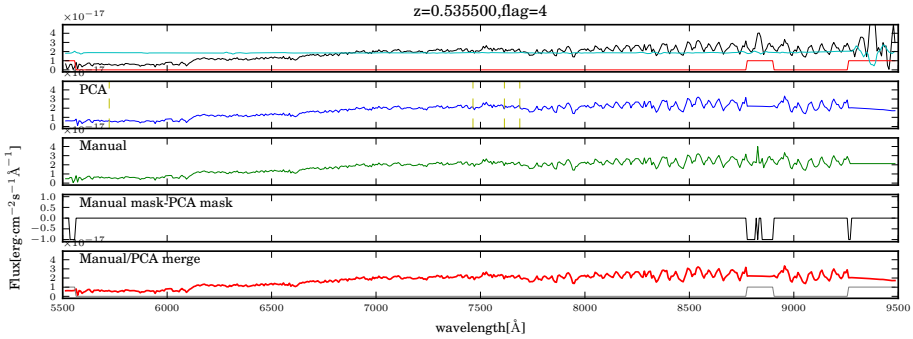


Figure 7.18: Continued... Eight example spectra with flag 4







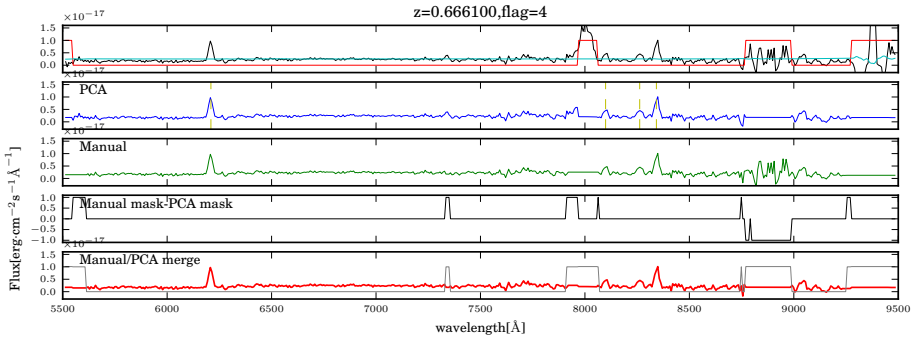
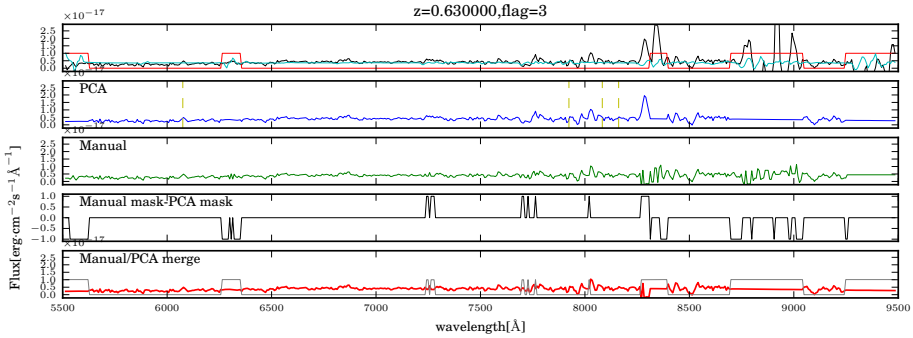
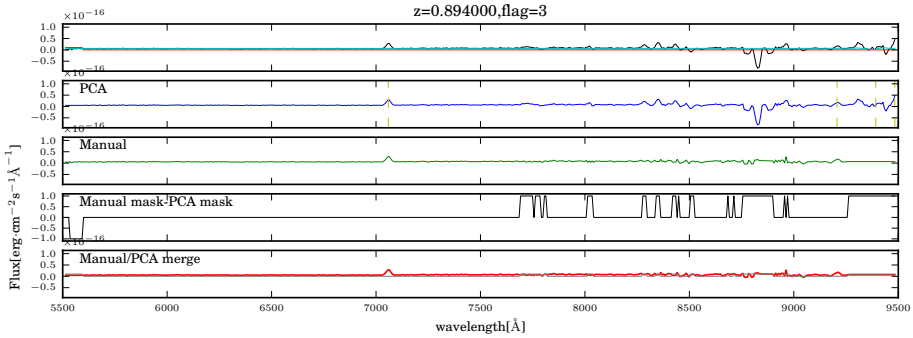
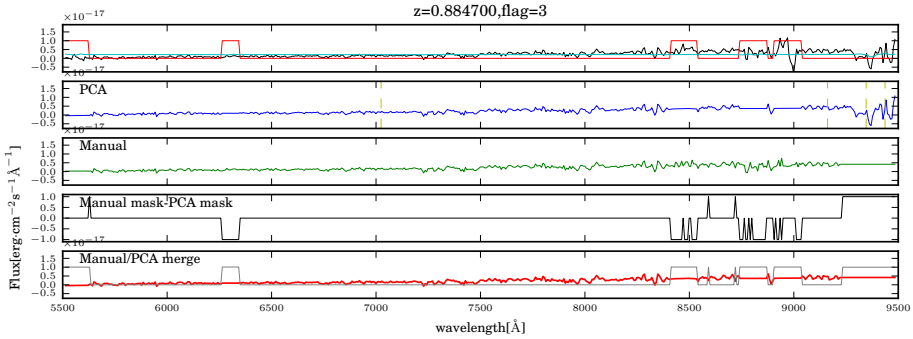


