Eötvös Loránd Tudományegyetem
Természettudományi Kar

Korai és késői típusú galaxispopulációk vizsgálata a Sloan Digital Sky Survey alapján

Doktori értekezés

Győry Zsuzsanna

Témavezető: **Csabai István**
Komplex Rendszerek Fizikája Tanszék

Budapest
2008

Győry Zsuzsanna
Komplex Rendszerek Fizikája Tanszék
1117 Budapest, Pázmány Péter sétány 1/A
gyory@complex.elte.hu
Eötvös Loránd Tudományegyetem
Természettudományi Kar
Fizika Doktori Iskola
Részecskefizika és Csillagászat Doktori Program
Doktori Iskola vezetője: dr. Horváth Zalán
Programvezető: dr. Csikor Frerenc
Témavezető: dr. Csabai István
2008. szeptember 1.

Eötvös University
Faculty of Science

Exploring the early and late-type galaxy populations using the Sloan Digital Sky Survey

PhD thesis

Zsuzsanna Győry

Advisor: **István Csabai**
Department of Physics of Complex Systems

Budapest
2008

Contents

Chapter 1

Introduction

1.1 Exploring the Universe through galaxies

Galaxies are the most common homes of the luminous matter in the Universe. Their light originates from stars and is modified by interstellar matter to get the final form that travels over cosmic distances before being observed. The properties of their shape, light, as well as spatial distribution have been widely studied for centuries, as these might give us clues to understanding our Universe, its composition, fundamental processes and evolution.

The galaxies have been primarily classified accodring to their morphological properties. Elliptical or spherical galaxies with smooth structure are historically called early types. They were ordered into a sequence on the basis of their ellipticity. The Hubble classification assings to them ratings from E0 to E7, in the direction of the increasing elongation. Galaxies also having a flat disk besides the central bulge, with prominent structure like spiral arms or bars in the disk, are usually referred to as late types. In the Hubble classification scheme, spiral galaxies are listed as Sa, Sb, Sc, the spiral structure becoming tighter, the bulge-to-disk ratio decreasing along the sequence. The lenticular or S0 galaxies are an intermediate form between ellipticals and spirals, they are generally considered as early types. Galaxies not meeting any of the above classification criteria are irregulars, labeled Irr.

One can observe a close relation between the shape and the light of the galaxies. The early type galaxies have typically redder color, the late types bluer. The differences in colors of galaxies reflect their different stellar populations. Early type, elliptical systems consist primarily of old red stars. The blue color of the late type spirals originates from young stars generated by ongoing (or very recent) star formation. The morphology of the

galaxies reflects their dynamics: the spheroid of stars in early type galaxies is supported largly by random motions, while the disk of stars in late type galaxies is rotationally supported. There have been different hypotheses on the origin of the variability observed in types and the correlations of dynamical and stellar properties within them. These ranged from viewing the Hubble sequence as an evolutionary sequence from early to late types, through assuming no evolutionary connection, but environmental effects only, to theories with late type systems being progenitors of early types. The evolution of galaxies is still intensely studied.

The first step towards understanding the physics of galaxies is to place them into a cosmological context. In the present day standard cosmological model, cold dark matter (CDM) dominates gravitation on large scales, and a cosmological constant acts as a negative pressure. On this background, after having started its evolution at the Big Bang, having passed through the early stages of thermalization, particle formation, nucleosynthesis and recombination, the Universe expands, cools, and forms large scale structure from the primordial fluctuations. According to the standard scenario, galaxies are formed in the most overdense regions of the cold dark matter, thus tracing its distribution. Therefore studying the large scale structure of the spatial distribution of galaxies is of great importance in cosmological tests.

For these studies we need to measure the distances of galaxies. As the Universe is expanding, the light of distant objects is shifted towards red. The redshift increases with distance following a simple relation inferred from the cosmological model, at the smallest distances this translates into the linear Hubble's law. The redshifts of galaxies are usually determined from their spectra, based primarily on spectral lines. By measuring redshifts we obtain the information of distances. Provided a large enough statistics, these data enable detailed studies of large scale structure, correlation functions, precision measurements of distributions of types, luminosity function, and, as objects at larger distances are observed in earlier epochs, even direct observations of evolution.

1.2 The Sloan Digital Sky Survey

With the advent of the large sky surveys of the recent years, these information have become available on galaxy samples of unprecedented sizes, sampling statistically significant volumes of the Universe. The Sloan Digital Sky Survey (SDSS) (Szalay 1999; York *et al.* 2000) is one of the most challenging of them. The aim was to construct a 3D map of 25% of the sky. It is a digital photometric and spectroscopic survey, covering one quarter

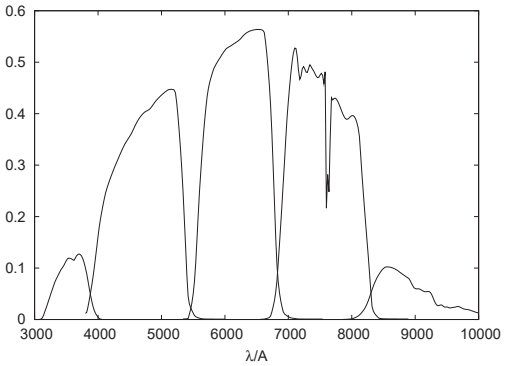

Figure 1.1: The SDSS filter profiles, from left to right: u, g, r, i, z.

of the Celestial Sphere toward the Northern Galactic cap, and produce a smaller area ($\sim$225 deg^2) but much deeper survey toward the Southern Galactic cap. Until now it has published more than 800,000 galaxies having spectroscopic measurements, and over a hundred million galaxies with photometric observations. The digital images are taken by a 2.5 m telescope in five broad-band filters: u, g, r, i, z(Fukugita *et al.* 1996). The filter curves are plotted in Figure 1.1, they cover the wavelength range $3,000 - 10,000$Å.

The images are then processed by the photometric pipeline which identifies the sources, measures their magnitudes, sizes, probabilities of being stars or galaxies and other parameters. Based on the output of the photometric pipeline, the brightest galaxies are targeted for spectroscopy. The spectra are taken through 3 arcsec diameter optical fibers plugged into holes in an aluminium plate; over 640 spectra can be recorded at the same time. The spectroscopic galaxy survey consists of the main sample, with a magnitude limit $r < 17.77$ with a mean redshift $z = 0.1$ (Strauss *et al.* 2002), and a deeper sample of the luminous red galaxis (LRG, Eisenstein *et al.* (2001)) sample with $r < 19.5$ nearly volume limited with $z < 0.4$. The observational strategies make the SDSS robust in almost all areas of extragalactic research.

1.3 Overview of the thesis

In my thesis, I address various aspects of the basic properties of galaxies already mentioned in Section 1.1 as derived primarily based on their photometric data. The differences and connections between the galaxy types regarding their distribution, evolution, structrure,

stellar population, possible photometric applications are specifically in focus.

In Chapter 2 we study in detail the optical color distribution of the SDSS photometric sample galaxies. We show that the distribution of galaxies in the $g-r$ vs. $u-g$ color–color diagram is strongly bimodal with an optimal color separator of $u-r=2.22$. The bimodality and this simple cut is justified by Bayesian probability analysis in the four dimesional color space. As expected from their different stellar populations, the two classes correlate with visual morphology, spectral classification and other morphologically derived parameters. While it has been well known that late type galaxies are bluer than early type galaxies, this was the first detection of a local minimum in their color distribution. Evolutionary effects in color and number densities different in red and blue types were detected too.

These findings support the idea that the red and blue subpopulations are not just incidental two halves of a continuous sequence, but rather represent two substatnially different categories of the underying physical properties. Since the discovery of the strong color bimodality, numerous studies of galaxy types have been based on it, investigating the distribution and evolution of galaxies. From deep redshift studies of the evolution of the red and blue luminosity function it has become clear that at least half of the red sequence galaxies have come to their present form since $z\sim 1$. These studies showed that there must be an evolutionary transition from early to late types to reproduce the changes in densities of both types with time. The transition involves merging of later types and the subsequent truncation of star formation.

One of the key qualitative predictions of the merger hypothesis for the origin of early-type (elliptical and lenticular) galaxies is that tidally-induced asymmetric structure should correlate with signatures of a relatively young stellar population. Such a signature was found by Schweizer & Seitzer (1992). In Chapter 3 we revisit this issue with a substantially larger sample of galaxies. We parameterize tidal structure using a repeatable algorithmic measure of asymmetry, and correlate this with color offset from the early-type galaxy color–magnitude relation. We find a correlation, in the sense that more richly tidally-structured galaxies are bluer than less asymmetric early-type galaxies; we demonstrate for the first time that this effect is driven by age variations. More asymmetric objects tend to have younger stellar population, metallicity does not vary with asymmetry. The existence of this correlation between asymmetry and age of the stellar population is in qualitative agreement with the expectations from the merger hypothesis. We address this question quantitatively by statistical comparison with merger motivated spectral models. With revised approach based on the recent years' observational evidence,

we readdress the historical question of the 'King gap', the missing link between the present day old ellipticals and the few very young merger remnants.

Galaxies with blue colors indicating star formation have usually two physically different components appearing together in their spectra. The spectrum of young stellar populations and emission lines emitted by interstellar gas nebulae are both signatures of recent star formation. The quantitative analysis of their connection is important from the aspect of modelling galaxy spectra and interpretation of photometric data. In Chapter 4, we present a statistical analysis of the nebular emission and its variation with spectral type. We explore the dimensionality of the emission line pattern and present an efficient dimensional reduction to two or three parameters. Making use of this technique we investigate the correlations of the stellar population characteristics and the emission lines. We establish an empirical relation between them. It can be used for emission line corrections of spectral models which can play an important role in photometric applications.

The most motivating application of this study was the photometric redshift estimation. With this method, redshift information can be obtained from photometric data for a much larger and deeper set of galaxies than the spectroscopic sample. We overview several aspects of this technique in Chapter 5. We discuss a physically motivated approach of using model spectra as spectral templates for comparison with the rough spectra of galaxies obtained with photometry, and thus deriving an estimate of their redshifts (Section 5.2). Using specially selected reference sets, we test and recalibrate the hybrid redshift estimator applied in SDSS. Finally we complete the method with an application of the emission line correction technique introduced above. We end up with a new version of photometric redshift estimator that is more accurate at higher redshifts and more homogeneous in spectral type. The changes have been implement the in the SDSS since the fourth data release (DR4) and can be used in a number of scientific applications.

Chapter 2

Distribution of galaxies in color space

2.1 Introduction

It has been known at least since the late 1930s that colors of galaxies reflect their dominant stellar populations and thus correlate with morphology (Humason 1936, Hubble 1936). Morgan and Mayall (1957) examined the spectra of 47 nearby galaxies and found that stellar systems with spectra dominated by A, A+F, and F stars[1] are exclusively classified as Sc and Irr morphologically, F+G dominated stellar systems correspond to Sb galaxies, and the K stellar systems are a mix of predominantly early type (E, S0, Sa) galaxies, with a sizeable fraction of Sb spirals. De Vaucouleurs (1961) used a sample of 148 galaxies to establish the dependence of galaxy color on morphological type. Since then, studies of the color distribution of galaxies have helped reveal their dominant stellar populations and star formation histories. Typical studies of galaxy colors in recent years have been based on samples of ∼1,000 galaxies. Fioc & Rocca-Volmerange (1999) used optical and near infrared colors of 1,000 galaxies to establish relations between the colors, morphological types, inclinations or shapes, and the intrinsic luminosities of galaxies. Ferreras *et al.* (1999) used color–magnitude and color–color analysis of HST photometry of ∼1,000 galaxies to infer the existence of non-negligible star formation in ellipticals and bulges at medium redshift (z ∼0.2). Brown *et al.* (2000) studied the dependence of clustering of galaxies on color. They used a catalog of $\sim 4 \times 10^5$ galaxies and selection rules based on synthetic colors given by Fukugita *et al.* (1995), and found that the galaxy correlation function is strongly dependent on color, with red galaxies more strongly clustered than

[1] In the stellar classification scheme OBAFGKM, O is the bluest, M the reddest type.

blue galaxies by a factor of $\gtrsim 5$ at small scales.

Since SDSS is generating accurate photometry for an unprecedentedly large and uniform sample of galaxies, we can expand galaxy color studies and extend them in new directions. The SDSS galaxy data have been previously used in a number of studies. For example, Blanton *et al.* (2001) analyzed 11,275 galaxies with redshifts and photometry to calculate the galaxy luminosity function (LF) and its dependence on galaxy properties such as surface brightness, intrinsic color, and morphology, Yasuda *et al.* (2001) derived the galaxy number counts. One of the scientific goals of the SDSS is to study the dependence of galaxy properties such as the luminosity function, size distribution, evolution, and large scale distribution on morphological type. While morphological types can be assigned with some certainty to nearby, well resolved galaxies, this is not possible for the fainter and more distant galaxies imaged by SDSS. If we can find a relationship between observed color and morphological type, this relation (if fairly independent of k-corrections[2]) can in principle replace morphological segregation in studies of galaxy properties and distribution all the way to the limit at which we can do reliable star–galaxy separation. In this analysis we study the color distribution of a large, uniform sample of galaxies. We show that the $u^* - r^*$ color distribution is bimodal, with a clear separation between the two classes down to the SDSS imaging faint limit[3].

2.2 Data

In the analysis of the galaxy colors we use the model magnitudes as measured by the SDSS photometric pipeline *Photo* (version 5.2, Lupton *et al.* 2001). The model magnitudes are calculated by fitting de Vaucouleurs and exponential models, convolved with the PSF, to the two-dimensional images of galaxies in the r band, and computing the total magnitude corresponding to the better fit. This r band fit is applied in all five bands, yielding galaxy colors measured through the same aperture. The uncertainty in the photometric calibration for this data is of the order of 0.3^m. The photometric errors are less than 0.1^m for $u^* \lesssim 19$, $g^* \lesssim 22$, $r^* \lesssim 21$, $i^* \lesssim 21$, and $z^* \lesssim 18$. We correct the data for Galactic extinction determined from the maps given by Schlegel *et al.*(1998).

Our *photometric* sample is a magnitude limited ($g^* \leqslant 21$) sample of 147,920 galaxies

[2]These are the corrections one needs to add to the observed (redshifted) multi-color magnitudes in order to transform them into rest frame.

[3]We refer to the measured magnitudes in this work as u^*, g^*, r^*, i^*, and z^* because the absolute calibration of the SDSS photometric system was still uncertain at the time of the study at the $\sim 0.03^m$ level.

over 101.4 deg^2 of SDSS imaging data. This sample is used to study the distribution of galaxies in SDSS color space.

The *spectroscopic* sample, used for the detailed comparison of color and spectral classification, contains 500 galaxies from a single SDSS spectroscopic plate. Of the 500 galaxies, 443 have $r^*_{\rm petro} < 17.8$ and are part of the main galaxy spectroscopic sample (for more details on the galaxy spectroscopy target selection, see Strauss *et al.* (2002). An additional 57 galaxies are part of the Luminous Red Galaxies (LRGs) sample, comprised of fainter, $0.25 < z < 0.50$, ellipticals.

The *morphological* sample is a subsample of 287 bright galaxies ($g^* < 16$) from the photometric sample, which were classified by eye, and which allows studies of the correlation between color and visual morphology.

2.3 Bimodality of the color distribution

The color–magnitude and color–color diagrams of galaxies in the photometric sample are presented in Figure 2.1. The left panel displays the $g^* - r^*$ vs. $u^* - g^*$ color–color diagram. The distribution of galaxies from the photometric sample is shown as contours. A sample of stars with $15 < u^* < 18$ extracted from the same area of the sky as the galaxy photometric sample is plotted as dots for comparison. The u^* magnitude limits on the stellar sample were selected to ensure a high signal-to-noise ratio. The galaxy distribution has two peaks, with the line connecting them almost perpendicular to lines of constant $u^* - r^*$. This suggests that the $u^* - r^*$ color is nearly optimal for separating galaxies into the two color types.

The g^* vs. $u^* - r^*$ color–magnitude diagram for the photometric sample of galaxies is presented in the right panel of Figure 2.1. We will refer to the subsample of galaxies on the left of the green short-dashed line as "blue" and the one on the right as "red". When plotted as a histogram, the $u^* - r^*$ color distribution has two maxima separated by a well-defined minimum. The positions of these three extrema are only weakly dependent on the sample magnitude cut. We quantify the dependence of the three extrema on the sample magnitude limit by binning the photometric sample in g^* and fitting a sum of skewed "Lorentzian" ($\propto [1 + (x - x_o)^2/(ax + b)^2]^{-1}$) to the $u^* - r^*$ color distribution of the resulting subsamples. The blue and red peaks are then given by the maxima of the "Lorentzians". We define the separator between the two as the point at which the two "Lorentzians" (with areas individually normalized to unity) have the same value. Note that this is not equivalent to finding the minimum between the blue and red peaks, since

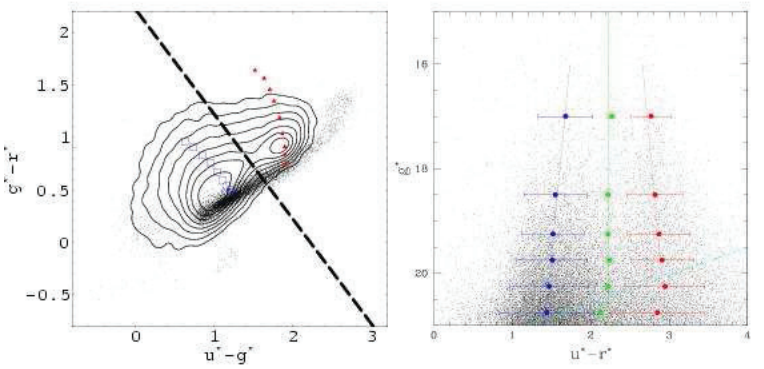

Figure 2.1: Left panel: Distribution of galaxies (contours) and stars (dots) in the $u^* - g^*$ vs. $g^* - r^*$ color–color diagram. The contours enclose $\sigma/4$ (20.8%) to 2σ (95.5%) of all galaxies, in steps of $\sigma/4$ (σ corresponds to the equivalent Gaussian distribution). The thick long-dashed line is the $u^* - r^* = 2.22$ separator. The evolution of spiral (blue open squares) and elliptical (red filled triangles) theoretical colors are given for $0 < z < 0.4$ at every 0.05 in redshift. Right panel: $u^* - r^*$ vs. g^* color–magnitude diagram of the photometric sample. Solid circles show positions of the red and blue peaks and the separator at each mean g^* of six subsamples (see text). Thick lines give linear regressions to each peaks' variation, while the green short-dashed vertical line is $u^* - r^* = 2.22$ separator. The slanted cyan long-dashed line is a $u^* = 22$ cut.

the ratio of red to blue galaxies is a function of sample magnitude cut, and decreases for fainter limiting magnitudes. Sample selection using g band limiting magnitudes or redder bands guarantees a sizeable fraction of red galaxies in fainter samples. The $u^* = 22$ cut, represented by the slanted long-dashed line in the right panel of Figure 2.1, strongly decreases the number of red relative to blue galaxies at fainter magnitudes.