Figure 7.19: Eight example spectra with flag 3







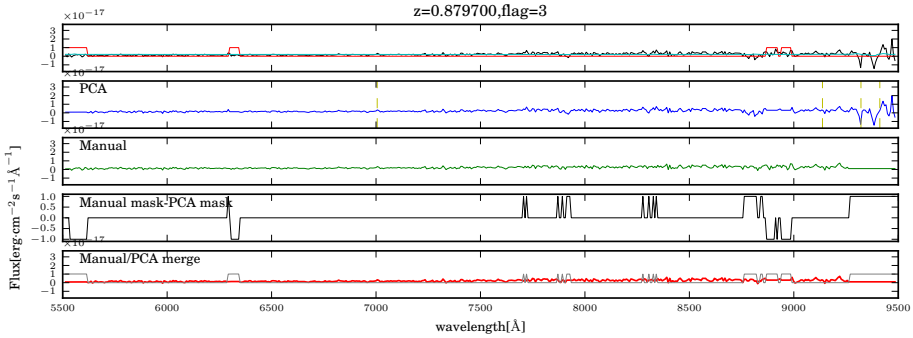
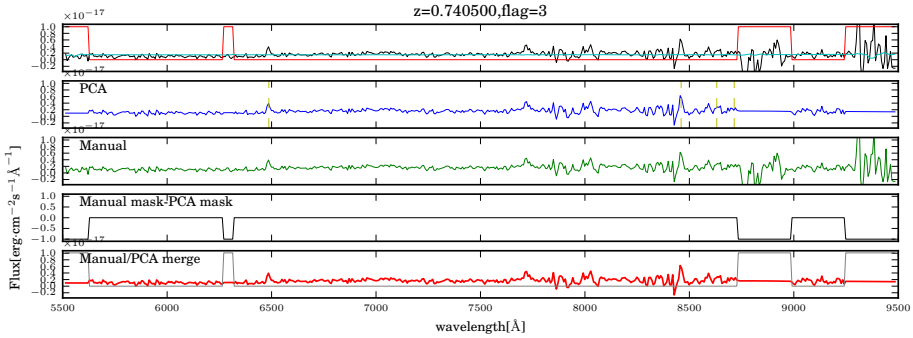
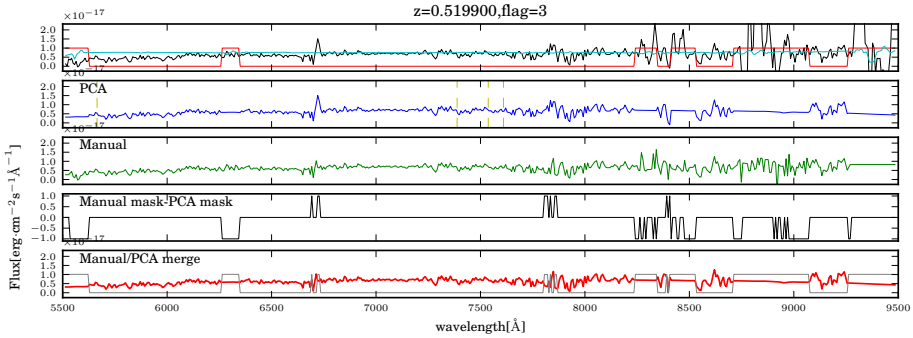
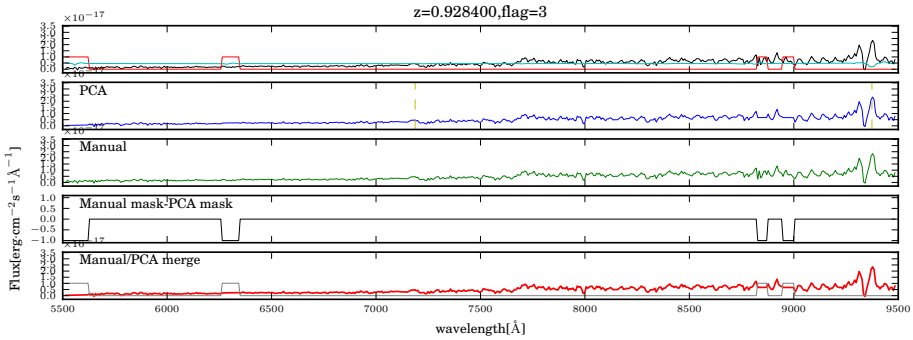


Figure 7.20: Continued... Eight example spectra with flag 3







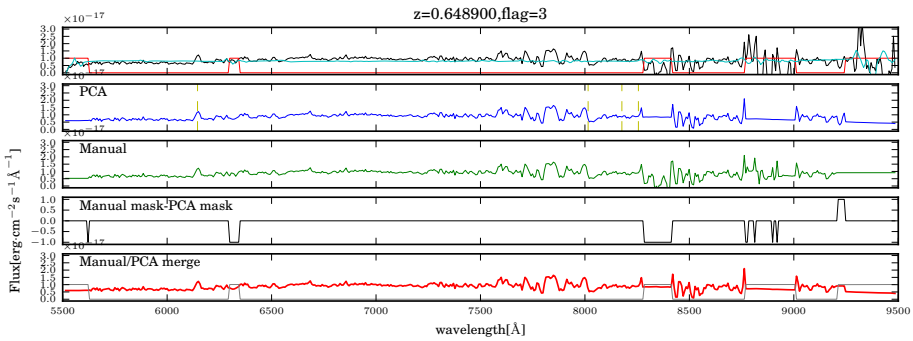
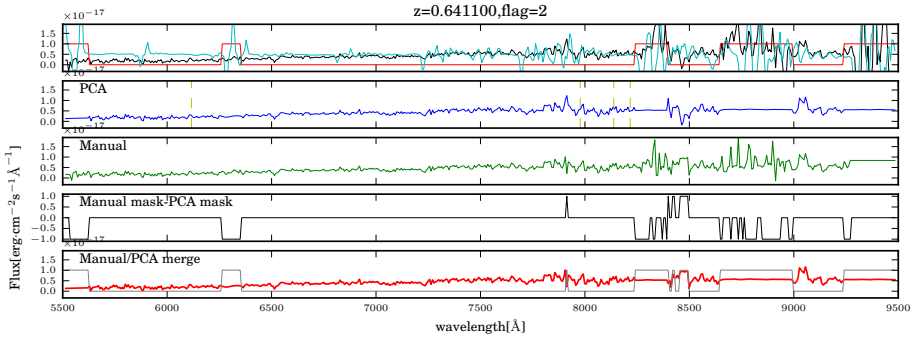
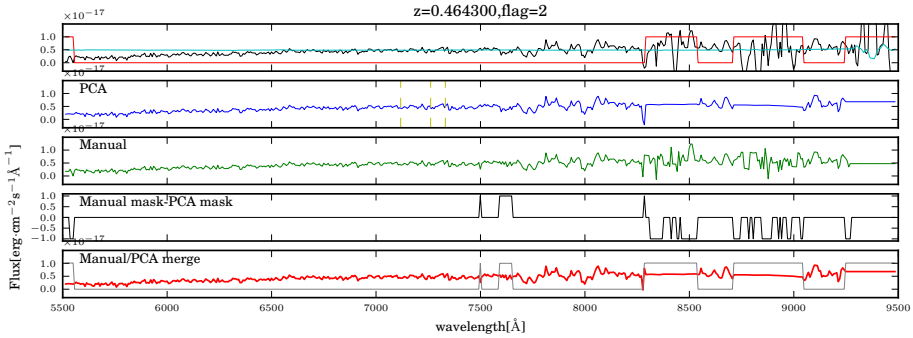
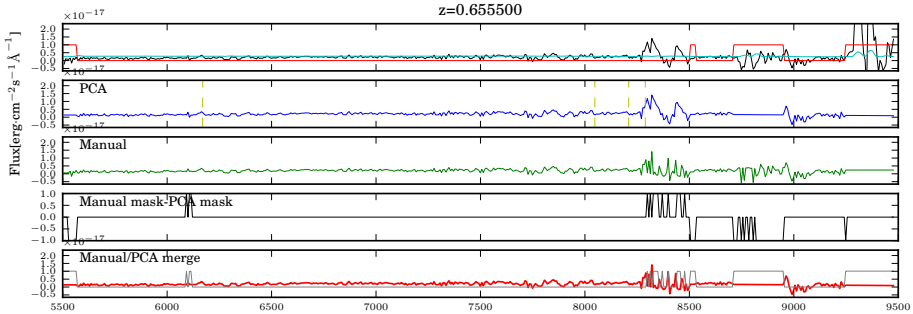