The fitted positions of the three extrema for six magnitude subsamples ($16 < g^* < 21$) are plotted as filled circles in the right panel of Figure 2.1 for each mean value of the g^* bin. The error bars indicate the FWHMs of the best-fit "Lorentzians". The straight lines are the lines fitted through the red and blue peak points. The variation in the position of the $u^* - r^*$ separator is very small (0.1^m) over the g^* range, showing that this criterion for separating blue and red galaxies is valid over a large range of magnitudes. The green vertical line through the separator points is $u^* - r^* = 2.22$, corresponding to the mean and median of the six fits.

Two trends in the $u^* - r^*$ color distribution with fainter g^* magnitudes are visible in the right panel of Figure 2.1: 1. a shift of the blue peak towards bluer and of the red peak towards redder $u^* - r^*$ colors, quantified by the line fits given above, and 2. an overall increase of the density of blue galaxies relative to the red. The variation of the color distribution of the blue and red subsamples with apparent magnitude is due to the fact that we sample galaxies at increasing redshifts when selecting fainter magnitude cuts. The variations are caused by the color variation with increasing redshift (K-corrections and galaxy evolution) and the dependence of the sample on the galaxy number counts and luminosity function. Qualitatively, based on K-corrections alone (Fukugita *et al.* 1995), we expect elliptical galaxies to get redder with increasing redshift ($z \lesssim 0.3$) which is consistent with the observed slope of the $(u^* - r^*)_{\rm red}$ line in the right panel of Figure 2.1. Based on K-corrections only, we would expect the blue galaxies to get initially redder in $u^* - r^*$ by a few tenths of a magnitude and then stay fairly constant with increasing redshift for $z \lesssim 0.5$ (Here we already anticipate the identification of the blue peak with spiral stellar population and the red peak with elliptical; see bellow). If we include galaxy evolution through stellar population synthesis, we expect the $u^* - r^*$ color of blue galaxies to stay almost constant up to z ∼0.4. The color–color diagram of Figure 2.1 presents the expected $u^* - g^*$ vs. $g^* - r^*$ color evolution for a late spiral (blue open squares) and an elliptical galaxy (red filled triangles). The symbols are plotted in 0.05 redshift intervals from z = 0 (redder $u^* - g^*$ and bluer $g^* - r^*$) to z = 0.4 (bluer $u^* - g^*$ and redder $g^* - r^*$). Beyond z ∼0.4, galaxies in the SDSS photometric system evolve roughly perpendicular to the $u^* - r^*$= constant cut in a $u^* - g^*$ vs. $g^* - r^*$ color–color diagram, and the $u^* - r^*$ color is not a good separator. These evolutionary tracks were computed using the evolutionary synthesis model PEGASE (Fioc & Rocca-Volmerange 1997). The evolutionary prescriptions (star formation rate, initial mass function and metallicity) were selected to reproduce the spectra of nearby galaxies at z = 0. The color evolution with redshift is almost parallel to $u^* - r^*$= 2.22 (thick dashed line), keeping the separator fairly constant with redshift up to z ∼0.4. However, the observed $u^* - r^*$ color of the blue peak gets bluer with fainter magnitudes (see the right hand side of Figure 1), departing from the expected evolutionary behavior of a late spiral galaxy. This departure is most likely due to an increasing degree of star forming activity (for which the $u^* - r^*$ color is a sensitive index) with redshift, *i.e.*, a larger number of galaxies with higher star formation activity than the modeled galaxy presented in the Figure are seen at high redshift than at low redshift.

The second trend of the $u^* - r^*$ color distribution with magnitude, namely, the increase

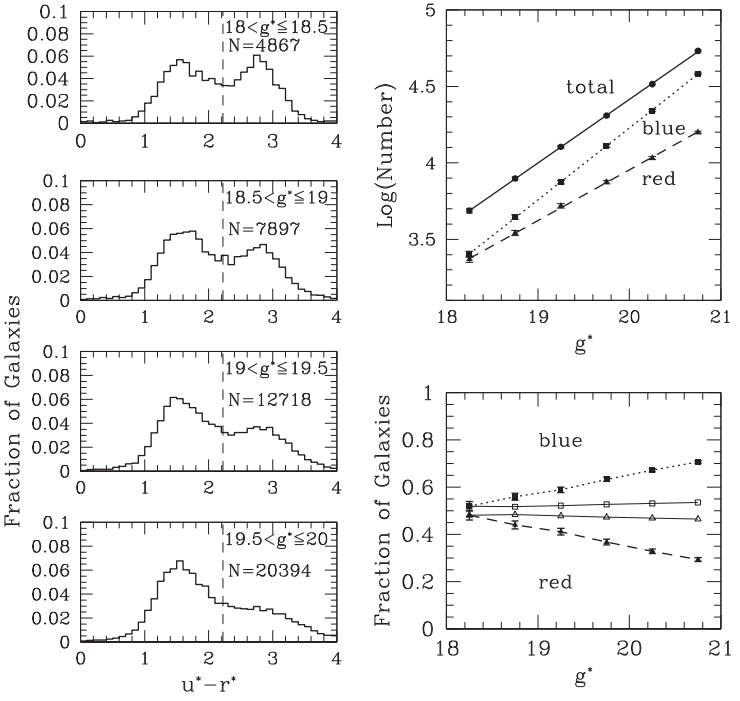

Figure 2.2: Left panel: $u^* - r^*$ color distribution as a function of g^* magnitude of the galaxy sample. Top right panel: The fraction of blue galaxies (filled squares) increases relative to the red (filled triangles) for fainter g^* samples. Bottom right panel: photometric errors cannot account for the dependence of the red and blue galaxy fractions on magnitude cut. The open symbols correspond to the predicted fraction (assuming only photometric errors change with magnitude), the filled symbols to the observed.

of the fraction of blue galaxies, is displayed in the four color histograms for different g^* magnitude bins in the left panel of Figure 2.2. For $18 < g^* < 21$, the slopes of the blue and red galaxy number counts in the top right panel are 0.47 and 0.33 respectively, amounting to a factor of $\sim$2 increase in the number ratio of blue to red galaxies. Since k-correction acts in the opposite direction (*i.e.*, to increase the number of red galaxies), it cannot be responsible for the observed increase of blue galaxies. Moreover, this trend is present if we bin the data in r^*, i^*, or z^* bands, which sample the redder stellar populations. Two instrumental effects could cause the increase of the blue fraction as measured in apparent g^* magnitude bins: the galaxy color distribution could get wider for fainter red galaxies due to increased photometric errors in the u^* band (which would affect red galaxies more strongly, since they are fainter in u^*), or alternatively red galaxies could be artificially "leaking" towards blue $u^* - r^*$ colors, because the asinh magnitudes (Lupton et al. 1999) used by SDSS cannot get fainter than a limit determined by the sky brightness in that band. We investigated both possibilities by assuming no evolution and no k-correction, *i.e.*, the intrinsic distribution of galaxies in faint magnitude samples is similar to the observed at the bright end ($18.0 < g^* < 18.5$), and allowing the galaxies to spread as random Gaussian deviates in the flux with fainter apparent magnitude. We simulated the changes in the $u^* - r^*$ color distribution with g^* magnitude due to those two effects. The lower right panel of Figure 2.2 compares the simulated and observed fractions of red and blue galaxies, demonstrating that those effects are not sufficient to explain the change in the observed distribution. We found that the increased photometric errors cannot explain the large difference observed between the two peaks. The $u^*(0) = 24.63$ sky limit has a negligible effect for the galaxies in our photometric sample, since even the reddest are more than a magnitude brighter than this limit. Increasing the photometric errors by 50% and/or lowering the value of the sky limit to $u^*(0) = 24.00$ do not help in reconstructing the observed change in the color distribution with apparent magnitude, suggesting that we are indeed seeing an increased number of blue galaxies with fainter apparent magnitude. We conclude that the observed evolution of the $u^* - r^*$ color distribution with apparent magnitude is a real evolutionary effect, caused by the dependence of the red and blue galaxy luminosity functions or comoving volume number density on redshift.

The two panels in Figure 2.3 compare the $r^* - i^*$ and $i^* - z^*$ color distributions for the two $u^* - r^*$ color-selected subsamples with $g^* < 21$. The histograms represent the data distributions and the curves show Gaussian fits. It is evident that these two colors are quite similar for "blue" and "red" galaxies, with significant overlap between the two subsamples. The peak separation in both colors is only about 0.1^m - 0.15^m, much smaller than the

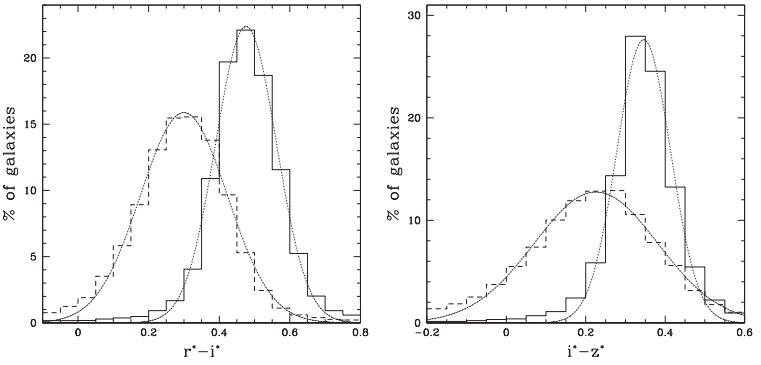

Figure 2.3: $r^* - i^*$ and $i^* - z^*$ color distributions for two subsamples separated by their $u^* - r^*$ color (see text). The smooth curves are Gaussian fits to the data.

wide $u^* - r^*$ color peak separation ($\gtrsim 1^m$). Since the r^*, i^*, and z^* fluxes of galaxies are dominated by the old/low-mass stellar populations present in all morphological types, it is indeed expected that the $r^* - i^*$ and $i^* - z^*$ colors will not show much difference for early and late type galaxies.

However, the fact that the $r^* - i^*$ and $i^* - z^*$ colors are not identical for the two $u^* - r^*$ color selected galaxy types suggests that it may be possible to use this additional information to perform a better galaxy classification based on all four SDSS colors. We used the program AutoClass (Goebel *et al.* 1989; Cheeseman & Stutz 1996) for an unsupervised search for structure in the galaxy color distribution. AutoClass employs Bayesian probability analysis to automatically separate a given database into classes, and is an efficient tool for analyzing multidimensional color diagrams. For example, Ivezić & Elitzur (2000) used AutoClass to demonstrate that the sources from the IRAS Point Source Catalog belong to four distinct classes that occupy separate regions in the four-dimensional space spanned by IRAS fluxes. We searched for self-similar classes in the galaxy color distribution by using a random subset of 25,000 galaxies from the photometric sample. While the algorithm proposed 4 distinct groups, most of the galaxies (82%) are included in only two classes. One of the remaining two classes represents outliers (5%), and the fourth one shows considerable overlap with one of the first two classes. We conclude that the bimodality is an excellent description of the galaxy distribution in the SDSS color space even when all four colors are used. The boundary between the two galaxy types

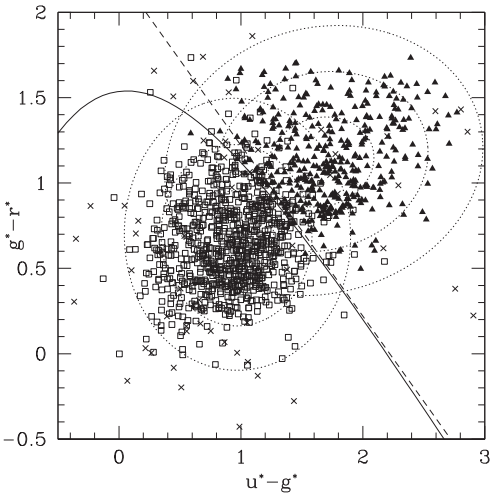

Figure 2.4: Comparison of the Bayesian cut and the $u^* - r^*$ cut in a projection of the four-dimensional color space. Open squares and solid triangles represent the two main classes found by the clustering algorithm (sparse sample). The crosses belong to the outlier class. The thin dotted ellipses are the 1, 2 and 3σ contours of the projected class probability ellipsoids. The diagonal dashed line is the $u^* - r^* = 2.22$ plane which closely follows the Bayesian separator (solid line).

inferred from the $u^* - r^*$ distribution diagram is strongly supported by the AutoClass results. Figure 2.4 compares the Bayesian cut (solid line) and the $u^* - r^*$ cut (dashed line). The close agreement between the unsupervised classification and the simple $u^* - r^*$ color cut is evident.

2.4 Colors and morphology

The data presented in the previous section indicate that the $u^* - r^*$ color distribution is bimodal, and that galaxies can be divided into "blue" and "red" subsamples, as expected based on the differences in the dominant stellar populations for different morphological galaxy types (*e.g.*, de Vaucouleurs 1961). We used independent morphological classification schemes to show that the blue galaxies are indeed dominated by late types (spirals)

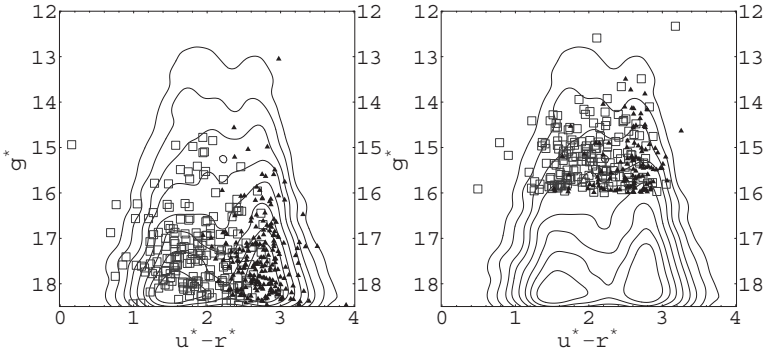

Figure 2.5: Bimodality in the photometric galaxy sample (contours) corresponds to early (filled triangles) and late (open squares) types of galaxies. The 500 galaxies in the left panel are classified spectroscopically, the 287 bright galaxies on the right are classified by visual inspection of images.

while the red galaxies are dominated by early types (ellipticals). This is achieved by classifying a subsample of 287 galaxies using visual appearance at the bright end ($g^* < 16$) and a fainter (${r^*}_{\rm petro} < 17.7$) subsample of 500 galaxies using spectra, and comparing the results to the separation based on $u^* - r^*$ color.

The 500 galaxies in the spectroscopic sample were classified by visually comparing their spectra with templates from Kennicutt's spectrophotometric atlas (Kennicutt 1992). The galaxies are separated into six types: E(0), S0(1), Sa(2), Sb(3), Sc(4), and Irr(5).

The correspondence between spectral classification and $u^* - r^*$ color is shown in the left panel of Figure 2.5. For simplicity, the galaxies are grouped into "early" types (E, S0, Sa), shown as filled triangles, and "late" types (Sb, Sc, and Irr), shown as open squares. Practically all (97.6%) galaxies spectroscopically classified as early types have $u^* - r^* \geqslant 2.22$, and the remaining 3.1% are early spirals (Sa) with $u^* - r^* > 2.05$. That is, we find no examples of spectroscopically classified early type galaxies bluer than $u^* - r^* = 2.05$. The galaxies classified spectroscopically as late type show more scatter in their colors. While 153 of the 210 late type galaxies (73%) have $u^* - r^* < 2.22$, and over 90% (190/210) have $u^* - r^*$ color bluer than 2.5, there are still a small number (20/210) of the spectroscopically late galaxies with $2.5 < u^* - r^* < 3.0$ and presumably low star formation rate and/or internal reddening (the images for the majority of these are consistent with

either face–on spirals or ellipticals). We note that the 3 arcsec diameter spectroscopic fiber samples only the central regions of most of the galaxies in this sample. Because of this aperture bias the spectroscopic classification tends to estimates earlier types than the visual class. The estimated miss-classification due to this effect is about 6%.

The visual classification is based on the optical appearance in the g^* band, which is the closest of the SDSS bands to the standard B band used for classification. 287 bright galaxies ($g^* < 16$) from the morphological sample were classified into the same six types as in the spectroscopic sample: E(0), S0(1), Sa(2), Sb(3), Sc(4), and Irr(5).

The right panel of Figure 2.5 presents the visually classified galaxies separated into early and late types as points over the photometric sample given in contours. Of the 117 galaxies visually classified as E, S0 or Sa, 80% have $u^* - r^*$ colors redder than 2.22, consistent with being early type galaxies. Of the 170 galaxies visually classified as Sb, Sc or Irr, 112 (66%) have colors bluer than $u^* - r^* = 2.22$ consistent with being late type galaxies. Another 16% have colors bluer than $u^* - r^* = 2.5$, while the remaining 18% (30/170 galaxies) have colors in the range $2.5 < u^* - r^* < 3.2$ (the reddest spiral in this sample is NGC 4666, a dusty SABc LINER galaxy, in the upper right corner of the right panel of Figure 2.5, with a B−V color of 0.8 from RC2).

The SDSS photometric pipeline calculates a number of global morphological parameters for every object. These include the likelihoods of the best fit exponential or de Vaucouleurs' profiles, and the concentration index. Profile probabilities and concentration index both correlate well with morphology.

Using the exponential ($P_{\rm exp}$) and de Vaucouleurs' ($P_{\rm deV}$) profile likelihoods, early type galaxies (E, S0, Sa) can be selected by requiring $P_{\rm deV} > P_{\rm exp}$, and late types (Sb, Sc, Irr) by $P_{\rm exp} > P_{\rm deV}$, where both likelihoods are calculated in the r band.

The concentration index, defined as the ratio of the radii containing 90% and 50% of the Petrosian r galaxy light, $C \equiv r_{\rm p90}/r_{\rm p50}$, also correlates with galaxy type. Centrally concentrated ellipticals are expected to have larger concentration indices than spirals. For a classical de Vaucouleurs profile, $I_{\rm E}(r) = I_e e^{-7.67((r/r_e)^{1/4}-1)}$ the concentration index is ∼5.5, while the exponential disks of spirals ($I_{\rm S}(r) = I_s e^{-r/r_s}$) have concentration index ∼2.3. Both estimates correspond to the seeing–free case; the observed values are somewhat lower. The dependence of concentration index on morphological type found in our spectroscopic and morphological samples is weak, with large scatter in C for each morphological type. This indicates that the concentration index is not a robust morphological separator, except in a very crude sense: it can be used to separate early (E, S0, Sa) from late type (Sb, Sc, Irr) galaxies. We adopt a $C = 2.6$ separator, which is optimized

for completeness of subsamples selected from both the spectroscopic and morphological samples and gives equally reliable types in both.

The performance of cuts in profile likelihood and concentration index have been analyzed, we do not go into details here. Overall, the concentration index criterion is can be used at the very bright end ($g^* < 16$), where galaxies are also easy to classify both visually and spectroscopically. The profile likelihood criterion is currently applicable to galaxies at intermediate magnitudes ($16 < g^* < 18$) with a somewhat low completeness for late type selection and low reliability for early type selection. In this range the concentration index and color criteria both give better results. The color criterion is applicable for all magnitude ranges considered, and we specifically recommend its use for fainter samples ($g^* > 18$).

2.5 Conclusions

This study indicates that galaxies have a bimodal $u^* - r^*$ color distribution corresponding to early (E, S0, Sa) and late (Sb, Sc, Irr) morphological types, that can be clearly separated by a $u^* - r^*$ color cut of 2.22, independent of magnitude. The peak-to-peak width of this $u^* - r^*$ color separation is much larger than the separations between the two populations other color combinations ($r^* - i^*$, $i^* - z^*$, $g^* - r^*$, and $u^* - g^*$). The SDSS $u - r$ color is a unique combination of an ultraviolet u band bluer than the Johnson-Morgan U band and a broader and redder r band compared to Johnson-Morgan V. This makes possible a more sensitive comparison of the blue and red portions of the galaxy spectra, relevant for isolating star formation rates. The cut is stable to evolution effects up to z $\sim$0.4, which shows that the separation is applicable to all redshifts relevant for the main sample of SDSS galaxies.

Among the SDSS parameters calculated automatically by the photometric pipeline *Photo*, concentration indices and profile likelihoods are also sensitive to morphology and correlate with $u^* - r^*$ color. Unlike the concentration index and profile likelihood methods, the $u^* - r^*$ color separation does not require well resolved images.

Since star formation rate (SFR) is one of the defining characteristics of the Hubble sequence, it is expected that $u^* - r^*$ color, as an indicator of recent star formation, correlates with Hubble type. Another defining parameter is the mass distribution, which is reflected in bluge-to-disk ratio or concentration index, profile likelihood or alternative parameters of the profile fit (see e.g. Sérsic index n in Section 3.2.2). As these quantities are less subjective than the Hubble classification and can be determined in an automated

way, they are expected to more beneficial to the statistical studies of galaxy properties and formation.

Chapter 3

Correlation between color and asymmetry in the red sequence galaxies

3.1 Introduction

It has become clear in the last decade that the distribution of galaxies is bimodal in color, and that this bimodality correlates with the structure (or morphological type) of galaxies (Strateva *et al.* 2001; Kauffmann *et al.* 2003a; Blanton *et al.* 2003), see Chapter 2. Early-type galaxies are dominated by a spheroid of stars supported largely by random motions, and tend to have red colors and very low present-day star formation rates. They populate a tight color–magnitude relation (CMR) (Sandage & Visvanathan 1978; Bower *et al.* 1992; Ruhland *et al.* 2008); the tightness of this color–magnitude relation is an indicator of their rather ancient stellar populations (Trager *et al.* 2000; Gallazzi *et al.* 2006). In contrast, late-type galaxies are dominated by a rotationally-supported disk of stars, and tend to have significant ongoing star formation at the present day, and populate a reasonably broad (at optical wavelengths) blue cloud whose colors reflect a combination of recent star formation and dust (Schiminovich *et al.* 2007). The stellar mass functions for the two classes of galaxy are different; early-type galaxies can extend to much higher stellar mass than late-type galaxies (Bell *et al.* 2003). This bimodality persists at all redshifts up to $z \sim 1$ (Bell *et al.* 2004a,b; Faber *et al.* 2007).

One of the most remarkable features of the bimodality in the galaxy population is the observed correspondence between galaxy structure and stellar populations. Models of galaxy formation in a cosmological context predict that the structure of a galaxy reflects

the dynamical assembly (merger history) of a galaxy (Kauffmann, White & Guideroni 1993; Cole *et al.* 2000), whereas the cooling of gas and subsequent star formation is a property of the warm/hot reservoir of gas in a halo, which is relatively unaffected by galaxy interactions (Cole *et al.* 2000; Cattaneo *et al.* 2006). Galaxy formation models with 'standard' ingredients — the formation and evolution of dark matter halos, gas cooling, star formation and stellar feedback — predict that almost all galaxies should be forming stars at an appreciable rate (Cole *et al.* 2000; Croton *et al.* 2006; Cattaneo *et al.* 2006). This is in direct disagreement with observations of a red sequence, and is usually interpreted (for good reasons) as indicating not just a minor failing of (admittedly uncertain) prescriptions for these physical processes; rather, this deficiency is interpreted as the signature of a completely separate physical mechanism.

In this regard, the empirical correlation between a lack of star formation and a prominent stellar spheroid (bulge) is an important clue to the mechanism suppressing star formation (Kauffmann *et al.* 2006). A bulge appears to be a necessary (but not sufficient) requirement for the shut-down of star formation (Bell 2008). Taken together with the bulge mass–black hole mass correlation (Magorrian et al. 1998; Häring & Rix 2004), it is not unreasonable to postulate that feedback from accretion onto a supermassive black hole (AGN feedback) is an important mechanism by which galaxies quench their star formation (Croton *et al.* 2006; Bower *et al.* 2006; Cattaneo *et al.* 2006; Hopkins *et al.* 2008a,b; Somerville et al. 2008); although other physical mechansims may also play a role (see, e.g., Naab *et al.* (2007), Dekel & Birnboim (2008), Khochfar & Ostriker (2008), Dekel & Birnboim (2006), Birnboim *et al.* (2007) or Guo & Oh (2008) on gravitational heating, the influence of the development of virial shocks, and the heating of large halos with cosmic ray energy).

Galaxy merging plays a decisive role in this picture. Toomre & Toomre (1972) and Barnes & Hernquist (1992) argued that a natural formation route for early-type galaxies (in particular elliptical galaxies) was through the merger of two pre-existing galaxies. In the intervening time, an impressive variety of evidence has been amassed showing that the properties of early-type galaxies, at least at the broad level, are consistent with those expected for merger remnants: e.g., the widespread existence of kinematically-decoupled cores (Emsellem *et al.* 2007), the distribution of isophotal shapes (Naab & Burkert 2003; Naab, Khochfar, & Burkert 2006), and low-level tidal debris around many ellipticals (Malin & Carter 1983; Schweizer & Seitzer 1992; van Dokkum 2005) are dynamical signatures of mergers and accretion events being an important part of their formation process. Observations of merger remnants support this picture also, having

surface brightness profiles, velocity dispersions and sizes in the near-infrared that are similar to intermediate-luminosity early-type galaxies (Dasyra et al. 2006).

It is argued that shocks and non-circular motions in a galaxy merger cause the gas to lose angular momentum, leading to the growth of the black hole mass and conditions suitable for efficient coupling of energy from an AGN to the diffuse gas in a galaxy (Sanders & Mirabel 1996; Hopkins *et al.* 2008a,b). Circumstantial support for this picture of merger-driven AGN activity and feedback is seen in the tendency for the AGN hosts to have intermediate colors (between blue cloud and red sequence; (Schawinski *et al.* 2007)) and rapidly-outflowing gas ($\gtrsim 1000\,\mathrm{km\,s^{-1}}$) in post-starburst galaxies — such velocities are not expected to result from starburst-driven winds and were interpreted as being relics of a quasar-driven wind (Tremonti *et al.* 2007). In such a picture, the suppression (or quenching) of star formation is reasonably rapid and follows the merger event (Kauffmann & Haehnelt 2000; Di Matteo *et al.* 2005).

Another aspect that has come into focus in the last years is that the formation of red sequence (and early-type) galaxies is an ongoing process; in particular, more than half of the early-type galaxy population has come into place since $z \sim 1$ (Bell *et al.* 2004a,b; Faber *et al.* 2007; Brown *et al.* 2007). Thus, one expects that the mergers (that create the early-type galaxies) and the physical processes that lead to the truncation of star formation (that make those early-types red) continue to happen at relatively recent times. In this case, these physical processes, and their late-time signatures, should be observable at the present day.

In this context, the work of Schweizer & Seitzer (1992) is of much importance. Motivated by the merger hypothesis for the origin of early-type galaxies, they sought a correlation between fine structure (from a tidal origin) and color offset from the CMR (in Schweizer *et al.* 1990 they did a similar work with a smaller sample with absorption line indices). If gas-rich mergers are the primary way in which early-type galaxies are made, then one expects a correlation between color offset and fine structure: Schweizer & Seitzer (1992) found such a correlation with $\sim 4\sigma$ confidence, albeit with considerable scatter. Schweizer & Seitzer (1992) used stellar population models to explore this result in more detail, finding that the early-type galaxy colors could be reproduced by a variety of post-merger ages from 2 to 6 Gyr, depending on the model parameters used.

Yet, a number of issues are left open by this work. Number statistics was a clear issue (they had a total of 69 galaxies in their sample), and testing with a larger sample is clearly desirable. Their definition of fine structure is motivated by the type of tidal features regularly seen around nearby early-type galaxies (Malin & Carter 1983; Schweizer

& Seitzer 1988; van Dokkum 2005), but has two limitations: firstly, it is impossible to measure their fine structure parameter automatically; secondly, high-contrast tidal features are often the result of *minor* mergers or accretions, not major mergers. Finally, at the time at which Schweizer & Seitzer (1992) was written, it was impossible to disentangle age and metallicity influences on spectral indices or colors; it is important to verify that age is in fact the driving parameter of any correlations seen.

Cognizant of these issues, we have initiated an effort to confirm and expand on the landmark results of Schweizer & Seitzer (1992). We have defined a volume and K-band luminosity limited sample (§3.2) of $0.01 < z < 0.03$ early-type galaxies using the Two Micron All-Sky Survey (2MASS; Skrutskie *et al.* (2006)), the Sloan Digital Sky Survey (SDSS; Abazajian *et al.* (2005)), supplemented with redshifts from the NASA/IPAC Extragalactic Database (NED). We have used asymmetry as an automated and repeatable metric for tidally-induced structure, and have analyzed it in conjunction with SDSS colors, and, crucially, ages and metallicities derived from absorption-line spectroscopy (Gallazzi *et al.* 2005); these parameters are described in §3.3. We present our results in Section 3.4. Finally, bearing in mind the insight gleaned from redshift surveys of red sequence and early-type galaxy evolution, we were in a position to model the spectral properties of early-type galaxies more completely than Schweizer & Seitzer (1992); this discussion is presented in Section 3.5. The casual reader is invited to skip to Section 3.4 directly, focusing in particular on Figs. 3.7 and 3.12.

3.2 Data and fitting

3.2.1 Sample selection

Our aim is to test the relationship between fine structure and color deviation from the color–magnitude relation for early-type galaxies. In order to test this relationship, a representative (but not necessarily complete) sample is required, and at distances close enough to allow us to discern both the global and tidal structure of early-type galaxies. Accordingly, we have chosen galaxies within a relatively thin, and nearby, distance shell $0.01 < z < 0.03$ for study. The bulk of our sample can be drawn from the SDSS, but the relatively bright galaxies in this sample lack SDSS spectra (for a variety of reasons, but partially because their fiber magnitudes were too bright to target as part of the SDSS main survey).

Accordingly, we have taken a hybrid approach for this work. The primary object selection was made from the 2MASS All-Sky Extended Source Catalog. We selected

galaxies brighter than 13.5 mag in K. We retrieved redshifts for these objects either from NED[1] or SDSS, matching the objects by position; NED was used to fill in objects lacking spectroscopy from the SDSS. In order to achieve a good redshift completeness we only included galaxies from regions of the sky that are fully covered by spectroscopic sample of the SDSS DR3. We formulated this criterion in the terms of a minimal object density of the SDSS spectroscopic sample at the given point of the sky (see upper panel of Figure 3.1). However, this selection might exclude underdense regions (if we set the density limit too high) and oversample clustered objects at the edges of the covered area. We found that a cut requiring at least 4 objects / $(0.25\ \mathrm{deg})^2$ is both high enough to achieve a good redshift coverage and still not too high to exclude large areas surveyed by SDSS. The positions of the objects selected by this cut are plotted on the lower panel of Figure 3.1. With this selection, 40% of the $K < 13.5$ objects have NED redshifts, 89% SDSS redshifts and 91% at least one of them.

As described in detail by McIntosh *et al.* (2006), from a comparison to the SDSS main galaxy sample, it turns out that the K-band incompleteness of 2MASS is modest and the only major source of incompleteness are blue low surface brightness disk galaxies. In order to check the completeness of our particular 2MASS selection we follow a similar philosophy to Bell *et al.* (2003) and McIntosh *et al.* (2006), by exploring the 2MASS properties of galaxies selected in the optical regime. We selected a sample of galaxies with known redshifts and SDSS photometry from NED. With the same procedure as described in Bell *et al.* (2003), the SDSS u, g, r, i, z magnitudes and the redshifts were fit with stellar population template spectra derived using the PEGASE stellar population code Fioc & Rocca-Volmerange (see 1997, for an earlier version of the model). These template fits were used to estimate k-corrections for the sample, and permitted estimation of the K-band magnitude of a galaxy from the optical fluxes alone. In this way we could explore whether 2MASS detected all the galaxies which one would *a priori* expect to be detected. Fig. 3.2 shows the color dependence of the completeness for two redshift cuts, $z < 0.1$ and $0.01 < z < 0.03$. The samples are less complete at bluer colors. The integrated incompleteness is approximately 3% and 5% for the two redshift cuts, respectively.

Figure 3.3 shows the completeness of the redshift data coming from the two sources. SDSS is most incomplete at the bright magnitudes, whereas in NED the completeness decreases with K magnitude. This trend dominates the combined catalogue, but the decrease of the completeness is much more moderate. With our selection criteria, at least one of the redshift data (SDSS or NED) is available for 91% of the $K < 13.5$ galaxies,

[1]The NASA/IPAC Extragalactic Database, `http://nedwww.ipac.caltech.edu/`

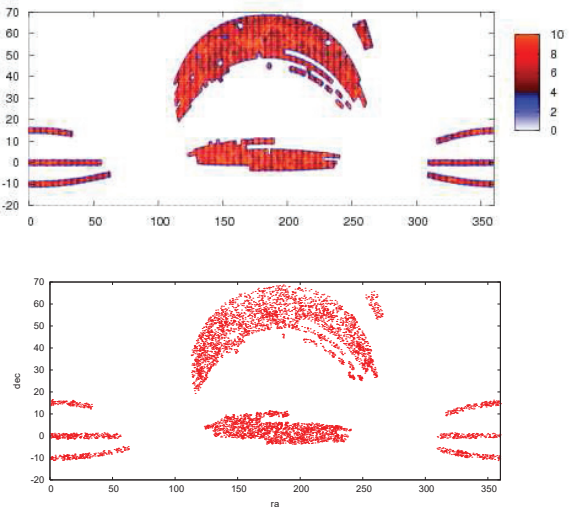

Figure 3.1: Upper panel: Density map of the SDSS DR3 spectroscopic sample, z axis: counts per $(1/4\,deg)^2$. Olny regions with at least 4 objects are included to the sample. Bottom panel: The sample selected by the density mask.

the incompleteness mainly coming from faint, $K > 12$ galaxies.

For the structural studies we select a nearby subsample which is volume limited between $0.01 < z < 0.03$. The corresponding and K-band absolute magnitude limit is $M_K < -22.31$ (which approximately corresponds to $L > L^*/5$, (Bell *et al.* 2003)). This sample contains 3006 galaxies. Its estimated completeness is approximately 90%; this completeness is dominated by the spectroscopic incompleteness (largely from SDSS, and this is primarily a geometric effect owing to the inability to place fibers on all objects) coupled with a modest trend to miss out low surface brightness, blue objects; for red galaxies, the focus of this work, the sample is essentially limited only by SDSS fiber placement.