Figure 7.21: Eight example spectra with flag 2







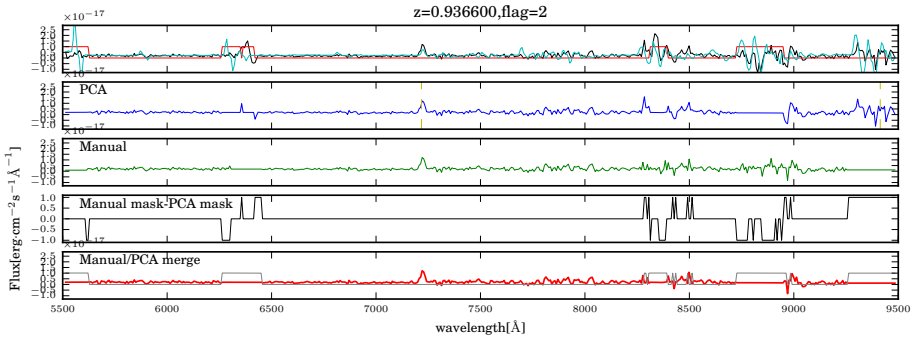
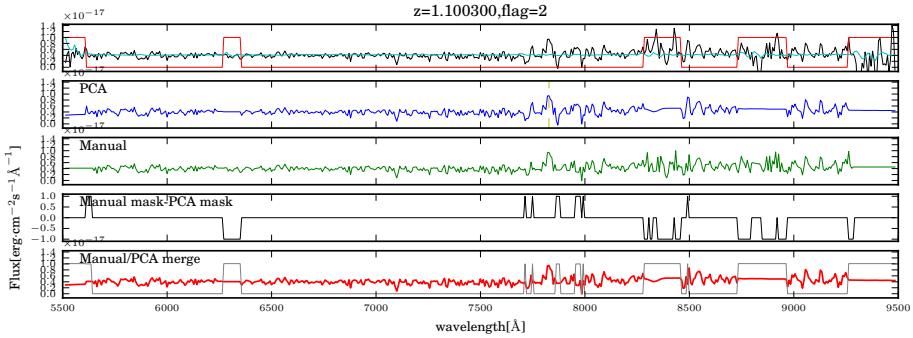
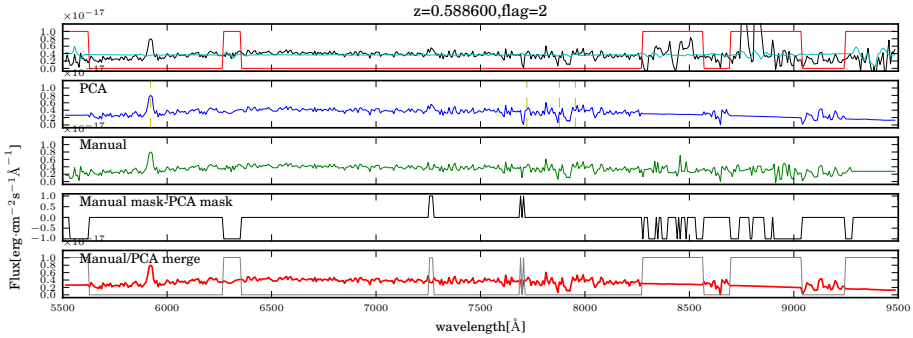
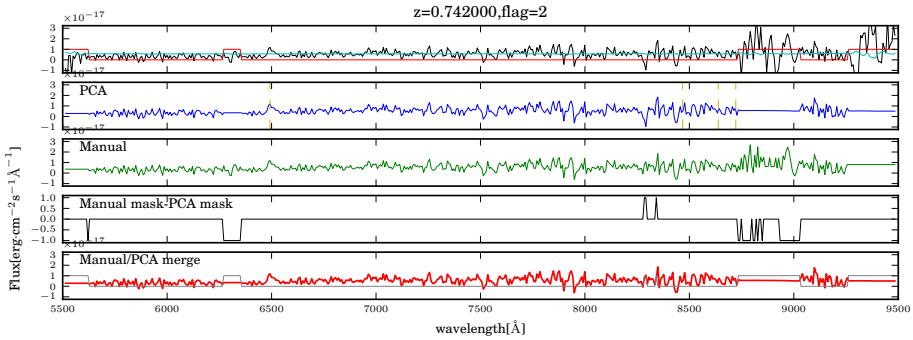


Figure 7.22: Continued... Eight example spectra with flag 2







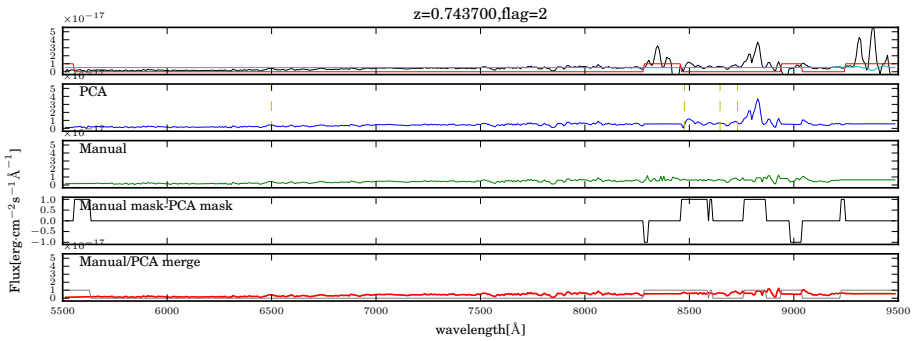
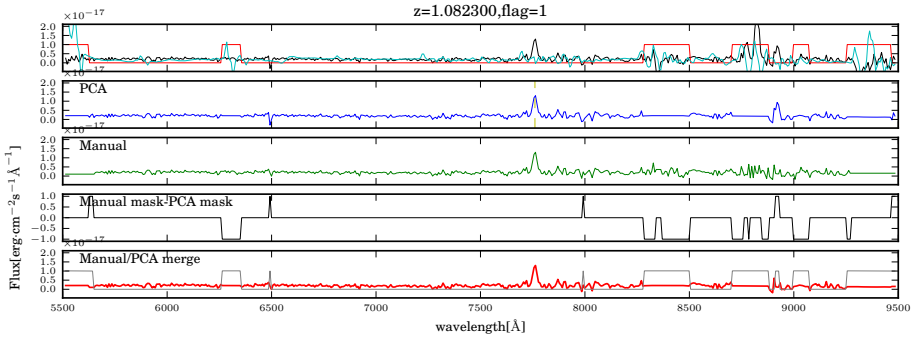
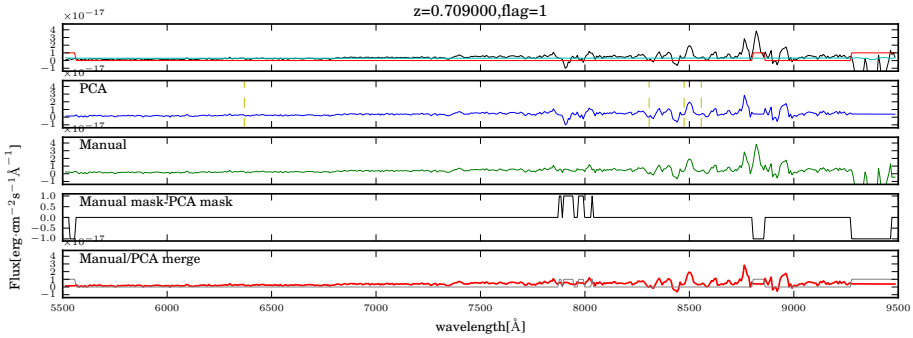
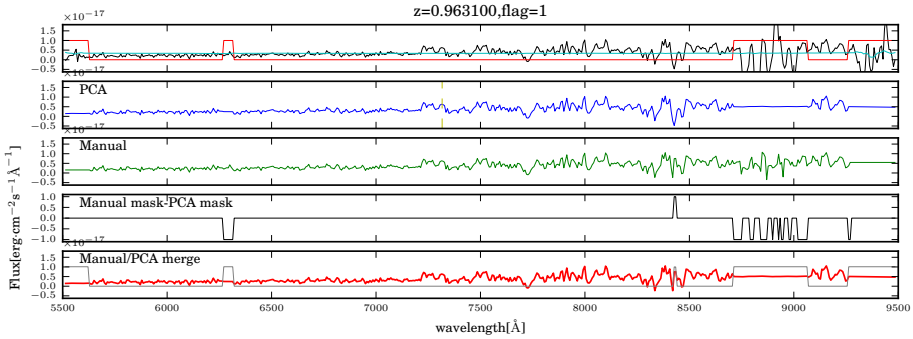