3.2.2 Fitting

We carried out image fitting with GALFIT to select early-type candidate systems. For fitting the galaxy profiles we constructed an automated pipeline involving programs Sex-

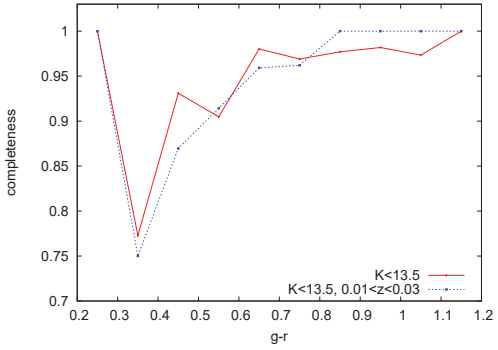

Figure 3.2: The estimated completeness of a 2MASS K-band selected sample, as assessed by comparison with synthesized K-band apparent magnitudes of a deeper optically-selected catalog from the SDSS. Solid red line: $K < 13.5, z < 0.1$, blue dotted line: $K < 13.5, 0.01 < z < 0.03$. The 2MASS K-selected sample is less complete for blue objects.

tractor version 2.3.2 (Bertin & Arnouts 1996) and GALFIT version 2.0.3. (Peng *et al.* 2002). SExtractor creates a catalog of sources from an astronomical image. The output parameters can be used to create an initial setup for light profile fitting with GALFIT.

We pipeline was run on SDSS r-band images. For each image, SExtractor was used to look up all sources and to identify the target galaxy, using SDSS positions as input. Besides our primary galaxy, we fit the nearby objects that could influence the light profile fit of the main galaxy. The selection of these secondary targets was made by distance and intensity: the object was ignored if the distance between the two objects exceeded 1.5 times the sum of their sizes, or the intensity of the secondary object was lower than 1% of the main object. In some cases SExtractor identifies parts of the galaxies, such as large HII regions as separate objects. In order to correct these cases, check images were created with labeled fit targets. The images were visually inspected and unnecessary sources were untargetted. These results were used to create the GALFIT initial setup.

The main target galaxies were fitted with a Sérsic (1968) profile

$$\Sigma_b(r) = \Sigma_e e^{-c(r/r_e)^{1/n}-1} \tag{3.1}$$

where r_e is the effective radius of the galaxy, Σ_e is the surface brightness at this radius, n is the Sérsic index (in the special case when $n = 1$ we get the exponential profile, $n = 4$

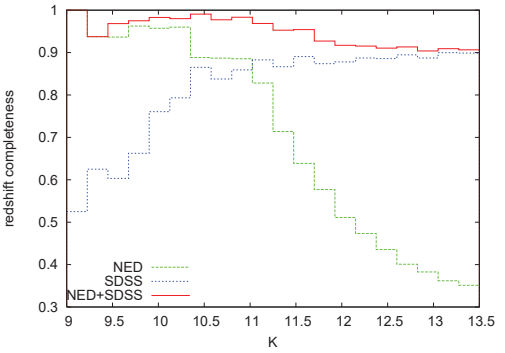

Figure 3.3: K-band magnitude dependence of redshift completeness of the sample. The three histograms show the redshift completeness of the 2MASS-selected sample using data from different sources: green dashed line: NED, blue dotted line: SDSS, solid red line: NED and SDSS together.

corresponds to de the Vaucouleurs profile) and $c = 1.9992n^{-0.3271}$ so that the half of the flux is within r_e. The nearby objects (secondary targets) were fitted either by Sérsic or, if more star-like, by the PSF. The PSF was created from stars of each field image separately, based on Moffat profile fits in the form:

$$\Sigma_b(r) = \frac{\Sigma_0}{[1 + (r/r_d)^2]^m}, \tag{3.2}$$

where r_d is the dispersion radius and m is the power-law index.

Initial values for sky level, flux, size and axis ratio were set using SExtractor output parameters. The Sérsic index and axis ratio for galaxies were set to initial values common for all objects, $n = 1.5$, $b/a = 0.9$. The fit parameters were constrained using empirical (observational) and SExtractor output information: n was required to lie in the interval $0.5 < n < 7$, axis ratio $0.3 < b/a < 1$, position within 2 pixels of the SExtractor value, and half-light radius between 0 and $1.5r_{\rm SExtractor}$. The pipeline first fits the objects separately, keeping the other objects fixed, and then finally fits the objects all together in the end, using their separate fit results as the initial guess. The sky is fixed at all times.

One way to test the fitting pipeline is to fit simulated objects of known parameters. We used a suite of simulated galaxies composed from a Sérsic bulge + exponential disk, with a wide range of magnitudes, bulge to total ratios and light profile parameters. We selected a test sample similar to our sample of interest: bulge dominated systems, having bulge to

total ratios $B/T > 0.6$ and magnitudes brighter than 16.5 (the $K < 13.5$ cut corresponds approximately to $r < 16.4$). The simulated galaxies were fit by the pipeline described above. The fits give us an important sanity check of the fitting procedure. Our results shows that Sérsic fit works relatively well, the simulated magnitudes are reproduced with an error ~ 0.2 mag for the test sample.

When fitting the real images, the Sérsic fit is successful and gives results similar to the SDSS photometric parameters for the majority of our galaxies. It also corrects the known error of the SDSS imaging pipeline that it tends to split up large objects into multiple sources. This affects 4.6% of the objects in our sample, mostly objects with $n < 1$. The Sérsic fit fails in 1.5% of the cases because the fit did not converge or the image file was corrupt. 1.9% of the fit results are unreliable because the fit did not move out from the initial setup or reached the fitting boundary. Figure 3.4 shows the fitted sizes (circularized $r_e^o = r_e * \sqrt{(b/a)}$) compared with the SDSS R_{50} radii. The fit gives results consistent with the SDSS sizes within 15%, when $> 3\sigma$ outliers are discarded.

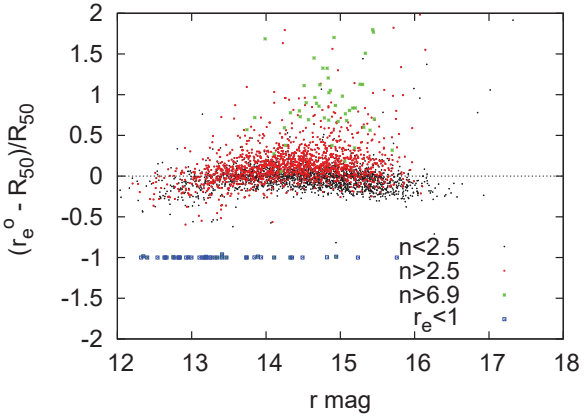

Figure 3.4: The relative difference of the fitted circularized Sérsic r_e vs. SDSS R_{50} as a function of r magnitude. Small symbols: black: $n < 2.5$ (a proxy for disk-dominated; $n = 1.5$ excluded), red: $n > 2.5$ (a proxy for bulge-dominated). Larger symbols: fit hit a limit, green: fit hit the upper limit in n, blue: fit hit the lower limit in r_e.

We choose $n > 2.5$ galaxies for further study, which appears to be a reasonable way to automatically reject a significant fraction of the disk-dominated galaxies (Bell *et al.*

2004b). We further exclude all objects with failed fits, as described above. The criterion $n > 2.5$ is deliberately generous, and includes a large number of galaxies with prominent disks (Bell *et al.* 2004b). Accordingly, we classified the $n > 2.5$ subsample by eye, classifying 553 objects as elliptical (pure bulge; E), 498 as lenticular (smooth, non-star-forming disk; S0), 479 as later types (Sa-Irr) and 27 as mergers.

We choose to visually exclude all galaxies with stars or companion/projected galaxies which made a major contribution to the asymmetry in the area in which asymmetry is measured. The spirit of this cut is to excise all systems where we were concerned that the asymmetry measurement would be corrupted by a non-interacting (or non-strongly-interacting) object (be it a projection or an interaction). While some form of masking might have been desirable, we adopted this conservative approach to account for objects missed by Sextractor and because the extent of the masked area is never straightforward to choose. We lose 1/3 of the sample using this cut (it is our most important cut), and the loss is primarily geometric (and therefore introduces no bias) for the galaxies lost from bright stars and projections, and those galaxies thrown out because of a pre-merger satellite. There is some natural fuzziness in this process, and some advanced mergers where the second nuclei (and large amounts of tidal disturbance) were within the asymmetry measurement area were included in the sample. After this final selection our sample contains 345 E and 347 S0 type galaxies, together 692 objects.

3.3 Asymmetry and stellar population measurements

For the investigation of the correlation between the post-merger structure and and stellar population we need to construct two broad classes of metric: descriptions of the (hopefully tidally-induced) non-equilibrium structure of an early-type galaxy (§3.3.1), and parameters characterizing the stellar population, like color, age, and metallicity (§3.3.2).

3.3.1 Structure

We choose to use asymmetry as a repeatable diagnostic of non-equilibrium structure. This measure is insensitive to symmetric bars and rings, which are important structures in some S0's but are not (at least unique) indicators of a recent perturbation. Our metric is not as tailored as the fine structure parameter used by Schweizer & Seitzer (1992), but the asymmetry measurement is repeatable and can be automatically applied to survey

datasets.

We adopt an asymmetry measure similar to methods discussed by Conselice, Bershady, & Jangren (2000). We measure asymmetry within an elliptical annulus of 0.8–3 r_e. In this way we minimize the impact of the bright galaxy core on our asymmetry measure which would otherwise be very sensitive to bad centering. A different choice of the aperture does not influence the results qualitatively, except for the inner $0.5r_e$. The asymmetry calculating algorithm is carried out as follows in the steps 1-7, for an illustration see Figure 3.5.

1. Rotate the image around the GALFIT center by 180°.

2. Subtract the rotated image from the original one.

3. Integrate flux in quadrature within the elliptical annulus between 0.8-3 r_e. This measure will have an additive contribution from noise which can be separated and corrected for. Let us separate the flux f into signal and noise term, $f = s + n$ and analogously for the rotated image $f' = s' + n'$. Then the asymmetry flux $f - f'$ integrated in quadrature over the image will read

$$\begin{aligned} \sum (f - f')^2 &= \sum (s + n - s' - n')^2 = \\ \sum (s - s')^2 &+ \sum (n - n')^2 = \sum (s - s')^2 + \sum n^2 + \sum n'^2, \end{aligned} \tag{3.3}$$

   as signal and noise, as well as the both noise terms are uncorrelated. This will result in a non-zero asymmetry even in the case of $s = s'$, i.e. perfectly symmetric underlying signal. We can correct for this by estimating the noise contribution and subtracting it from the integral. We estimate $n - n'$ using the GALFIT model image with noise similar to the observed data.

4. Apply Poisson noise + readnoise to the (symmetric) GALFIT model image.

5. Measure asymmetry in the same way as for the actual image (steps 1-3, rotate, subtract, integrate in quadrature).

6. Subtract model asymmetry from the object asymmetry.

7. In the end we make the asymmetry measure independent of the total flux: normalize by the object (sky subtracted image) flux squared and integrated within the same area.

To summarize, our asymmetry measure is defined as

$$A = \left(\sum (f - f')^2 - \sum (m - m')^2\right) / \sum (f^2). \tag{3.4}$$

where f denotes the flux of the galaxy, m is the fitted Sérsic model, the primes denote the corresponding quantities rotated by 180 degrees, and the summation goes over pixels within the elliptical annulus 0.8-3 r_e. Because the asymmetry is determined from the square of the asymmetry image, values of ~ 0.01 are already very highly significant, suggesting asymmetries involving $\sim 5\%$ of the flux (this definition counts asymmetric flux from both the positive and negative parts of the asymmetry image; other definitions divide the asymmetry by two to account for this effect).

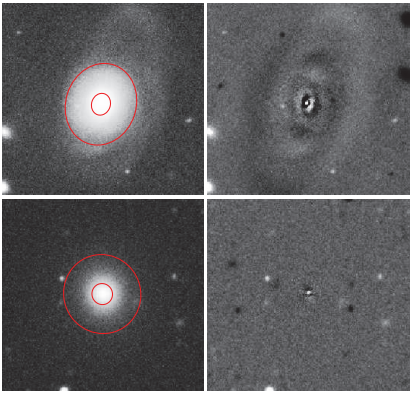

Figure 3.5: Two examples of asymmetry measurement. Left panels: the original image, with the elliptical annulus overplotted; right panels: asymmetry images.

3.3.2 Stellar population diagnostics

We choose for consideration in this work two types of stellar population diagnostics: galaxy colors, as an indicator of broad changes in the stellar population; and luminosity-weighted stellar ages and metallicities — available for roughly 2/3 of the sample — as model-dependent but more information-rich stellar population diagnostics.

We use SDSS model $g - r$ color corrected for foreground extinction and AB offset, then k-corrected using method of Bell *et al.* (2003). Note that these colors are very similar to colors we measured ourselves within $0 < r/r_e < 3$, and the results change only

imperceptibly when such colors are used. We adopt the model colors because these are readily and publicly available. The color combination $g - r$ was used because it is the bluest well-measured SDSS color combination. While, in principle, $u - r$ color is more sensitive to stellar population (as it straddles the 4000Å break), u-band has poor S/N, and suffers from increased systematics compared to g-band.

In order to aid in disentangling the effects of the age/star formation history of a galaxy from metallicity in driving its color, we use estimates of luminosity-weighted age and metallicity from Gallazzi *et al.* (2005). Gallazzi *et al.* (2005) derived ages and metallicities from all SDSS spectra with median S/N$>$ 20 per pixel. This stringent S/N cut selects only $\sim 1/4$ of all galaxies, primarily those with high surface brightness cores. Our sample consists of E/S0 galaxies only, and therefore is much more complete than the whole sample: of the 345 Es and 347 S0s in the sample, 230 Es and 223 S0s have SDSS spectra, from which 224 Es and 215 S0s have age/metallicity estimates (i.e., more than 97% of the Es and S0s in our samples with SDSS spectra have age/metallicity estimates from Gallazzi *et al.* (2005)).

Ages and metallicities are derived by comparing observed spectral absorption features (three primarily age-sensitive indices, 4000Å break strength D4000, Hβ, Hγ_A+ Hδ_A, and two primarily metallicity-sensitive indices [Mg_2Fe] and [MgFe]$'$) with a library of Bruzual & Charlot (2003) models with a wide range of star formation histories and metallicities. The models have solar abundance ratio patterns (i.e., they do not allow for α-element enriched abundance patterns), but the impact of this shortcoming on the derived ages and metallicities is not that severe owing to the choice of combined Magnesium and Iron indices as the metallicity-sensitive indices (see Gallazzi *et al.* (2005) for details and more discussion). In this section, the main emphasis is on the age estimates (recall that the hypothesis being tested is that richly-structured early-type galaxies have younger ages), which depend primarily on the Balmer lines and the strength of the 4000Å-break. These ages are *luminosity-weighted*, i.e., the ages are the mean luminosity-weighted ages of the models that provided a reasonable fit to the five observed spectral indices, and are correspondingly younger than the mass-weighted ages of those model populations. It is worth noting that the luminosity-weighted ages are reasonably model-dependent, and depend on when star formation is assumed to start and to the timing and amplitudes of bursts of star formation. For example, a model with given linestrengths could be fit by a population that started forming stars 12 Gyr ago, or only 5 Gyr ago; this will naturally lead to dramatic differences in luminosity-weighted age. Accordingly, in this work, we use luminosity-weighted ages primarily as a tool to understand *relative* trends in age, and to

attempt to disentangle the effects of age and metallicity on broad-band colors.

3.4 Results

Figure 3.6 shows the color-magnitude relation for E/S0 galaxies. We fitted a line to the main locus of the CMR, excluding blue outliers at $g-r<0.6$ from the fit. The equation of the fitted line is

$$g-r=-0.023\,M_r+0.28. \tag{3.5}$$

We subtract this linear relation from the color to get the color offset from the red sequence: $\Delta(g-r)$. The scatter of $\Delta(g-r)$ is 0.041 with 3σ-clipping applied; much of this scatter is intrinsic (not measurement error, the typical measurement error in model $g-r$ for our sample is ~ 0.015 mag and is dominated by flat fielding and calibration uncertainties; see also Schweizer & Seitzer (1992) and Ruhland *et al.* (2008) who also measure similar intrinsic scatter in the CMR[2]). The mean color offset is slightly negative, $\langle\Delta(g-r)\rangle=-0.005$, because of the clipping of very blue outliers from the CMR fit.

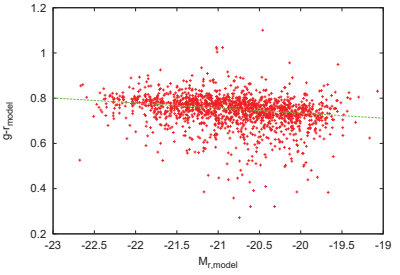

Figure 3.6: The color-magnitude relation in $g-r$ SDSS model magnitude. The green line shows the best fit.

Figure 3.7 shows the color offset $\Delta(g-r)$ as a function of our asymmetry measure, with bin averages plotted with blue dotted lines. The data show a correlation between asymmetry and color offset for both E and S0 galaxies, with approximately the same slope and zero point; this figure can be compared reasonably directly with Figs. 2 and 3 of Schweizer & Seitzer (1992), and confirm their detection of a significant correlation

[2]for the purposes of comparing scatter in $g-r$ with $U-V$, note that $\Delta(U-V)\sim 2.5\Delta(g-r)$

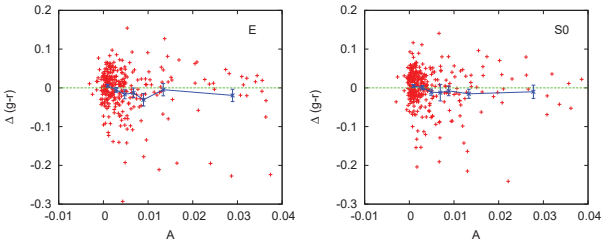

Figure 3.7: Color-asymmetry diagram. Left panel: E's, right panel: S0's. Blue dotted lines show averages in bins, error bars denoting the error of the average. A correlation between color and asymmetry is present in both types.

between substructure in early-type galaxies and an offset from the color–magnitude relation. The correlation is clearest for $A < 0.01$; at larger asymmetry values, in particular for elliptical galaxies, there are quite a few galaxies with significant asymmetry but relatively red colours. To help visualize the trend, we ranked the E/S0 galaxies by A and show how rank A correlates with color in Fig. 3.8. The correlation between $\Delta(g-r)$ and A for Es and S0s combined with $A < 0.04$ is significant at the 99.999% confidence level, as determined using the sample's Pearson correlation coefficient (see Table 3.1 for the full breakdown by galaxy type and stellar population diagnostic).

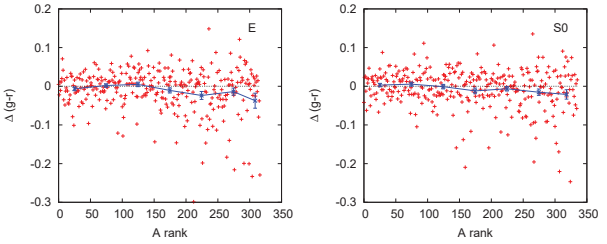

Figure 3.8: Color-asymmetry diagram. Left panel: E's, right panel: S0's. Symbols are same as in Fig. 3.7.

In Figures 3.9, 3.10, and 3.11 we show several examples of different asymmetry values

Sample	N	r	$P(t > \lvert t_{\rm obs}\rvert)$
Color–asymmetry			
E	319	-0.198	0.00018
S0	336	-0.106	0.026
all	655	-0.163	1.4×10^{-5}
Age–asymmetry			
E	208	-0.242	0.00002
S0	210	-0.312	2.0×10^{-6}
all	418	-0.270	9.8×10^{-9}
Metallicity–asymmetry			
E	208	-0.037	0.30
S0	210	-0.011	0.44
all	418	-0.021	0.34

Table 3.1: Probability of observing a given stellar population–asymmetry correlation (for galaxies with asymmetry $A < 0.04$) in a completely uncorrelated dataset, as evaluated using Pearson's correlation coefficient. The probability of a given value of the correlation coefficient r being due to chance alone in an uncorrelated dataset is distributed as a t distribution with $N-2$ degrees of freedom, where $t = r/\sqrt{(1-r^2)/(N-2)}$.

for both morphological types, to allow the reader to generate some intuition for the asymmetries shown by the sample. In Fig. 3.9 shows a sample most symmetric E's and S0's, with asymmetry in the range $0 < A < 0.001$. Fig. 3.10 shows galaxies with $0.005 < A < 0.01$; galaxies in this range are those that drive the bulk of the correlation between color offset and asymmetry. Fig. 3.11 shows highly asymmetric galaxies from the asymmetry bin $0.02 < A < 0.04$. It is clear that the asymmetries being discussed here are not from isolated HII regions, or low-level spiral arms; rather, these asymmetries are to a great extent driven by large-scale asymmetries, many of them matching qualitatively the form expected for tidal features. Taken together with Fig. 3.7, these figures underline the primary result of this work (and that of Schweizer & Seitzer (1992)): early-type galaxies with asymmetric signatures of tidal interactions tend to be offset from the locus of the color–magnitude relation, in the sense that more asymmetric galaxies tend to be bluer than undisturbed galaxies.

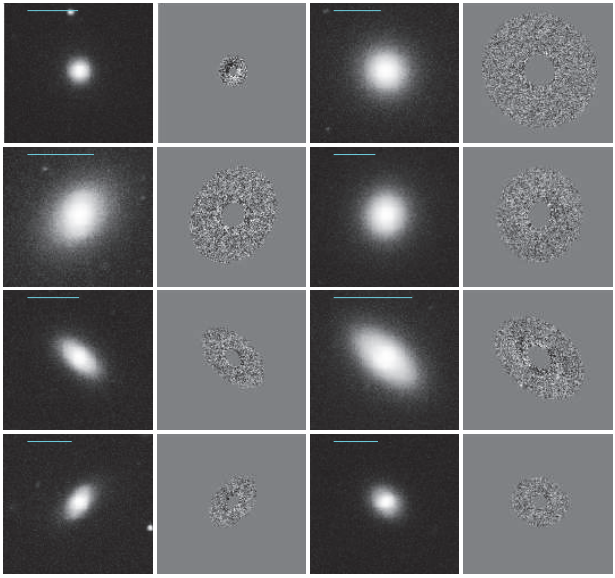

Figure 3.9: Galaxy images (first and third column) with their corresponding asymmetry images) in elliptical annuli (second and fourth column, respectively) for the $0 < A < 0.001$ asymmetry bin (most symmetric objects). First and second row: E's, third and fourth row: S0's. Each box is approximately 55 arcsec on a side. 10 kpc horizontal line plotted to each object image.

3.4.1 What drives these correlations, age or metallicity?

One of the unavoidable limitations of the work of Schweizer & Seitzer (1992) was that it was impossible to explore quantitatively whether it was age or metallicity that was driving the color–fine structure relation. Here, we make use of estimates of luminosity-weighted stellar age and metallicity, available for almost all of the $\sim 2/3$ of our sample with SDSS spectra (Gallazzi *et al.* 2005), to address this question quantitatively.

Figs. 3.12 and 3.13 show the run of luminosity-weighted age and metallicity as a function of asymmetry separately for E and S0 galaxies. There is a strong correlation between age and asymmetry for both E and S0 galaxies (combined E/S0 sample has an essentially zero chance of being due to chance alone, from a Pearson's correlation coefficient), and no such correlation for metallicity (34% probability of arising from chance alone). We

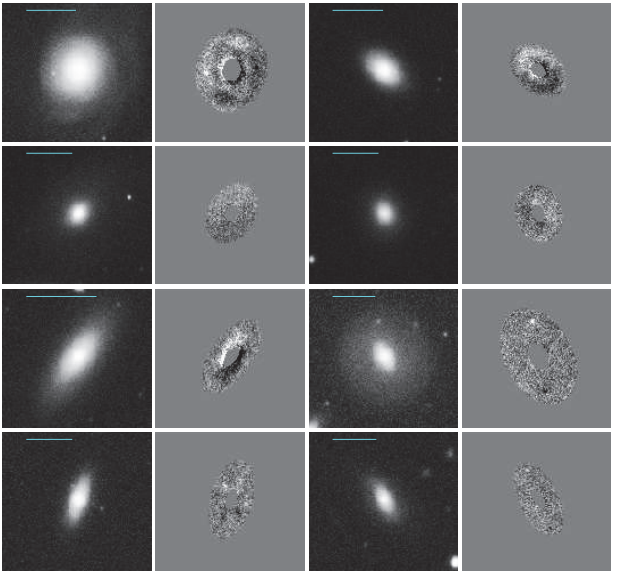

Figure 3.10: Same as Fig. 3.9 but asymmetry range $0.005 < A < 0.01$. These are typical asymmetry values for the galaxies that are driving the color–asymmetry correlation.

can therefore conclude that the color–structure correlation is, to the extent that we can tell, driven entirely by age effects; this *observation* verifies the fundamental *assumption* made by Schweizer & Seitzer (1992). The correlation is actually somewhat clearer in age–asymmetry space, as the contribution from random scatter in the metallicties to the color–asymmetry trend act to scatter out the colour–asymmetry relation.

3.5 Discussion

Motivated by the work of Schweizer & Seitzer (1992), we showed that the color offset from the early-type galaxy color–magnitude relation correlates with (primarily tidally-induced) asymmetry, although the scatter in the correlation is large. We showed, for the first time, that this correlation is driven primarily by age effects. In this section, we will explore further the meaning and limitations of these results. We will present a quantitative model-based discussion of the results, focusing on the bearing that these results have on

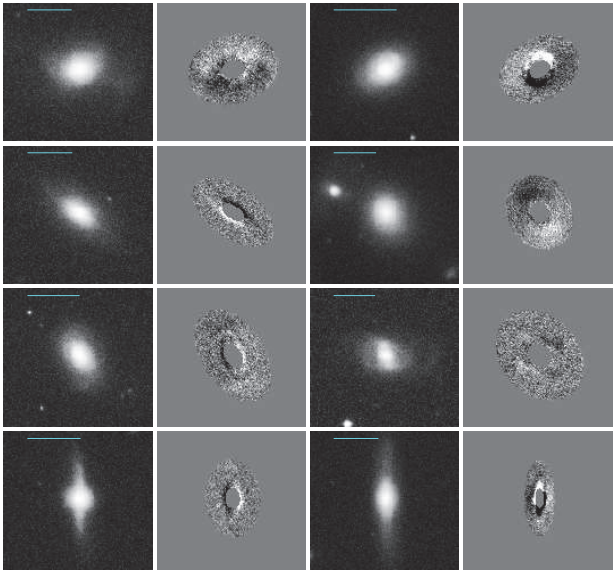

Figure 3.11: Same as Fig. 3.9 but asymmetry range $0.02 < A < 0.04$ – most asymmetric E/S0 galaxies.

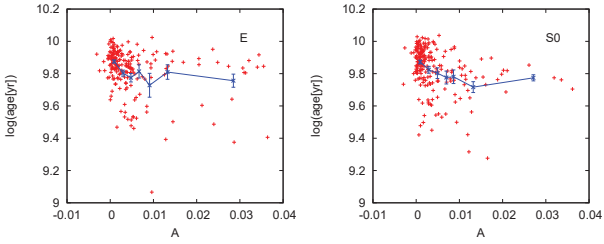

Figure 3.12: The logarithm of the luminosity weighted age (in years) versus asymmetry. Left panel: elliptical galaxies, right panel: lenticular (S0) galaxies. Blue line: average in bins; error bars: error of the mean.

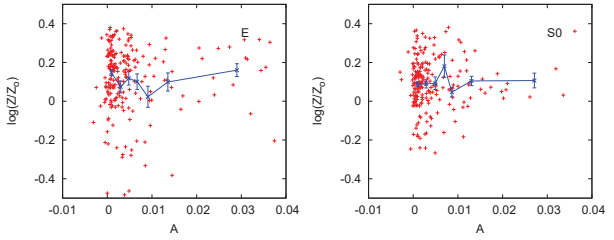

Figure 3.13: Metallicity ($log(Z/Z_{\odot})$) as a function of asymmetry. Left panel: E's, right panel: S0's.

the hypothesis that one of the primary formation routes of early-type galaxies is through galaxy merging. Then, we will then wrap up with a discussion of the limitations of this work, and scope for future improvements.

3.5.1 A quantitative model of red sequence offsets

Like Schweizer & Seitzer (1992), we show a relationship between color offset from the red sequence and asymmetric structure of a galaxy. In Schweizer & Seitzer (1992), it was argued that the 'fine structure' being measured was a strong indication of a previous gas-rich galaxy merger. Thus, in that paper, they presented a model in which early-type galaxies were the product of gas-rich major merging, and they studied the time taken for major merger remnants to age as a way of translating their color offsets into timescales for early-type galaxy evolution. Their model had a number of free parameters: the two key parameters τ_2 and ϵ controlled the duration and prominence of star formation after the merger event. They found that these parameters were of great importance in setting the timescales for fading of the remnant. In models with a large value of τ_2 the remnant reddened only slowly, reaching the red sequence only after many Gyr. In models with short-lived post-merger star formation, they found a rapid evolution onto the red sequence.

One of the motivations of their work was to address the 'gap' between clear remnants from relatively recent mergers ($\lesssim 1$ Gyr ago, e.g., NGC 3921 or NGC 7252) and 'ancient' early-type galaxies. Schweizer & Seitzer (1992) quote King's question to Toomre at the 1977 Yale conference: "You showed us 10 merging pairs and then asked us to look for, or at least accept the existence of, 500 remnants from so long ago that they no longer

bear the 'made by Toomre' label. I would be much more impressed if you showed us the 20 or 30 such systems in the box immediately adjacent in your histogram. What do these merged pairs look like in their next few galactic years?". While Schweizer & Seitzer (1992) argued that the blue, structured early-type galaxies in their study were reasonble candidates to fill the 'King gap', the considerable sensitivity of the evolution of the merger remnant to the properties of the merger and star formation after the merger made it difficult to reach a unique conclusion.

In this subsection, we revisit this issue with the benefit of more advanced stellar population models, and more critically, a more physically-motivated model for the color evolution of early-type galaxies. We will argue that the systems explored in this section (and in Schweizer & Seitzer (1992)) fill the 'King gap', but that owing to natural scatter in the merger history of early-type galaxies and in the stellar populations of these newly-formed ellipticals, it is impossible to connect systems one-to-one with a given merger time; King's question was ill-posed, and will be very difficult to ever unambiguously answer.

Model ingredients

Clearly, there is a huge parameter space that can be explored when modelling the evolving early-type galaxy population, and arriving at a single unique picture that is clearly superior to the others will be impossible (as clearly articulated and demonstrated by Schweizer & Seitzer (1992)).

We take an approach that is motivated by the last decade of research on the evolution of the early-type galaxy population. There are now a number of surveys that have empirically tracked the evolution of the stellar mass density in early-type galaxies, with the general result that the growth in stellar mass density is driven primarily by an increase in the *number* of early-type galaxies (Chen *et al.* 2003; Bell *et al.* 2004a; Brown *et al.* 2007; Faber *et al.* 2007). Star formation is shut off in blue star-forming galaxies, and the remnant fades and reddens onto the red sequence (Bell *et al.* 2007). The exact mechanisms that shut off star formation are still a matter of some debate (merging and AGN feedback, environment, or gas exhaustion are all possibilities); also open is the issue of how important any burst of star formation might be before the truncation of star formation (see Robaina et al., in prep, for a discussion of the average effects of galaxy mergers on star formation rate). Notwithstanding this uncertainty about mechanisms for shutting off star formation, it is not unfair to assume that the timescale for truncation of star formation is short, $\ll 1\,\mathrm{Gyr}$[3]. Evidence indicating the short timescale for star formation truncation include the detection

[3]We are assuming in the terminology of Schweizer & Seitzer (1992) that $\tau_2 \ll 1\,\mathrm{Gyr}$.

of rapid, arguably AGN-driven winds in post-starburst[4] galaxies (Tremonti *et al.* 2007), the pronounced bimodality[5] in the star formation rates of galaxies (Strateva *et al.* 2001; Bell *et al.* 2004a; Schiminovich *et al.* 2007), the high incidence of concentrated, post-starburst AGN galaxies in the 'valley' between the star-forming and non-star-forming galaxies (Schawinski *et al.* 2007), the modest population of blue morphologically early-type galaxies (if extended periods of star formation followed a merger, a large fraction of star-forming early-type galaxies would be observed; such a population is not common; Häußler 2007) and analysis of the spectra of early-type galaxies (Gallazzi *et al.* 2006). In the context of this section (which was motivated originally by the merger hypothesis), we will suppose that merging rearranges the stellar content of the remnant into a more spheroidal configuration and leads to the suppression of future star formation through AGN feedback or some other mechanism.

Thus, we model a constantly-growing early-type galaxy population as a population of galaxies with constant star formation rate, that then stop forming stars at times drawn from the following probability distribution: $P(t) \propto 1 - t(\mathrm{Gyr})/21$, where t is the age of the galaxy in Gyr. For concreteness, we choose to assign $t_0 = 12\,\mathrm{Gyr}$, i.e., the formation of early-type galaxies in this model started 12 Gyr ago and continues at a slowly-reducing rate to the present day. Our default models are truncation-only for simplicity; we show later that bursts of star formation on truncation affect the results very little. We assign an early-type galaxy a metallicity drawn from a distribution with $\sigma_{[\mathrm{Fe/H}]} = 0.1$ and a mean of solar metallicity[6], motivated by the metallicity scatter inferred by (Gallazzi *et al.* 2006). We use the multi-metallicity PEGASE stellar population model (see Fioc & Rocca-Volmerange (1997) for a discussion of the previous version of the code) to model the color evolution of the early-type galaxy population. We neglect, for the sake of simplicity, possible differences in formation history as a function of galaxy mass[7], while noting that the mean mass of our sample is somewhat less than $10^{11} M_\odot$.

Major merger model

The above model incorporates the effect of only one major merger: the one that created the early-type galaxy. The results of such a model are shown in Fig. 3.14; the results of

[4]Galaxies with signatures of no current star formation but substantial star formation <1 Gyr ago.

[5]such bimodality is washed out if transitions from blue to red take any longer than 1 Gyr

[6]The mean metallicity is of no importance in this section, because we are exploring *residuals* from the color–magnitude relation.

[7]We see no significant difference in the behavior of this sample, when splitting into a high-mass and low-mass subsample, justifying this oversimplification.

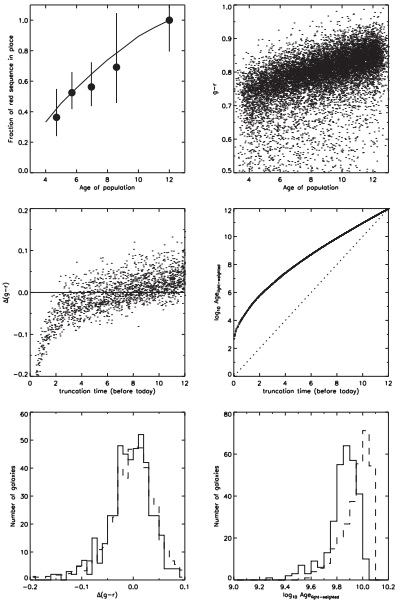

Figure 3.14: A model in which major mergers create the early-type galaxy population. The top left panel shows the creation rate of early-type (red, in this case) galaxies as a function of redshift from the models (solid line), and as measured by Borch *et al.* (2006). The top left panel shows the evolution of the predicted evolution of the color distribution of such a population as a function of the age of the population (the range $4-12\,\mathrm{Gyr}$ in this picture corresponds roughly to the interval $1 > z > 0$). The middle left panel shows the relationship between color offset $\Delta(g-r)$ and the truncation time, and the middle right panel between luminosity-weighted age and truncation time. The bottom panels show the predicted (dashed lines) distributions of color offset and luminosity-weighted ages, and the observations (solid lines). It is clear that the present-day color distribution of early-type galaxies is reproduced by a model in which the early-type galaxy population is being built up at the observed rate, and that in such a picture there is a broad but scattered correlation between merger=truncation time and color offset from the color–magnitude relation.