Figure 7.23: Eight example spectra with flag 1







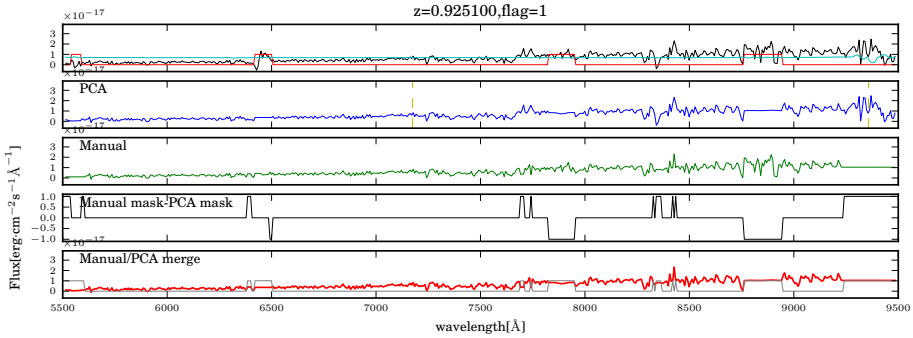
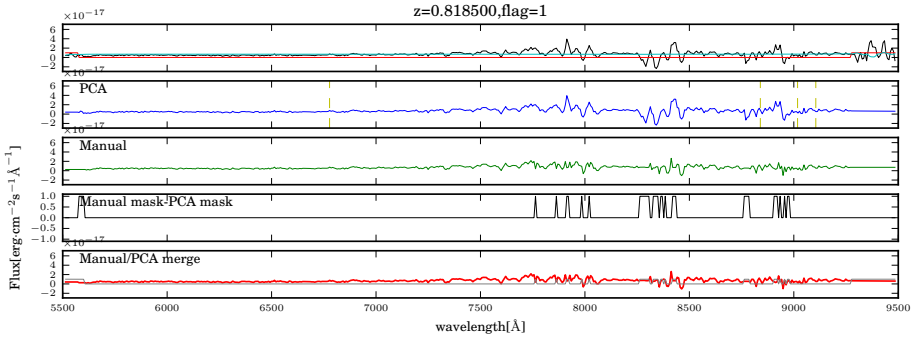
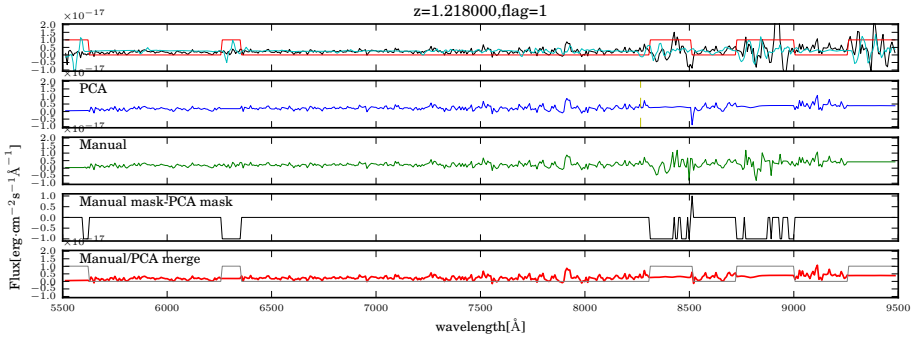
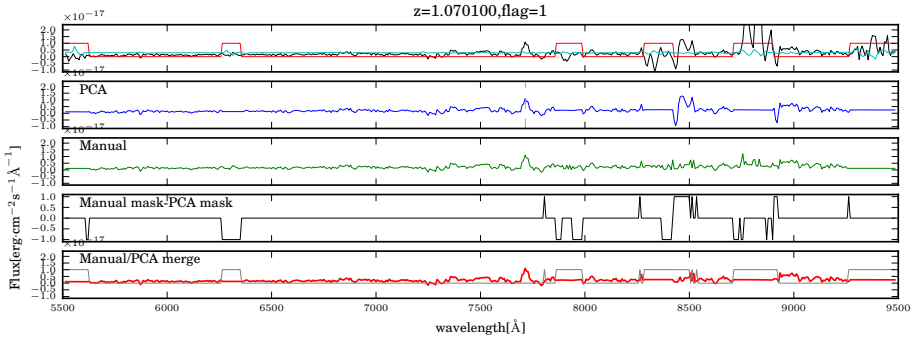