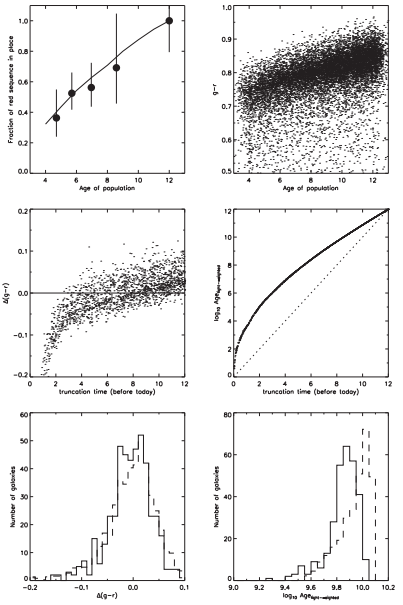

Figure 3.15: A major merger model, but with a 10% burst of star formation just prior to the truncation of star formation. Our modeling conclusions are robust to a burst of star formation on the final merger of a galaxy.

a very similar model where 10% of the stars are formed in a burst on merging are shown in Fig. 3.15, and are very similar to the truncation-only model. The upper left panel shows the build-up in the number density of model red sequence, early-type galaxies, compared with observations of the build-up of the red sequence stellar mass density from Borch *et al.* (2006). The upper right panel shows the predicted evolution in the color–magnitude relation zero point and scatter (think of this plot as a slice through the evolving color–magnitude relation at constant stellar mass): it is clear that there is a 'core' of red sequence galaxies, with a tail of blue early-type galaxies.

In the center left panel, we explore the relationship between truncation time (when star formation ceased) and color offset from the $z = 0$, t =12 Gyr red sequence. One can see that after truncation of star formation, the model galaxy reddens rapidly, within ~ 1 Gyr, to within 0.1 mag of the red sequence. The subsequent reddening of the population is much slower. 'Typical' structured early-type galaxies, with $\Delta(g-r) \sim -0.04$, would be interpreted to be between 2 and 8 Gyr from their truncation event (depending on metallicity)[8]. It is to be noted that this timescale (in its value and scatter) is in rough agreement with Schweizer & Seitzer (1992); early-type galaxies less than 2σ from the locus of the color–magnitude are expected to be more than 1 Gyr from their 'creation event'. The oldest red sequence galaxies are *redder* than the core of the color–magnitude relation. This is one of the reasons that the locus of the color–magnitude relation appears to evolve *slower* than an ageing stellar population in such models; the average color of an early-type galaxy is always being pulled bluewards by new arrivals to the red sequence.

We presented this relationship between truncation time and $\Delta(g-r)$, but can it have any relationship to reality (this, at some level, is the part of the question that Schweizer & Seitzer 1992 were less able to address quantitatively)? Put differently, in a model which is constrained to reproduce the number density evolution of early-type galaxies, is the color distribution of present-day early-type galaxies correctly predicted? The result of this exercise is shown in the lower-left panel of Fig. 3.14. The solid histogram shows the color offsets from the red sequence of the data, and the dashed histogram the major merger model. The two distributions are clearly very similar: a KS test yields a $> 10\%$ chance of the two distributions being drawn from the same distribution, which is a remarkable achievement for what is essentially a blind prediction, using only the observed red sequence number density growth and an assumed metallicity scatter.

In the center and lower right-hand panels, we show light-weighted ages of the systems

[8]Recall that many of the early-type galaxies with slightly blue colors have little asymmetry in Fig. 3.7, and the large magnitude of this metallicity-dependent scatter in truncation time for a given color offset presents a natural and appealing explanation for this phenomenology.

(comparable to the ages derived by Gallazzi *et al.* (2005)) as a function of truncation time (central panel), and the histogram of light-weighted ages (lower panel). It is clear that the luminosity-weighted ages are skewed considerably towards older ages than the truncation times, as is expected from any model galaxies that have a constant star formation rate before the truncation of star formation. The distribution of ages is reasonably encouraging, and has the asymmetry characteristic of the data. The average age of the model is offset from the data; inasmuch as luminosity-weighted ages are considerably more model-dependent than color this offset could be relatively easily remedied (but is not here, for the sake of simplicity). For example, one could bring the start of star formation to $\sim 10\,\mathrm{Gyr}$ in the past and adopt a different parameterization of the early stages of truncation of star formation in early-type galaxy progenitors.

Minor merger model

Of course, minor mergers are much more frequent than major mergers, and the minor merger of a gas-rich satellite onto a pre-existing early-type galaxy may drive the galaxy's color bluewards for a short time and induce asymmetry. In Fig. 3.16, we show a model that is identical to the major merger model above, but with the early-type galaxy population undergoing the accretion of a star-forming galaxy in the last 8 Gyr. The probability of accretion is modelled as being constant over the last 8 Gyr, with a total probability of 1 that a given galaxy accretes a star-forming satellite over that period. The mass ratio is randomly chosen from a uniform distribution over the interval [0.0,0.4]; the metallicity of the satellite is assumed to be identical to the primary galaxy for modelling convenience.

Fig. 3.16 shows many of the same phenomenologies as were seen in the major merger case (Fig. 3.14). The evolution of number of red sequence galaxies is similar, as is the color distribution produced at $t_0 = 12\,\mathrm{Gyr}$. The distribution of light-weighted ages is improved somewhat by the contribution of a small number of (bright) younger stars. Furthermore, the relationship between $\Delta(g-r)$ and the 'last' truncation time (the time of the last event to take place, be it a major or minor merger) is skewed considerably towards younger ages; it is already possible to be very close to the red sequence locus only $\sim 1\,\mathrm{Gyr}$ after a minor merger.

Bringing it together: our intepretation of the asymmety-color relation

Thus, the two key conclusions of the modelling effort are: *i)* the full color distribution of early-type galaxies at the present day is consistent with being built-up through truncation of star formation (as part of a merger event) at the rate actually observed (by observing the

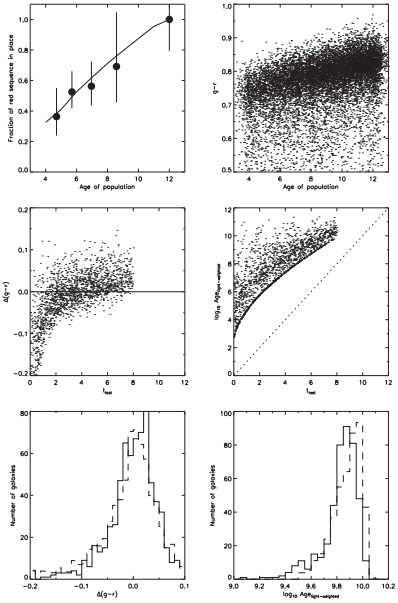

Figure 3.16: A model in which major mergers create the early-type galaxy population, with the addition of minor mergers. The figure is formatted similarly to Fig. 3.14, except that in this case the middle panels show population parameters as a function of the truncation time of the last event to take place, be it a major or a minor merger. Again, it is clear that the color distribution of early-type galaxies is well-reproduced by such a model, but that the addition of minor mergers scatters further the already scattered correlation between color offset from the color–magnitude relation and last truncation time.

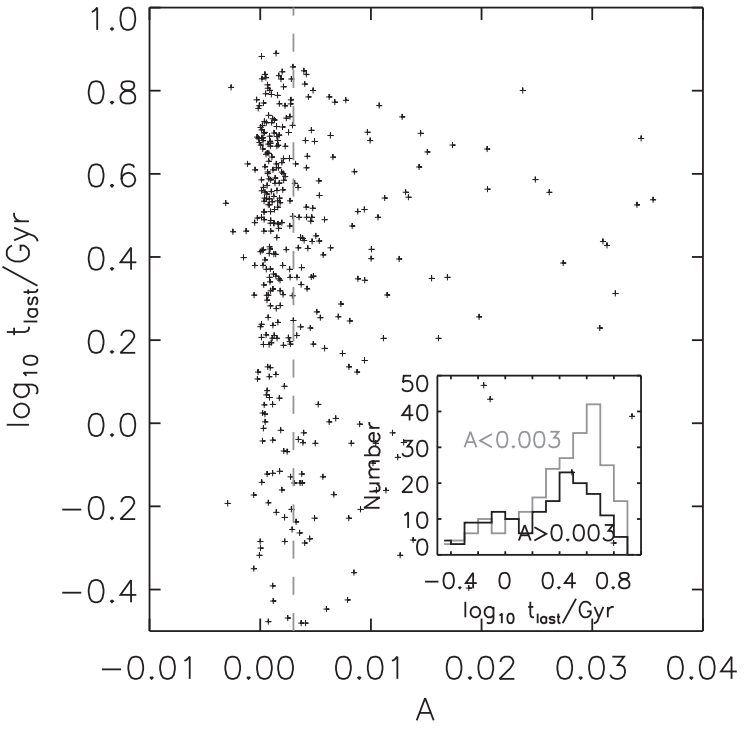

Figure 3.17: Time of last major or minor merger, inferred statistically from the luminosity-weighted age (from Gallazzi *et al.* 2005), as a function of asymmetry. The dashed line shows the dividing line between 'symmetric' galaxies with $A < 0.003$ and 'structured' galaxies with $A > 0.003$. The inset panel shows the histograms of inferred last merger times for the symmetric and structured samples.

build-up of the red sequence), and *ii)* minor mergers can lead to considerable short-lived color offsets (and presumably asymmetries), complicating considerably the interpretation of color offsets and asymmetries.

Despite the substantial scatter one expectes in a major+minor merger model, the correlations shown in the central panels of Fig. 3.16 are still substantially stronger than the observed correlations between color/age and asymmetry. We quantitatively illustrate this issue in Fig. 3.17. We show the quantity $t_{\rm last}$, the lookback time at which the last major or minor merger merger occurred in the major+minor merger model, statistically inferred from luminosity-weighted age (Gallazzi *et al.* 2005). For each galaxy, a model galaxy with very similar luminosity-weighted age is found, and its $t_{\rm last}$ is assigned to the galaxy of interest[9]. Such an inferred $t_{\rm last}$ is purely statistical, by design reproducing the trend and scatter in the middle right panel of Fig. 3.16; such a statistical estimate is for illustrative purposes only and does not represent an estimate of $t_{\rm last}$ that would be accurate in a galaxy-by-galaxy sense. We show the $t_{\rm last}$ inferred in this way as a function of asymmetry; the inset panels show the $t_{\rm last}$ distribution inferred for symmetric ($A < 0.003$) and structured ($A > 0.003$) galaxies. It is clear that asymmetry and $t_{\rm last}$ correlate only very weakly; while structured galaxies have a weak tendency towards lower $t_{\rm last}$, there are a number of structured galaxies with high $t_{\rm last}$ (i.e., rather red stellar populations). Put differently, this figure attempts to encapsulate quantitatively the fact that *neither* the stellar population parameters *nor* the asymmetry are perfect clocks: the scatter in star formation history, gas content, mass ratio and details of the merger orbits appear to add very significant scatter to the relationship between the stellar population and structural 'clocks'.

In this kind of picture, while one expects a broad correlation between color offset from the red sequence and asymmetry (as is observed), there will be a lot of scatter in this relationship for astrophysical reasons. Some relatively recent minor accretions, or interactions between already non-star-forming progenitors (dry mergers; van Dokkum (2005); Bell *et al.* (2006)), might be richly structured but already red; whereas some more ancient major interactions may remain blue for much longer, showing little asymmetry but blue colors. In such a picture, finding galaxies to fill the 'King gap' is not a particularly well-posed exercise. It is of course possible to define samples that will likely contain some fraction of the most recently-formed early-type galaxies. Such a candidate sample is presented in Fig. 3.18 and Table 3.2, where we present the 20 bluest objects in our

[9]By doing this, we end up with more scatter than is necessary in the final $t_{\rm last}$–asymmetry relation. We choose a model galaxy with a given luminosity-weighted age at random, whereas nature is likely to ensure that highly structured galaxies tend to have a lower than average $t_{\rm last}$.

sample, with $\Delta(g-r) < -0.14$. Inspection of Fig. 3.18 shows that many of them clearly show asymmetries generated by tidal effects; yet, some of this sample are not obviously asymmetric, illustrating the challenges of uniquely identifying a sample of 'young' early-type galaxies.

Yet, we would argue that there should be little concern about this difficulty to uniquely identify intermediate-age merger remnants. Much more important in our view is the finding that in this model, in which early-type galaxies are being produced at the right rate by major interactions (with or without the additional accretion of satellites), the correct distribution of early-type galaxy colors is straightforwardly reproduced. This indicates that as an ensemble, the (stellar population) properties of the early-type galaxy population are in accord with a model in which they are being built up by merging at the observed rate. Our result confirms a broad link between asymmetric (largely tidally-induced) structure and stellar population parameters, confirming a qualitative prediction of the merger hypothesis. Detailed N-body and hydrodynamical modeling will be necessary to test this link *quantitatively*, by providing directly the predicted distribution of galaxies in the color/age vs. asymmetry plane as predicted by assembly histories derived from simulations of galaxy formation in a cosmological context.

3.5.2 Musings on the role of dissipationless (dry) merging

In recent years, it has been demonstrated that largely dissipationless merging between already-formed early-type galaxies (dry mergers) play a significant role in the build-up of the early-type galaxy population (Bell *et al.* 2004a; van Dokkum 2005; Bell *et al.* 2006; Lotz *et al.* 2008; Faber *et al.* 2007). In particular, the most massive early-type galaxies appear to grow *only* by dry merging (Bell *et al.* 2004a; Faber *et al.* 2007; McIntosh *et al.* 2008), although what that growth rate is remains somewhat unclear (van Dokkum 2005; Bell *et al.* 2006; Masjedi *et al.* 2006; Scarlata *et al.* 2007; McIntosh *et al.* 2008).

In this context, it is of interest to note the population of relatively red ($\Delta(g-r) \sim 0$) and richly structured E and S0 galaxies in Figs. 3.7 and 3.12. We show a set of 20 randomly selected red highly asymmetric galaxies with $\Delta(g-r) > -0.1$ and $A > 0.02$ in Fig. 3.19. These systems are clearly undergoing (or have undergone) interactions, both minor and major, with already early-type galaxies. The relative prominence of this tail of red asymmetric E/S0 galaxies appears to be relatively independent of luminosity[10] (somewhat in contrast to the naive expectation that dry interactions/mergers are more

[10]This was checked by splitting the sample in half by luminosity. There was no significant difference between the red asymmetric early-type galaxy fractions of the bright and faint bins.

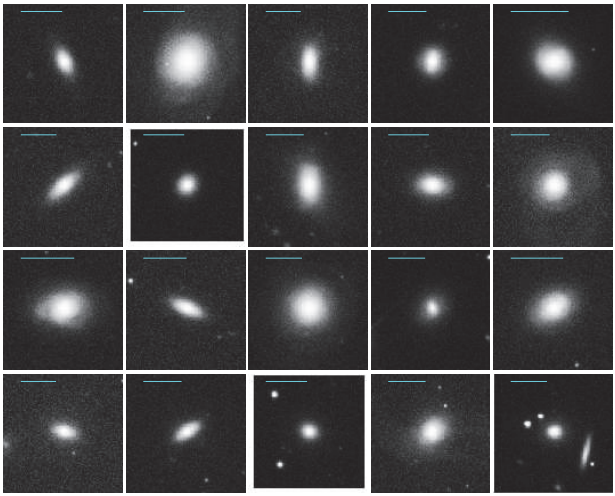

Figure 3.18: The 20 bluest E/S0 galaxies. Each box is approximately 55×55 arcsec. A 10kpc horizontal line shown for each galaxy.

Table 3.2: Parameters of the 20 bluest E/S0 galaxies

ra	dec	$\Delta(g-r)$	log (age [Gyr])	log ($Z/Z_\odot$)	A
224.1003	49.6960	−0.14			0.002
166.3744	3.6600	−0.14	9.56	−0.28	0.003
142.5555	49.4882	−0.14			0.005
17.9440	−0.6645	−0.15			0.002
169.7402	58.0566	−0.15			0.003
226.6311	40.6962	−0.16			0.013
202.2038	53.4430	−0.18			0.009
32.8899	13.9171	−0.19	8.84	−0.04	0.016
167.6620	3.7555	−0.19	9.51	−0.46	0.005
135.1501	46.6863	−0.19			0.007
50.8886	−0.4385	−0.20			0.002
11.2468	−8.8897	−0.20	9.49	0.03	0.024
153.9239	7.0522	−0.21			0.013
135.7590	40.4340	−0.21			0.012
199.9037	3.0327	−0.22	8.87	−0.20	0.037
158.8934	−0.3545	−0.23	9.38	−0.64	0.029
129.9133	3.8285	−0.24			0.022
210.1692	−1.9217	−0.29			0.004
120.8668	25.1026	−0.31			0.039
205.3000	1.7798	−0.42	9.39	−1.08	0.013

important for more luminous systems), indicating that such interactions are possible for systems of a wide range of luminosities from $L^*/5$ right up to the most luminous systems probed in our study $\sim 2-3L^*$.

It is of interest that many of the most structured early-type galaxies have this signature. Without a wide range of N-body/hydrodynamical simulations to guide us, it is difficult to properly interpret this behavior. Two aspects will clearly play a role in setting the relative prominence of blue vs. red richly-structured early-type galaxies: the ratio of blue vs. red merger partners typical of the mergers that create early-type galaxies, and the issue of when one classifies a merger remnant as an early-type galaxy (almost always for two interacting early-type galaxies, as opposed to only at very late times for the remnant of a merger between two gas-rich galaxies).

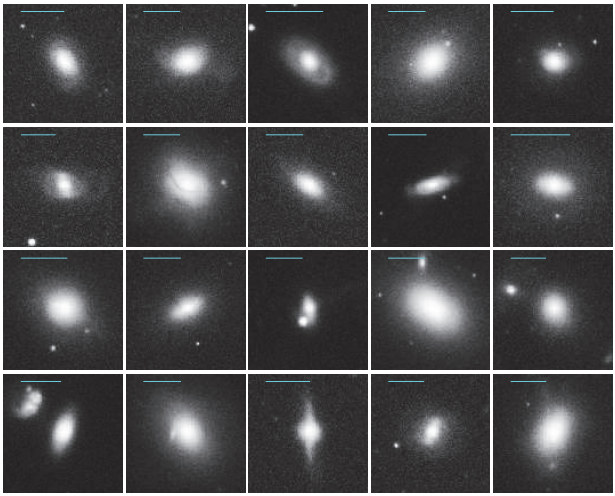

Figure 3.19: A sample of 20 red, asymmetric E/S0 galaxies. Each box is approximately 55×55 arcsec. A 10kpc horizontal line shown for each galaxy.