Figure 7.24: Continued... Eight example spectra with flag 1







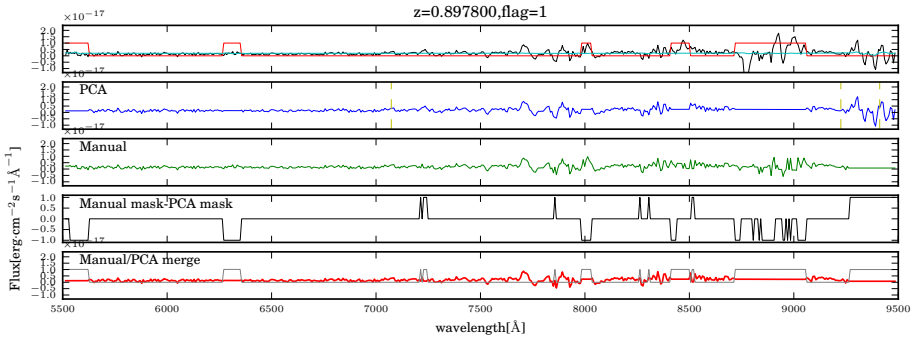


Figure 7.25: Eight example spectra with flag 0

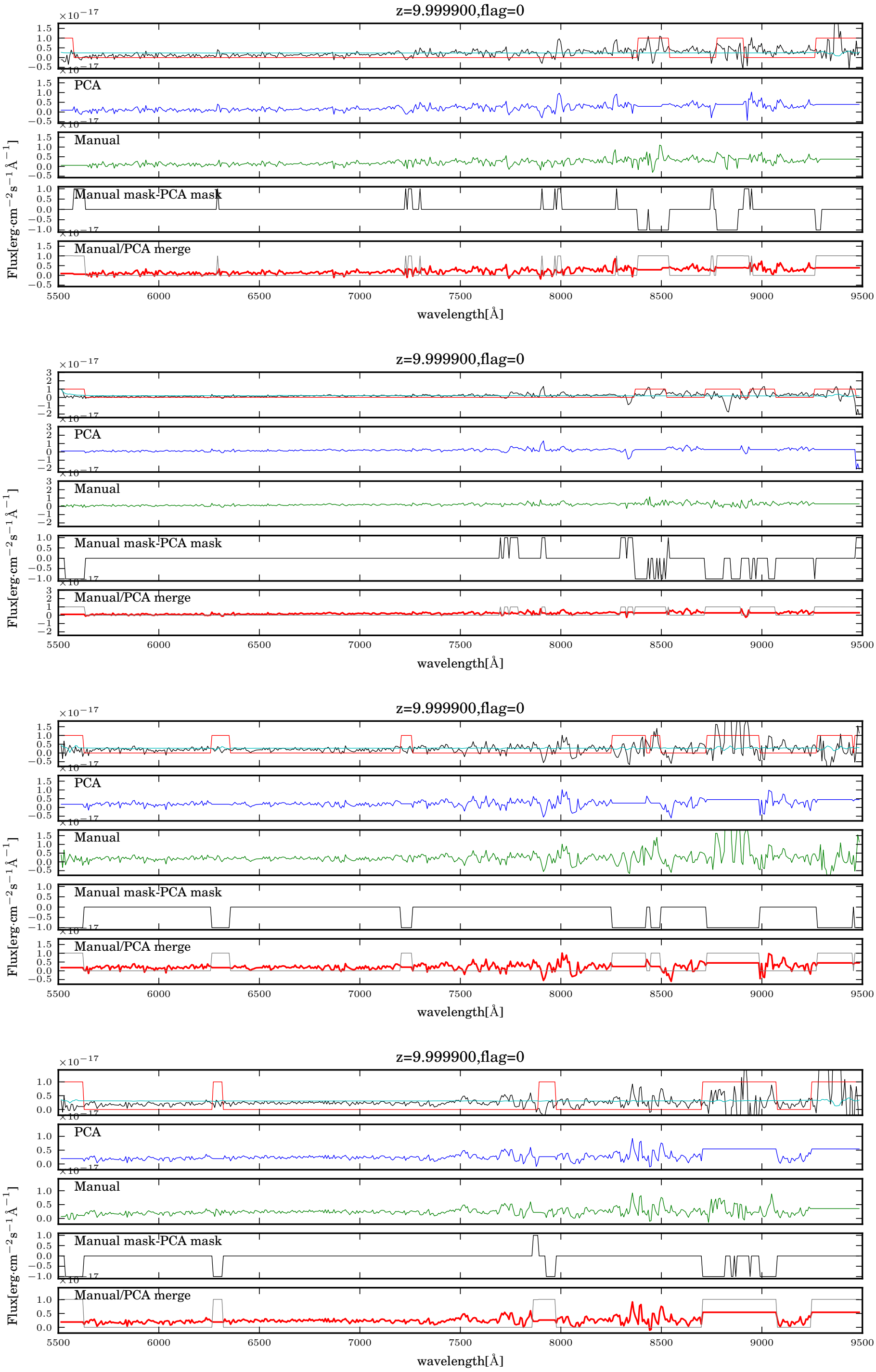


Figure 7.26: Continued... Eight example spectra with flag 0

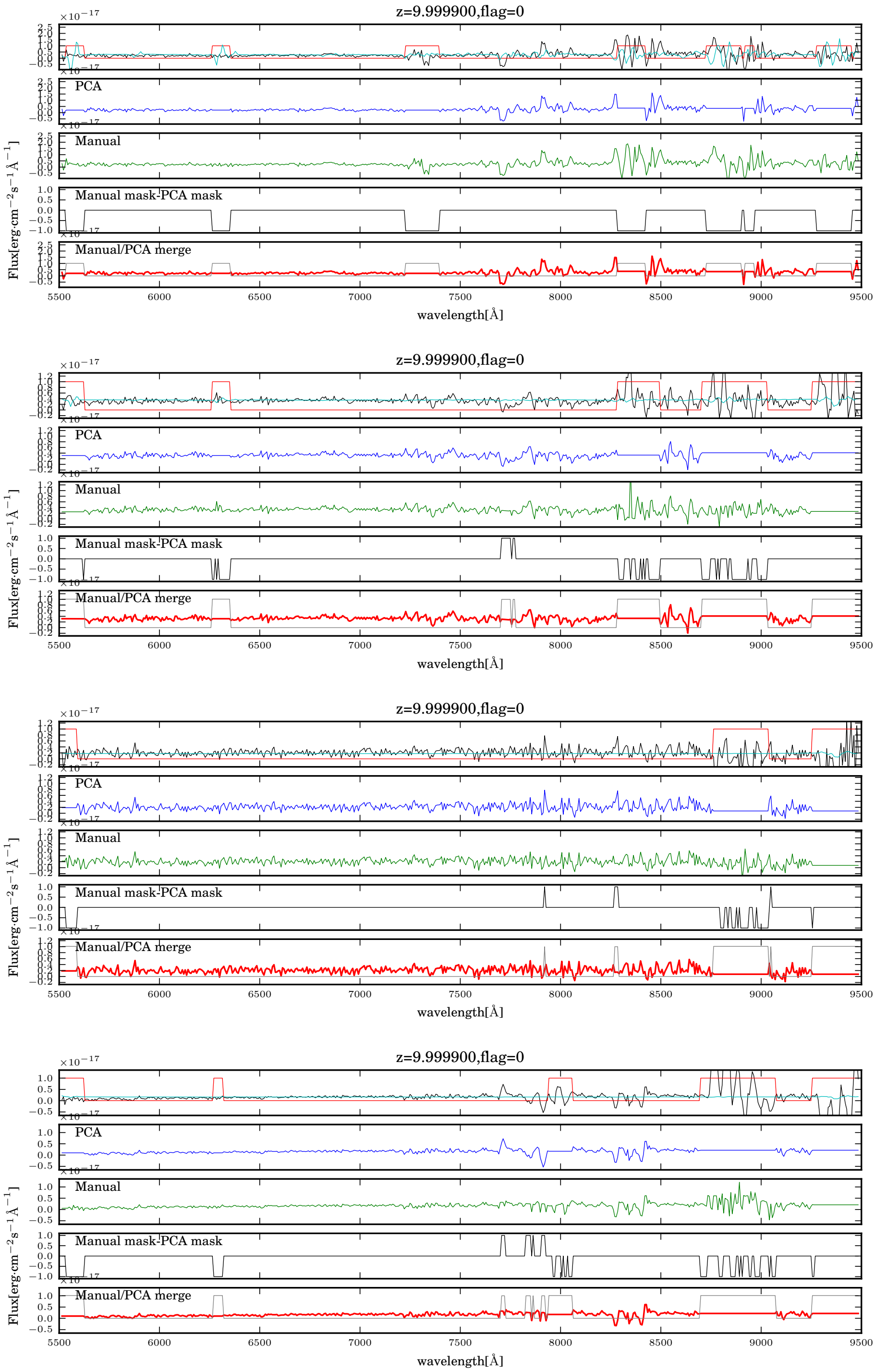
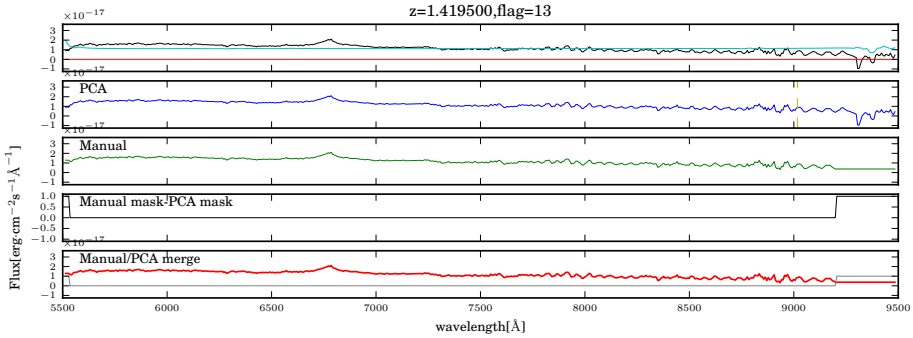
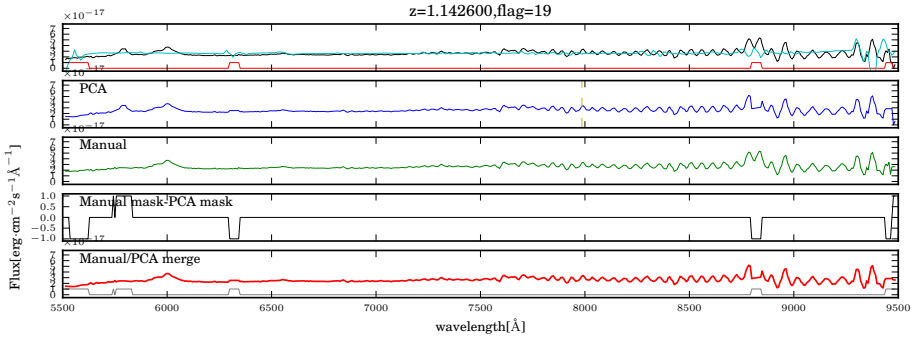
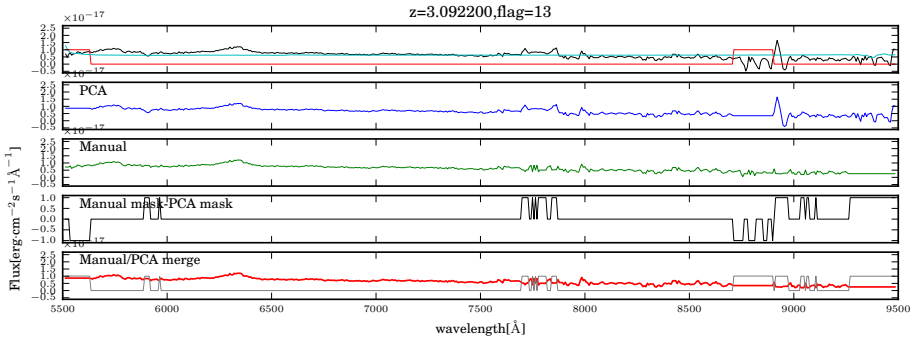


Figure 7.27: Eight example spectra with flag 10-19







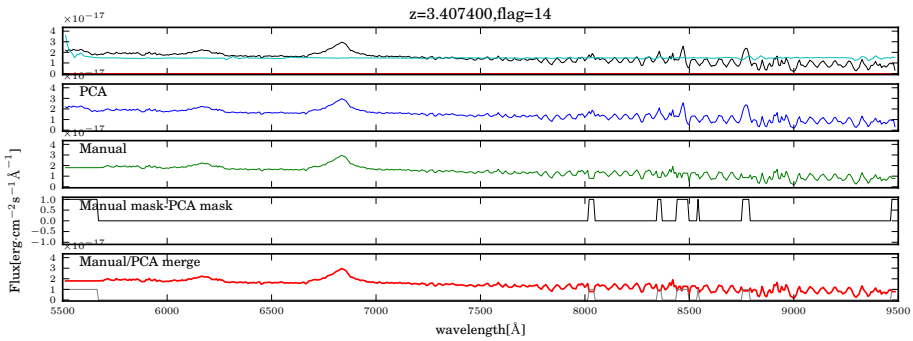
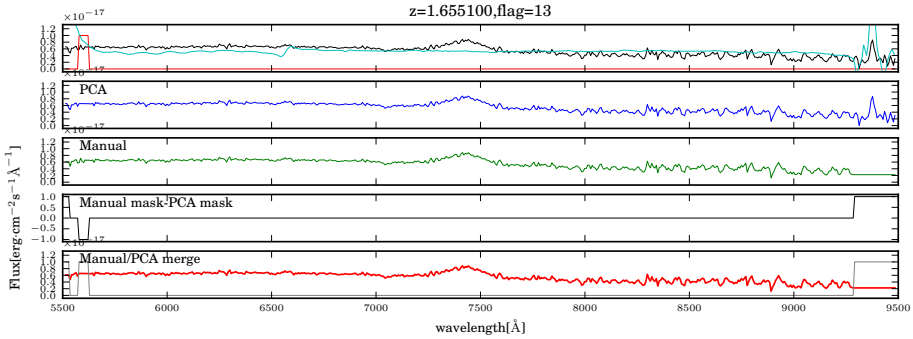
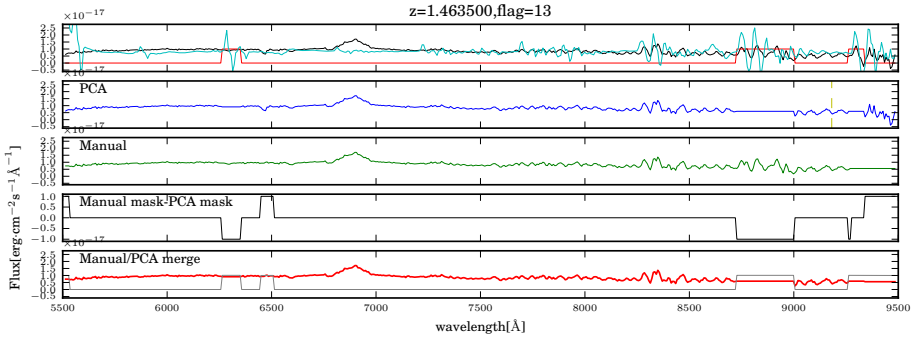
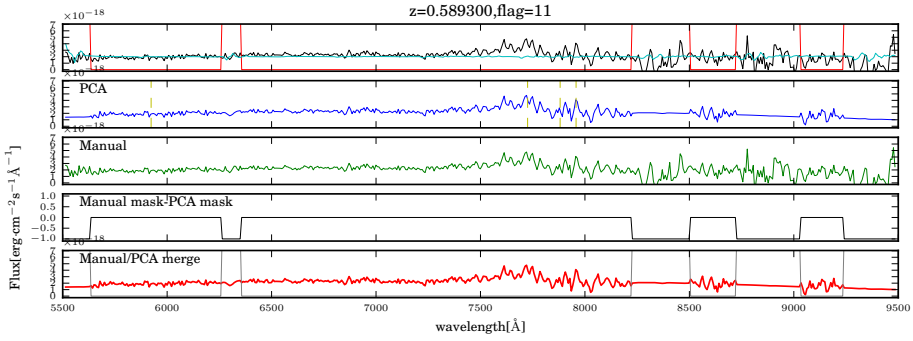