It is also of historical interest to ask why there were relatively few such systems in the work of Schweizer & Seitzer (1992). Number statistics clearly will play a role, but also important is the differing nature of our measures of tidal structure in early-type galaxies. Schweizer & Seitzer (1992) attempted to measure *fine* structure characteristic of mergers between gas-rich systems, whereas we explore asymmetry, which is also sensitive to large-scale asymmetries and debris fields, more characteristic of mergers between early-type galaxies. This aspect, along with details of initial sample selection, is likely to play a role in this small difference between our two studies.

3.5.3 Limitations of this approach, and outlook

As alluded to in the previous sections, a significant (and largely unavoidable) limitation of our study is the choice of asymmetry as a metric of tidally-induced disturbances. Our choice of asymmetry was motivated by the overriding consideration that the measurement be algorithmic and reproducible; asymmetry was chosen over other reproducible descriptions of structure (such as residuals from a smooth model fit; e.g., McIntosh *et al.* (2004)) to avoid symmetric bars or rings from contributing to our measure of structure. Yet, a key limitation of our metric is that systems can be asymmetric for a variety of reasons: asymmetry may be produced by fly-bys or mergers, major or minor interactions, ongoing or past interactions, gas-rich or gas-poor mergers. As such, asymmetry is a valuable but ultimately blunt tool; in this respect, our work is a useful way to start attacking the problem, but is not capable of answering higher-level, more detailed questions about the origin of the early-type galaxy population.

In principle, one could attempt to construct a richer description of structure, which would be more clearly correlated with particular types of interaction (e.g., a measure of the spatial scale of tidal debris, or a measure of the luminosity in tidal debris). Calibration with simulations would be a critical step in the establishment of such metrics (indeed, we have not yet calibrated our measure of asymmetry as part of this work). Yet, in practice, such metrics have proven difficult to reliably construct; the fine structure parameter of Schweizer & Seitzer (1992) is a good example of such a metric, which may be more physically meaningful but is difficult to reproduce.

3.6 Conclusions

One of the key predictions of the merger hypothesis for the origin of early-type galaxies is that tidally-induced asymmetric structure should correlate, at least in a broad sense,

with signatures of a relatively young stellar population. Motivated by this argument, Schweizer & Seitzer (1992) studied the relationship between color offset from the color–magnitude relation and tidally-induced fine structure, finding a correlation between the two quantities. They modeled the colors of these structured early-type galaxies, finding that such colors can be reproduced by a variety of models in which star formation truncates reasonably quickly, with typical timescales of a few Gyr to reach the colors characteristic of ancient early-type galaxies (although with much model dependence).

In this section, we have re-examined this issue, incorporating a number of improvements over the work of Schweizer & Seitzer (1992): a sample from 2MASS/NED and the SDSS that is ten times larger with well-measured colors; the use of a non-ideal, but reproducible and meaningful metric for tidally-induced structure, asymmetry; the use of luminosity-weighted ages and metallicities from the work of Gallazzi *et al.* (2005) using the SDSS; and, the ability to construct a more physically-motivated model of early-type galaxy evolution with which to interpret the results.

We found, in agreement with Schweizer & Seitzer (1992), a correlation between offset from the color–magnitude relation and asymmetry. Inspection of asymmetric systems gave weight to the notion that the main driver of asymmetries was galaxy interactions (although a variety of interactions can create asymmetries, leading to a natural ambiguity when interpreting the results). We demonstrated, for the first time, that age effects are driving this correlation. A fraction of asymmetric early-type galaxies have normal colors/ages, characteristic of mergers between already-formed early-type galaxies (dry mergers). The empirical correlation between stellar population age and tidally-induced asymmetries, for the bulk of the population, is the key result of this section, and is consistent with the basic prediction of the origin of early-type galaxies through galaxy merging, and their modification by accretion of gas-rich satellites.

We interpreted these results in the context of a model in which the number of early-type galaxies is increasing constantly with time, in quantitative agreement with measurements of the evolving number density of red sequence (early-type) galaxies. Assuming that the ultimate effect of a galaxy merger is to truncate star formation (through either gas consumption or AGN feedback), we find that such a model reproduces the distribution of color offsets from the color–magnitude relation; put differently, the merger hypothesis appears to give the correct distribution of present-day early-type galaxy colors. Under the assumption that both light-weighted stellar age and asymmetry are good 'clocks' of the time elapsed since the last major or minor merger, such a model predicts a rather tighter correlation between age and asymmetry than is observed. This indicates that

astrophysical sources of scatter, e.g., the star formation histories of the progenitors, their gas content, and mass ratios, play an important role in setting the properties of the remnants. Thus, the broad correlation between age and asymmetry is in qualitative (but not yet quantitative) agreement with such a model. The existence of such a large scatter precludes the unambiguous identification of 'young' early-type galaxies from stellar population and asymmetry data alone; confirming the merger hypothesis using a few individual examples will be a challenging and ambiguous exercise. However, there is a common feature in all models: the necessity of recent truncation times (or last merger times) for achieving the observed color distribution. Hence, in the statistical sense, the properties of the early-type galaxy population imply that an important fraction of them are the product of relatively recent mergers.

We conclude that the properties of the early-type galaxy population as a whole are consistent with the bulk of them being formed by galaxy merging (some of these mergers are between already-formed early-type galaxies). In the case of stellar populations parameters, there is a quantitative match between toy models of merger-driven growth and the observations. In the case of asymmetries, the existence of a broad and scattered correlation between stellar population parameters and asymmetry is in qualitative agreement with the expectations from the merger hypothesis, but our toy models (and even more sophisticated models) are currently incapable of providing expectations for how asymmetry should vary with time and interaction phase. Further work, using simulations to predict and calibrate the distributions of asymmetries, and using more sophisticated descriptors of galaxy structure, will help to more deeply explore this issue in the next years.

Chapter 4

Correlations between nebular emission and the continuum spectrum

4.1 Introduction

Galaxy spectral models play an essential role in the interpretation of observational data. It is important to be able to characterize the models with a small number of parameters as accurately as possible especially when working with photometric measurements. In multicolor surveys like SDSS it is possible to match the multi-band photometric data of the galaxies with model spectral energy distributions (SED's). This matching is commonly used to determine the k-corrections and photometric redshifts. The procedure implicitly involves the estimation of the spectral type. When investigating the physical properties of the galaxies, especially in evolutionary studies, it is essential to have a reliable method for template SED fitting. An example of an application of model-based k-correcting is given in Chapter 3 (Sections 3.2.1 and 3.3.2). The photometric redshift estimation will be discussed in detail in Chapter 5, for an overview, see e.g. Csabai *et al.* (2003)).

The spectral features that are reflected in the integrated fluxes, namely spectral continua and the strongest emission lines should be properly represented for all known galaxy types in the templates. This also holds if model SED's are used.

Spectroscopic observations show that there is a strong connection between the galaxy type and the intensity of the emission lines. Blue, late type galaxies with ongoing star formation, containing a large fraction of gas have strong nebular emission due to the photoionization from young stars. Early type galaxies have old stellar populations, they

are typically free of sigificant emission lines. While for early type galaxies, models with stellar continua only may work perfectly, in late type galaxies with intense star formation they are insufficient.

A spectral model suitable for description of all physical SED types should include the radiation of stars, ionized gas, and the effect of dust. There are recent works by Stasinska & Leitherer (1996); Moy *et al.* (2001); Charlot & Longhetti (2001); Panuzzo *et al.* (2003) that couple these components. Stellar continua are usually modeled using population synthesis models, e.g. Bruzual & Charlot (2003), PEGASE (Fioc & Rocca-Volmerange 1997). Emission lines in star forming H II regions are generated by photoionization codes, e.g. PHOTO (Stasinska 1990), CLOUDY (Ferland 1996). In general, a particular model is defined by age, metallicity, star formation history and initial mass function of the underlying stellar population, furthermore chemical composition, density and geometry of the ionized gas as well as dust content and certain characteristics of dust. In order to reduce the number of free parameters one usually applies simplifying assumptions and self-consistency constraints, i.e. empirical relations between the physical quantities. This enables to produce models described with a reasonable accuracy by about three stellar and three gas parameters.

Principal component analysis (PCA) proved to be a powerful tool in exploring the correlations between emission line properties and continuum spectral characteristics. Sodré & Stasinska (1999) and Stasinska & Sodré (2001) used it for statistical analysis of spectral features of nearby spiral galaxies. They identified the trends of emission line equivalent widths as a function of spectral type obtained by PCA.

The Sloan Digital Sky Survey (SDSS) provides data suitable for statistical analyzes of nebular emission on a large sample of galaxies. There have been numerous studies addressing the physical properties of SDSS emission line galaxies (Brinchmann *et al.* (2004), Kauffmann *et al.* (2003b),Tremonti *et al.* (2004)). The aim of our present study is to elucidate the statistical description of galaxy emission lines. This is complementary to the previous studies as it allows one to explore, for example, the contamination of photometric magnitudes by emission lines in an efficient way.

In this study we perform PCA on emission line equivalent widths of SDSS galaxies in order to find the minimal number of independent parameters describing the emission line pattern with a reasonable accuracy. We explore the strong connections in separation of galaxy types by color, spectral type and emission line PC coefficients. We analyze in detail their correlations with the continuum spectrum PCA parameters and determine the most probable emission line pattern and its variations as a function of the emission-free

continuum spectrum. These relations enable to add emission lines to population synthesis model spectra in an empirical way. This prescription has been used to improve the empirical spectral templates used in SDSS photometric redshift estimation since SDSS fourth data release (DR4, Adelman-McCarthy *et al.* (2006). For the details of this application we refer to 5.3.5.

4.2 Data

4.2.1 Description of the SDSS sample

We studied the emission line data of galaxies selected from the SDSS DR6 database (Adelman-McCarthy *et al.* 2008). The description of the spectroscopic sample was given in Section 1.2. The spectra are taken using 3 arcsec diameter fibres, thus in the most low redshift galaxies the spectroscopy only samples the central region. The data include redshifts, spectral type, measured characteristics of spectral lines etc. (For a technical overview of SDSS see York *et al.* (2000).)

The spectral line characteristics published in the spectroscopic catalog are evaluated by an automated pipeline. The pipeline is able to identify 48 spectral lines. In order to determine the fluxes the lines are fitted by Gaussian profiles. The database lists the fit parameters including position, height, width, equivalent width and spectral continuum flux to each line. We use these values for our analysis.

The SDSS spectroscopic catalog also provides quantities that carry useful information on the spectral shape of galaxies in a compact form. Connolly *et al.* (1995b) showed that the spectra of galaxies form a low-dimensional manifold. Using only few parameters (say three) the spectra can be described with very good precision (99% accurate) and an objective spectral classification of galaxies is also possible. The spectroscopic pipeline applies this dimensional reduction technique as detailed in Connolly & Szalay (1999) and Yip *et al.* (2004) to obtain the parameters, $\texttt{ecoeff}_i$, $i = 0..4$ for each galaxy. They are the weights of the first five PCA eigenspectra of the SDSS galaxies. The derived quantity $\texttt{eclass} = atan(-\texttt{ecoeff}_1/\texttt{ecoeff}_0)$ characterizes the shape of the spectral continuum very well. Its increasing value corresponds to rising blue end of the spectrum and decreasing 4000 Å break, i.e. small/large `eclass` values indicate early/late types. (For illustration, see the left panel of Figure 4.11, which shows the correlation of `eclass` with the color $u - r$.)

Line name	Rest wavelength [Å]
[O II] λ3727	3727.09
[O II] λ3730	3729.88
Hγ λ4342	4341.68
Hβ λ4863	4862.68
[O III] λ4960	4960.29
[O III] λ5008	5008.24
[N II] λ6550	6549.86
Hα λ6565	6564.61
[N II] λ6585	6585.27
[S II] λ6718	6718.29
[S II] λ6733	6732.67

Table 4.1: The analyzed emission lines. Reference SDSS table, see also SpecLineName table in the catalog science archive.

4.2.2 Sample selection

The sample for the present study was selected as follows. We selected objects that were classified by the SDSS photometric pipeline as galaxies. We restricted our investigations to the strongest 11 emission lines listed in Table 4.2.2. In order to include only reliable data we required the median signal to noise ratio in g^* band to be more than 10, $\texttt{sn}_0 > 10$ in the SpecObj data table of the SkyServer[1] catalog science archive. For all galaxies included in our dataset we required that all lines listed in the Table 4.2.2 are measured (125832 objects). In order to exclude measurements with extremely large errors we set an upper cut in errors of equivalent widths at 5 Å, this affected 5% of the parent sample. Our typical errors are between 0.2-0.4Å, depending on line. The cleanness of data manifests e.g. in correct ratios of the known doublet lines. In some cases the line profile fit does not resolve the broad Hα λ6565 and its two close neighbors [N II] λ6550 and [N II] λ6585 correctly. In order to get rid of this systematic error, we excluded objects with blended Hα λ6565. An undesired effect of this sampling is the exclusion of galaxies that have some broad component around Hα, e.g. Seyfert I galaxies. We also required $\texttt{sigma} < 2$ for the [O II] $\lambda\lambda$3727,3730, in order to exclude cases where the fit did not resolve the two lines but captured the blended doublet instead. We found these particular constraints the most suitable for obtaining a both clean and representative dataset.

[1] http://skyserver.sdss.org

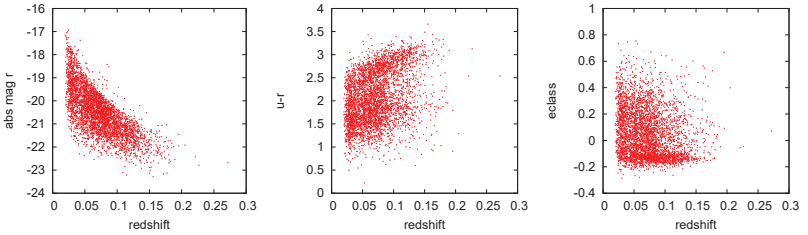

Figure 4.1: Distribution of absolute magnitude (left), color $u-r$ (middle) and the continuum spectral type parameter `eclass` (right) versus redshift shown for a 10% random subset of our sample. Two distinct groups of galaxies can be identified in both color and spectral type.

Our final sample contains 40312 galaxies. Since our selection criteria did not include any constraints on the sign of the emission line fluxes, our sample includes both emission line galaxies and objects with mainly absorption features. Some characteristics of the sample are shown in Figure 4.1, namely distribution of the data in redshift, absolute magnitude, color and `eclass`. Redshift ranges up to $z = 0.3$, with average value $z = 0.07$. In both color $u-r$ and continuum type `eclass` there are two groups of objects visible: red, early types at higher $u-r$ and low `eclass` and an other group of blue, late types at smaller $u-r$ and larger `eclass`. As described by Strateva *et al.* (2001) based on a study of SDSS galaxies, the two underlying groups of the bimodal distribution in color space can be separated by a single cut at $u-r = 2.22$. According to this criterion, 41% of our galaxies are red and 59% blue. The corresponding distribution of spectral types is manifested in `eclass` too. For more details on the distributions of galaxy spectral types in SDSS see Yip *et al.* (2004). We can distinguish between star forming (SF) galaxies and AGN using emission line diagnostics based on the line ratios $N2 = \log([\mathrm{N\,II}]\,\lambda6585/\mathrm{H}\alpha)$ and $O3 = \log([\mathrm{O\,III}]\,\lambda5008/\mathrm{H}\beta)$, which was introduced by Baldwin, Phillips, & Terlevich (1981) (BPT). (For the distribution of our data on the BPT diagram see Figure 4.13, however, the explanation of the different symbols comes later in Section 4.3.4.) In order to identify the two groups in our sample we adopt the AGN/SF separator of Kauffmann *et al.* (2003b)

$$O3 = 0.61/(N2 - 0.05) + 1.3. \tag{4.1}$$

Objects with $O3$ values above this line are classified as AGN, the rest as SF. Based on this cut, nearly 50% of the sample are SF galaxies, 18% have AGN-like emission pattern

and over 32% of the objects cannot be classified as either because they have at least one of the four lines of equation (4.1), in most cases Hβ having non positive EW's. These are objects with weak overall emission.

4.3 Analysis of spectral lines

4.3.1 Equivalent width and spectral type distribution

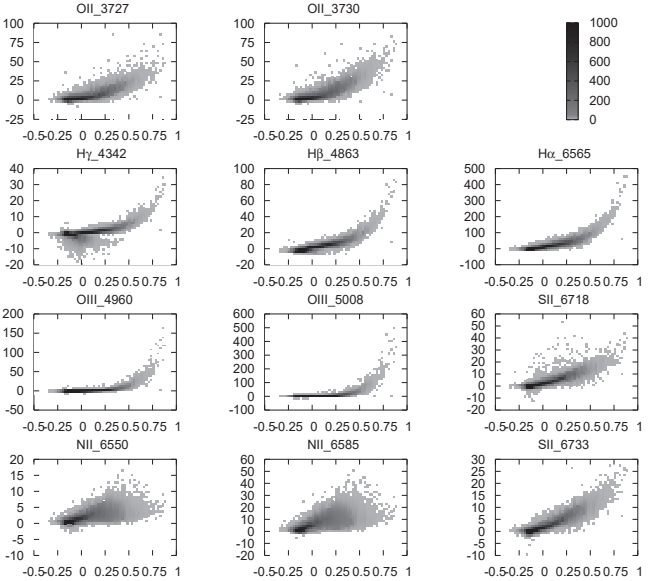

Figure 4.2: Equivalent width - `eclass` distribution of the selected 11 spectral lines. Galaxy counts are plotted in grayscale. The spectral type parameter `eclass` (x axis) is small/large for early/late type galaxies. The EW's (in Å, y axis) of all lines show a strong type dependence. The absorption dominated early type objects are situated at negative `eclass` values. At positive `eclass` the EW's of the emission line galaxies increase with type.