Figure 7.28: Continued... Eight example spectra with flag 10-19







Bibliography

Abbas U., et al., 2010, MNRAS, 406: 1306
Albert J., et al. 2008, Nucler Instruments and Methods in Physics Research A, 588, 424
Ascasibar Y., Sánchez Almeida J., 2011, MNRAS, 415, 3: 2417-2425
Baldry I.K., 2004, ApJ 600: 681
Baldwin J.A., Phillips M.M., Terlevich R., 1981, Astronomical Society of the Pacific, 1981PASP, 93, 5B
Beals C.S., 1953, Publications of the Dominion Astrophysical Observatory, Vol. 9, p. 1
Balogh M.N., Navarro J.F., Morris S.L., 2000, ApJ, 540, 113
Bell E.F., 2004, ApJ, 608: 752
Blanton M.R., et al. 2003, ApJ, 594: 186
Brammer G.B., et al., 2009, ApJ, 706L: 173
Bromley B.C., Press W.H., Lin H., Kirshner R., 1998, ApJ, 505: 25
Brown M.J.I., Dey A., Jannuzi B.T., Lauer T.R., Tiede G.P., Mikles V.J., 2003, ApJ, 597: 225
Bruzual G., Charlot S., 2003, MNRAS, 344,4: 1000-1028
Budavari T., et al. 2003, ApJ, 595, 59
Cabré A., Gaztañaga E., 2009, MNRAS, 393: 1183
Calzetti D., Kinney A.L., Storchi-Bergman T., 1994, ApJ, 429: 582
Cardelli J.A., Clayton G.C., Mathis J.S., 1989, ApJ, 345: 245, 256
Cimatti A., et al., 2002, A&A, 381L: 68
Coil A.L., Newman J.A., Cooper M.C., Davis M., Faber S.M., Koo D.C., Willmer C.N.A., 2006, ApJ, 644: 671
Coil A.L., et al., 2008, ApJ, 672: 153
Colless M., Dalton G., Maddox S., et al. 2001, MNRAS, 328, 1039
Connolly A.J., Szalay A.S., Bershady M.A., Kinney A.L., Calzetti D., 1995, ApJ, 110,3: 1071-1082
Connolly A.J., Szalay A.S., 1999, ApJ, 117: 2052-2062
Coupon J., et al., 2012, A&A, 542A: 5
Cristianini N., Shawe-Taylor J., 2000, An introduction to supprot vector machines: and other kernel-based learning methods (Cambridge University Press)
Cucciati O., Iovino A., Marinoni C., et al. 2006, A&A, 458, 39
Daddi E., et al., 2003, ApJ, 588: 50
Davidzon I., Bolzonella M., Coupon J., et al. 2013, A&A, 558, A23
Davis M., Geller Margaret J., 1976, ApJ, 208: 13
de la Torre S., et al., 2011, MNRAS 412, 2: 825-834
de Vaucouleurs G. 1961, ApJS, 5, 233
de Vaucouleurs G., 1962, Problems of Extra-Galactic Research, Proceedings from IAU Symposium no. 15. Edited by George Cunliffe McVittie. International Astronomical Union Symposium no. 15, Macmillan Press, New York, p.3
Dobos, L., Csabai, I., Yip, C.-W., et al., 2012, MNRAS, 420: 1217
Everson R. & Sirovich L., Karhunen-Loeve procedure for gappy data, 1995, J. Opt Soc. Am. A J2, 1657-1664
Faber S.M., et al., 2007, ApJ, 665: 265
Ferreras, I., Pasquali, A., de Carvalho, R. R., de la Rosa, I. G., & Lahav, O., 2006, MNRAS, 370: 828
Folkes S., Lahav O., Maddox S.J., 1996, MNRAS, 283: 651
Francis P.J., Hewett P.C., Foltz C.B., Chaffee F.H., 1993, ApJ, 398: 476

Franzetti P., et al., A&A, 465: 711
Fritz A., et al., 2014, A&A, 563: A92
Galaz G., de Lapparent V., 1998, A&A, 332: 459
Garilli B., Fumana M., Franzetti P., Paioro L., Scodeggio M., Le Fèvre O., Paltani S., Scaramella R., 2010, PASP
Gao D., Zhang Y. & Zhao Y., 2009, Research in Astron. Astrophys. 2009 Vol. 9 No. 2, 220–226
Garilli B., Guzzo L., Scodeggio M., Bolzonella M., et al., 2014, A&A, 562: A23
Gavazzi G., Boselli A., Donati A., Franzetti P., Scodeggio M., 2003, A&A, 400: 451
Giovanelli R., Haynes M., Chincarini G.L., 1986, ApJ, 300: 77
Goranova Y., Hudelot P., Magnard F., et al., 2009, The CFHTLS T0006 Release, `http://terapix.iap.fr/cplt/table_syn_T0006.html`
Graves Genevieve J., Faber S.M., Schiavon Ricardo P., Yan Renbin, 2007, ApJ, 671: 243
Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1994a, ApJ, 426, 340
Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1994b, Vistas Astron., 38, 293
Guzzo G., Strauss M.A., Fisher K.B., Giovanelli R., Haynes M.P., 1997, ApJ, 489: 37
Guzzo L., Scodeggio M., Garilli B., et al., 2014, A&A, 556A.108G
Guzzo L., et al. 2008, Nature 451, 541-544
Hubble E.P., 1926, ApJ 64, 321
Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971
Juneau S., Dickinson M., Alexander D.M., Salim S., 2011, ApJ, 736:104
Karhunen H.., 1947, Ann. Acad. Science Fenn, Ser. A.I. 37
Kinney A.L., Calzetti D., Bohlin R.C., McQuade K., Storchi-Bergman T., Shmitt H.R., 1996, ApJ, 467: 38
Le Fèvre O., et al. 2005, A&A, 439: 845
Lilly S.J., Le Brun V., Maier C., et al. 2009, ApJS, 184, 218
Lamareille F., 2010, A&A, 509, A53
Loève M., 1948, Processus Stochastiques et Mouvement Brownien (Hermann, Paris, France)
McDonald P., Seljak U., 2009, jcap, 10, 7
Maddox S.J., et al., Large Scale Structure: Tracks and Traces, ed. Muller (Singapore: World Scientific), 99
Madgwick D.S., et al. 2002, MNRAS, 333: 133
Madgwick D.S., Coil, A.L., Conselice C.J., Cooper M.C., et al. 2002, MNRAS, 333: 133
Marchetti A., Granett B. R., Guzzo L., et al. 2013, MNRAS, 428, 1424
Małek K., Solarz A., Pollo A., et al. 2013, A&A, 557, A16
Marulli F., Bolzonella M., Branchini E., Davidzon I., de la Torre S., et al., 2013, Astronomy & Astrophysics, Volume 557, id.A17
Mendez A.J., Coil A.L., Lotz J., Salim S., Moustakas J., Simard L., 2011, ApJ, 736, 2: 110-133
Meneux B., et al., 2008, A&A, 478: 299
Meneux B., et al., 2009, A&A, 505: 463
Mignoli M., et al., 2009, A&A, 493, 39:49
Morgan W.W., Mayall N.U., 1957, Publications of the Astronomical Society of the Pacific, 69, 409 : 291-303
Murtagh F., Heck A., 1987, Multivariate Data Analysis. Reidel, Dordrecht
Norberg P., et al., 2002, MNRAS, 332: 827
Odewahn S.C., Stockwell E.B., Pennington R.L., Humphreys R.M., Zumach W.A. 1992, AJ, 103, 318
Phleps S., Peacock J.A., Meisenheimer K., Wolf C., 2006, A&A, 457: 145
Ronen S., Aragón-Salamanca A., Lahav O., 1999, MNRAS, 303: 284
Rogers, B., Ferreras, I., Lahav, O., Bernardi M., Kaviraj S., Yi S.K., 2007, MNRAS, 382: 750
Rogers, B., Ferreras, I., Pasquali, A., Bernardi M., Lahav O., Kaviraj S., 2010, MNRAS, 405: 329
Roweis S., EM algorithms for PCA and SPCA, 1997 Neural Inf. Proc. Syst, 10, 626-632
The Hubble Atlas of Galaxies. (Carnegie Institution of Washington)
Sandage A.R., 1975, Stars and Stellar Systems, Vol. 9, Galaxies and the Universe, ed. A.R. Sandage, M.Sandage, and J. Kristian (Chicago: University of Chicago Press), p. 761ff
Shawe-Taylor J., Cristianini N., 2004, Kernel methods fro pattern analysis (Cambridge University Press)
Singh H.P., Gulati R.K., Gupta R., 1998, MNRAS, 295: 312
Sodré, L. Jr., & Cuevas, H. 1994, Vistas in Astronomy, 38, 286
Sodré L., Cuevas H., 1997, MNRAS, 287: 137
Solarz A., Pollo A., Takeuchi T.T., 2012, A&A, 541, A50
Storrie-Lombardi M.C., Irwin M.J., von Hippel T., Storrie-Lombardi L.J., 1994, Vistas Astron., 38, 331
Strateva I., Ivezić Ž., Knapp G. R., et al., 2001, AJ, 122: 1861
Tinsley B.M.: Astron. Astrophys. 20, 383 (1972)

Tojeiro R., Percival W.J., 2010, MNRAS 405: 2534-2548
Tojeiro R., Percival W.J., Heavens A.F., Jimenez R., 2010, MNRAS 413: 434-460
Tojeiro R., et al., 2012, preprint (astro-ph/1202.6241)
von Hippel T., Storrie-Lombardi L. J., Storrie-Lombardi M.C., Irwin M., 1994, MNRAS, 269, 97
Wake D.A., Nichol R.C., Eisenstein D.J., et al., 2006, MNRAS 372: 537-550
Weiner B.J., et al., 2005, ApJ, 620: 595
Wild V., Hewett P.C., 2005, ApJ, 358:1083W
Woźniak P.R., Williams S.J., Vestrand W.T. & Gupta V., 2004, AJ, 128,2965
Yip C.W., Connolly A.J., et al., 2004, ApJ, 128: 585-609
York D.G., et al., 2000, AJ, 120: 1579
Zehavi I., et al., 2011, ApJ, 736: 59-89
Zhang, Y., & Zhao, Y. 2003, ChJAA (Chin. J. Astron. Astrophys.), 3, 183
Zhang, Y., & Zhao, Y. 2004, A&A, 422, 1113
Zhao, Y., & Zhang, Y. 2007, Advances in Space Research, 41, 1955

List of Publications

Primary author

The VIMOS Public Extragalactic Redshift Survey (VIPERS). Spectral classification through Principal Component Analysis
A. Marchetti, B.R. Granett, L. Guzzo, A. Fritz, B.Garilli, et al., 2013, MNRAS, 428, 1424-1437

Cosmological Data Analysis of f(R) Gravity Models
Z. Gironés, A. Marchetti, O. Mena, C. Peña Garay, N. Rius, JCAP11(2010)004

Participation as VIPERS member

The VIMOS Public Extragalactic Redshift Survey (VIPERS). Never mind the gaps: comparing techniques to restore homogeneous sky coverage
Cucciati, O., & VIPERS Team, 2014, A&A, 565, A67

The VIMOS Public Extragalactic Survey (VIPERS): First Data Release of 57 204 spectroscopic measurements
B. Garilli, L. Guzzo, M. Scodeggio, M. Bolzonella, U. Abbas, et al., 2014, A&A, 562, A23

The VIMOS Public Extragalactic Redshift Survey (VIPERS). An unprecedented view of galaxies and large-scale structure at 0.5$<$z$<$1.2
Guzzo, M. Scodeggio, B. Garilli, B. R. Granett, U. Abbas, et al., 2014, A&A, 566, A108

The VIMOS Public Extragalactic Redshift Survey (VIPERS): an unprecedented view of galaxies and large-scale structure halfway back in the life of the Universe
L. Guzzo, & VIPERS team, 2013, The ESO Messenger, 151, 39

The VIMOS Public Extragalactic Redshift Survey (VIPERS): Ω_{m_0} from the galaxy clustering ratio measured at $z \sim 1$
J. Bel, C. Marinoni, B. R. Granett, L. Guzzo, J. A. Peacock, et al., 2014, A&A, 563, A37

The VIMOS Public Extragalactic Redshift Survey (VIPERS). Galaxy clustering and redshift-space distortions at z=0.8 in the first data release
S. de la Torre, L. Guzzo, J. A. Peacock, E. Branchini, A. Iovino, et al., 2013, A&A, 557, A54

The VIMOS Public Extragalactic Redshift Survey (VIPERS). A Support Vector Machine classification of galaxies, stars and AGNs
K. Malek, A. Solarz, A. Pollo, A. Fritz, B. Garilli, M. Scodeggio, et al., A&A, 557, A16

The VIMOS Public Extragalactic Redshift Survey (VIPERS). Luminosity and stellar mass dependence of galaxy clustering at 0.5$<$z$<$1.1
F. Marulli, M. Bolzonella, E. Branchini, I. Davidzon, S. de la Torre, et al., 2013, A&A, 557, A17

The VIMOS Public Extragalactic Redshift Survey (VIPERS). A precise measurement of the galaxy stellar mass function and the abundance of massive galaxies at redshifts 0.5<z<1.3
I. Davidzon, M. Bolzonella, J. Coupon, O. Ilbert, S. Arnouts, et al., A&A, 558, A23

Acknowledgments

A special acknowledgment to Ben Granett for supporting, helping and pushing me, for working with me, giving many suggestions, directions and ideas, as if he was an official supervisor.

Thanks to Gigi for trusting me, even if I had no expertise in his working field (and still have not, having worked on PCA!) and rewarding me after the end of my fellowship, making me feel my work was really worth.

Thanks to Marco Bersanelli, for allowing me to work with Gigi's group all of this years, keeping interested in my developments, achievements and future perspectives.

Thanks to Andy Connolly, for refereeing this work with precision and insight.

Thanks to all the VIPERS team, especially Bianca, Marco and Angela for the suggestions, the interest and the help in carrying this work on.

Thanks to the LBT Team, for exempting me from going to Arizona, to let me finalize my thesis.

Thanks to Andrea Zanzani, for being the PhD student's guardian angel, always helping (often even before being asked) and supporting with documents, deadlines, examinations, courses and all the PhD bureaucracy.

Thanks to Petra for being a friend, more than the secretary we dreamed of for long.

A special thank to my present group, for encouraging and helping me to finish this chapter of my studies. Thanks to Sascha for reading my papers so carefully. Thanks to Marco (F.) for compensating for my lacks in my new job (his old job), while I was busy with the thesis, and for being with me during my first trip to Arizona. Thanks to Paolo and Paolo, for the job they will do in adding part of my work to a reduction pipeline. Thanks to Paolo, Sasha, Letizia and Kasia for your friendship and our relaxing lunches outside.

Thanks to my previous officemates, Matteo, a real genius and a potential wonderful teacher, Davide, a dear friend, and guitar player in our OAB country band, and of course humble and brilliant scientist. Thanks to Lorenzo (wonderful OABand country singer), Faizan, Stefano, Melita, Jihan-Hua, Adam, Maria Grazia (keep on singing!!), Tullia, Elisabetta, Martino, Coti, Perri, Giuseppe, Riccardo and all the Merate people. A special thank the OAB choir: Giacomo, Cinzia, Cristina, Matteo, Clara, Maria Rosa and Daniele; all these musical expressions born during my PhD were so infinitely precious, for someone like me who cannot live without music.

Thanks to Cristina and Samuele, and all my dearest friends through these years, Michele (praying for me from Siberia), Filippo (Pips), Stefano, Valentina, Marina, Valeria, Sonja, MariaChiara, Nicholas and Chiara, Caterina, Chiara and Raffaele, Lucia and Paolo, Annalisa and Gianluca.

Thanks to the wonderful choir "Sursum Corda", born during my PhD from my passion, my husband's and a couple of friends', for those years of truly beautiful music, capable of touching the heart of people (and ours) even through our imperfect efforts.

Thanks to the 'Wedding Brothers" band, for the funny (and profitable!!) performances played in those years.

Finally, the dearest thank to my family: Filippo, for supporting me over these years, giving me advices and strength, to Giacomo, now 2 years old, born during my PhD and growing with a very-busy mother! Thanks to my brother, cousins and to all the "grannies", especially to my mum and dad, for the irreplaceable help and support. Without you it wouldn't have been possible to reach this achievement.