Figure 4.2 shows the distribution of the equivalent width (EW) values of our sample plotted against the type parameter `eclass` (the smaller the redder). The data show the

well-known tendency that the emission lines get stronger when going from early to late types, see in Figure 4.3.1 for a few examples. This trend is obvious for all 11 analyzed lines. Early type galaxies located at negative `eclass` values show mainly absorption features and almost no emission independently of `eclass`. This group can be distinguished from late type galaxies situated mostly at positive `eclass` values, whose emission tends to rise with `eclass`. However, this tendency is not the same for the different lines, which originates in differences of the physics of their formation. We also see galaxies that have apparently strong absorption at Hγ λ4342 but not at Hα λ6565, which is puzzling as typically one expects the absorption EW at Hα λ6565 to be about 60% of that at Hγ λ4342. Since our EW's are the sum of emission *and* absorption, this probably means that the Hα λ6565 absorption is filled up by emission for these galaxies. This effect is enhanced by measurement uncertainties as well. The apparent strong absorption values at [O III] and [S II] are also present due to measurement errors.

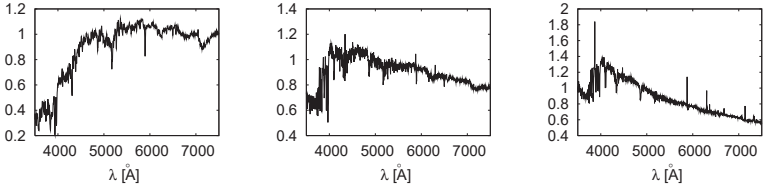

Figure 4.3: Composite galaxy spectra from early type with no emission (left) to emission rich late type (right).

4.3.2 Orthogonal approach

We try to quantify the common trends and differences in the variation of the nebular emission pattern using principal component analysis of the equivalent width data. They characterize the emission strength physically as they do not depend on distance and the effect of the galactic reddening is cancelled by normalization. At the same time, EW's are affected by the intrinsic reddening caused by inhomogeneous distribution of star, gas and dust components (Calzetti *et al.* 1994; Charlot & Fall 2000). As we are interested in data as they are observed in photometric measurements, we choose not to correct for the intrinsic reddening.

Principal component analysis (PCA) is a linear transformation of data vectors to the eigensystem of their correlation matrix. It results in an uncorrelated representation of

the data, and makes possible to identify the most relevant directions of variation by to ranking them by their information content.

The PCA of our data was carried out as follows. We represented the EW's of each galaxy by an $M = 11$ dimensional vector $\mathbf{y}$, with the average $\overline{\mathbf{y}}$ subtracted. Diagonalizing the covariance matrix we obtained the orthonormal set of M eigenvectors or principal components, PC's $\mathbf{e}^k$. We ordered them by their eigenvalues λ^k (normalized to unit trace) since they express the relative information content of each eigenvector. For each galaxy the expansion coefficients of the vector $\mathbf{y}$ form the new M dimensional vector $\mathbf{c}$,

$$\mathbf{y} = \sum_{k=1}^{M} c_k \, \mathbf{e}^k. \tag{4.2}$$

The transformation of vector $\mathbf{y}$ to $\mathbf{c}$ corresponds to a simple rotation of the data vectors to the basis where they are uncorrelated. Inverting the transformation we get back the original vectors. However, if we truncate the expansion coefficients at some $m < M$, the data will not be restored exactly. The effect of omitting the k-th principal component is the reduction of the variance of the truncated EW estimator

$$\mathbf{y}^{(m)} = \sum_{k=1}^{m} c_k \, \mathbf{e}^k, \tag{4.3}$$

by λ^k with respect to the original distribution of the data. Hence the eigenvalues are actually a measure of the importance of each eigenvector to reconstructing the real distribution of the data.

4.3.3 The principal components

Figure 4.4 shows the results of the PCA of the EW vectors. The average equivalent width for each of the 11 lines (vector $\overline{\mathbf{y}}$) are plotted on the top panel. Below are the first five eigenvectors $\mathbf{e}^k$ ordered by their eigenvalues. Their information content is 89.1%, 7.8%, 1.8%, 0.7% and 0.2% of the total variance, respectively. The numerical results are summarized in Table 4.3.3. The meaning of each principal component can be easily interpreted by comparing the weight of the i-th line in the k-th eigenvector e_i^k with the mean EW of the i-th line $\overline{y}_i$.

The first eigenvector $\mathbf{e}^1$ is very similar to the average vector $\overline{\mathbf{y}}$. It means, the most important variation in emission line equivalent widths is simply a constant multiplicative factor in the amplitude that varies from galaxy to galaxy. Larger c_1 means stronger nebular emission. The principal component is dominated by the strongest line Hα λ6565.

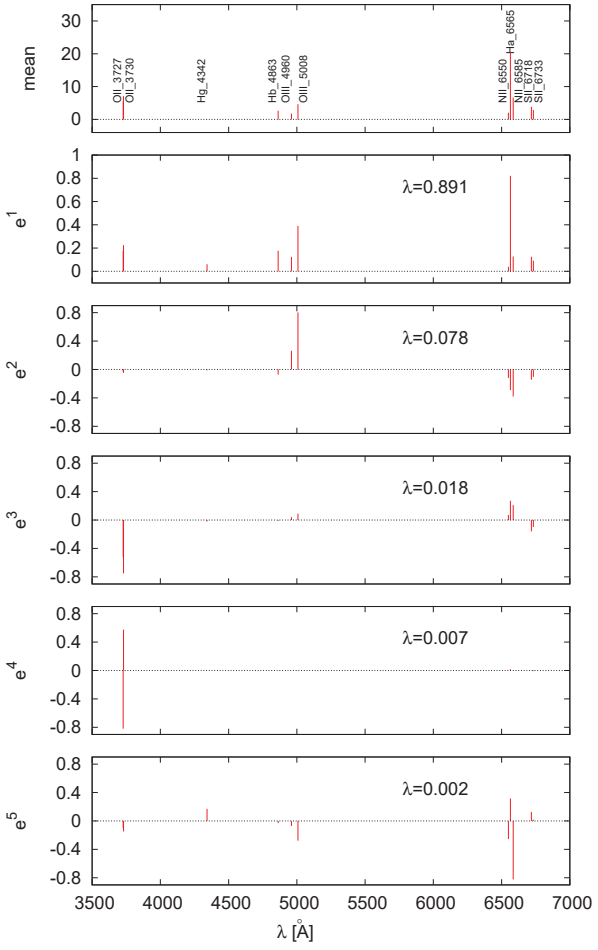

Figure 4.4: The mean vector $\overline{\mathbf{y}}$ and the first 5 eigenvectors of EW's. For each eigenvector λ denotes the relative information content. See explanation in Section 4.3.3.

Line name	e_1 (0.891)	e_2 (0.078)	e_3 (0.018)	e_4 (0.007)	e_5 (0.002)
[O II] λ3727	0.176	-0.033	-0.518	-0.819	-0.095
[O II] λ3730	0.225	-0.046	-0.750	0.574	-0.148
Hγ λ4342	0.061	-0.012	-0.021	0.000	0.171
Hβ λ4863	0.176	-0.072	-0.012	0.001	-0.031
[O III] λ4960	0.124	0.265	0.040	0.001	-0.070
[O III] λ5008	0.390	0.805	0.089	0.002	-0.278
[N II] λ6550	0.039	-0.118	0.072	-0.001	-0.254
Hα λ6565	0.821	-0.289	0.271	0.018	0.315
[N II] λ6585	0.127	-0.380	0.213	-0.003	-0.822
[S II] λ6718	0.125	-0.143	-0.159	-0.004	0.127
[S II] λ6733	0.091	-0.104	-0.098	-0.008	0.017

Table 4.2: The first five EW principal components ordered by their relative information content. The eigenvalue of each eigenvector is given in round brackets in the column header.

The second eigenvector $\mathbf{e}^2$ represents mostly the [O III] $\lambda\lambda$4960,5008 and nitrogen lines. The coefficients of [O III] $\lambda\lambda$4960,5008 and [N II] $\lambda\lambda$6550,6585 have opposite signs. The same holds for the [O II] $\lambda\lambda$3727,3730 and [N II] $\lambda\lambda$6550,6585 in the eigenvector $\mathbf{e}^3$. This enables the nitrogen emission lines to vary independently on oxygen in the reconstructed emission-lines spectrum. The EW data (Figure 4.2) show that [N II] emission grows continuously, slowly from early to late types, while the oxygen lines have a stronger type dependence getting steep especially for the extremely blue galaxies. Due to the higher ionization degree, the behavior of the [O III] doublet as a function of type is different from the other lines: while showing no significant emission at low and moderate `eclass` values, there is a steep rise at `eclass` > 0.5. Nitrogen in these components is relatively strong compared to the Balmer lines, thus both e_2 and e_3 influence the [N II]/Hα λ6565 ratio.

These eigenvectors do not change significantly the ratios of lines in the same doublet since their weights normalized by their average values are nearly equal. The constant ratios have physical reasons and it is a strong effect which persists in these components.

In $\mathbf{e}^4$ the two lines of the [O II] $\lambda\lambda$3727,3730 doublet are represented with opposite signs. The effect of this component is to change their ratio. This eigenvector reflects

the measurement errors of the [O II] $\lambda\lambda$3727,3730, which is difficult to deblend with the resolution of SDSS spectroscopy.

Vector $\mathbf{e}^5$ contains mostly [N II] $\lambda\lambda$6550,6585, with some weak Balmer and oxygen lines. Thus we expect that it might influence the precise reconstruction of [N II] lines.

Doublet lines in some higher principal components often appear with opposite sign, which is mainly the effect of errors. We do not detail the further components as their variance is less than 0.2% for each, they are dominated by noise.

4.3.4 Eigenspace representation of emission line data

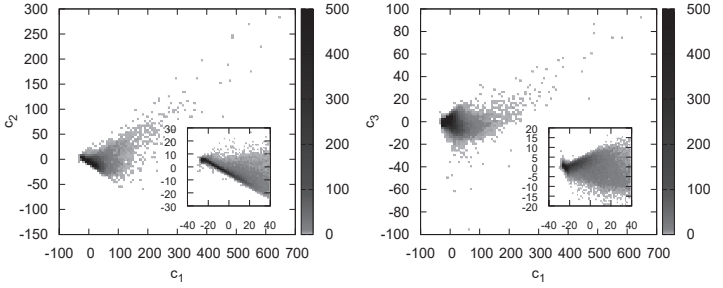

Figure 4.5: The distribution of the emission line galaxies in the subspace of the first and second (left) and first and third (right) principal components. PCA shows that the data form a roughly two dimensional manifold in the eleven dimensional EW space. The inset plots show the low c_1 region zoomed in - the distribution separates into a 'head' and a 'tail'.

The distribution of the data in the subspace of the first three principal components is shown in Figure 4.5. The range occupied by galaxies is approximately a two-dimensional curved manifold. Most of the objects form a triangular region that is closely parallel to the axes $\mathbf{e}^1$ and $\mathbf{e}^2$ and closely perpendicular to $\mathbf{e}^3$, having small c_3 coordinates. The distribution also has a 'head' at the lower end in c_1 and a 'tail' having large c_1 values. The galaxies of the 'tail' have significant contribution also from the third principal component but are still located on the curved surface which is a continuation of the main locus described above. The fact that PCA overestimates the dimensionality of the data is a

known limitation of the method. It is because PCA is linear transformation while the physics of emission lines generates nonlinear structure (see Figure 4.2).

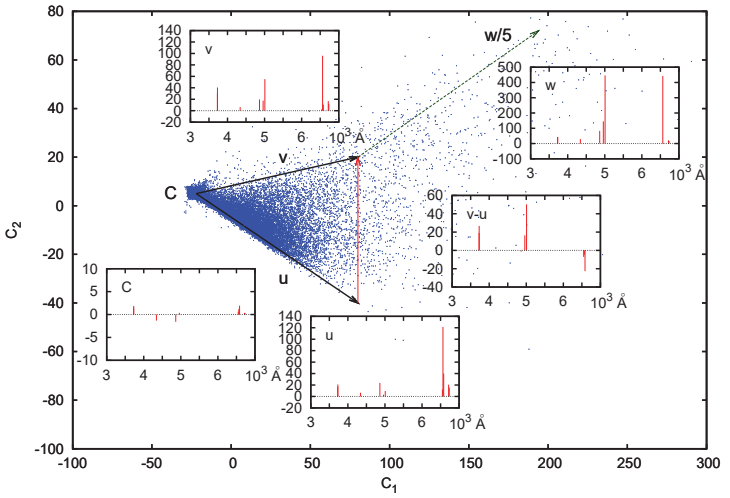

Figure 4.6: EW data projected to the subspace of the first two principal components. Inset plots show 3D-reconstructed EW's corresponding to the origin C and some representative directions. The two vectors **u** and **v** generate the main locus occupied by the majority of emission line galaxies. Their difference **v** − **u** contributes to the spectrum in the sense of enhancing oxygen, at the same time depressing nitrogen lines when going in the direction from **u** to **v**. The vector **w** together with the previous two vectors generates the strongest emission spectra. It ends out of the range of this figure in the point E= (500,250,80), the shown vector **w**/5 has the same direction and fifth the length of **w**. The spectrum of point E (not shown) is very similar to that of vector **w**. Very strong Balmer lines and [O III] λλ4960,5008, weak [O II] λλ3727,3730, as well as nitrogen deficiency can be observed.

As shown in Figure 4.6, the points in the triangular main locus embedded in the subspace of the first three eigenvectors can be generated by two vectors originated in a point C (Chan *et al.* 2003). Their coordinates in (c_1, c_2, c_3) space are: $C = (-22, 5, 0)$, $\mathbf{u} = (102, -45, 15)$, $\mathbf{u} = (102, 15, -15)$. The origin and the vectors correspond to certain emission pattern recovered from three principal components according to the equa-

tion (4.3). As the figure shows, the origin has almost no emission. Thus, since the region is situated approximately in the ($\mathbf{e}^1$,$\mathbf{e}^2$) plane, the EW values in these two vectors alone can give us some idea of physical interpretation of the first two principal components.

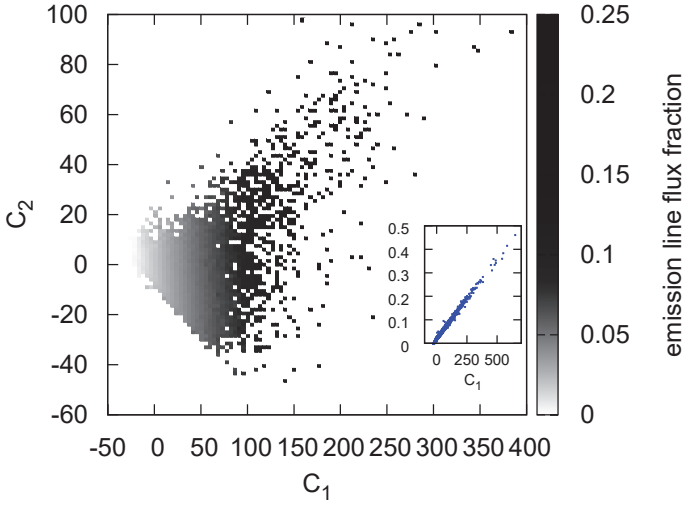

Figure 4.7: The variation of the relative emission line flux μ averaged over bins (grayscale) shows a linear relation with c_1 (inset plot).

As indicated in Section 4.3.3, the overall strength of the emission is manifested in the first coefficient c_1. Let us introduce the relative emission line flux fraction μ as the total emission line flux normalized by the continuum flux within the range $3728-6733$ Å. If we plot this quantity for each galaxy in the plane ($\mathbf{e}^1$,$\mathbf{e}^2$), we can see the monotonic growth of μ with c_1 (Figure 4.7). It is apparent that the gradient of μ is almost parallel to the axis $\mathbf{e}^1$. The inset plot shows a linear relation between μ and c_1. Specially, the 'head' of the distribution has emission flux fraction less than 1%, while the largest c_1 galaxies have $\mu \approx 0.5$, i.e. equal flux contribution from continuum and emission lines in the analyzed wavelength interval.

Another striking effect in Figure 4.6 is that while oxygen lines are relatively strong, the nitrogen emission is weak in vector $\mathbf{v}$, $\mathbf{v}-\mathbf{u}$, the difference, shows negative [N II] lines. This indicates a difference in metallicity. We estimated the ratio of oxygen and hydrogen abundances from the ratio of [N II] λ6585 and Hα fluxes using the empirical formula of Pettini & Pagel (2004).

$$12 + \log(\mathrm{O/H}) = 8.9 + 0.57 \log\left([\mathrm{N\,II}]\,\lambda6585/\mathrm{H}\alpha\right). \tag{4.4}$$

As this estimator is calibrated for H II galaxies we excluded from this analysis objects dominated by non-thermal emission by requiring $\log([\mathrm{N\,II}]\,\lambda6585/\mathrm{H}\alpha) < -0.3$. Figure 4.8 shows that the vector $\mathbf{v}-\mathbf{u}$ points approximately in the direction of the negative metallicity gradient: at a fixed contribution from vector $\mathbf{u}$ the objects tend to have smaller metallicities if the mixing ratio of $\mathbf{v}$ is larger, i.e. upwards in the diagram. The 'head' has large metallicity values which means old galaxies.

In Figure 4.6 data are plotted only up to $c_1 = 300$, however, the 'tail' of the distribution reaches nearly $c_1 = 700$. The vector $\mathbf{w}$ together with the previous two vectors generates the long tail. Its endpoint E[Å] = (500,250,80) lays out of the range of the figure, only the fifth of the vector, $\mathbf{w}/5$ is plotted. Point E represents galaxies with extremely strong nebular emission. Since vector $\mathbf{w}$ carries very strong EW values compared to vector $\mathbf{v}$ (or $\mathbf{u}$), the spectrum of the point E, as well as the points near the large c_1 end of the distribution, is very similar to that of vector $\mathbf{w}$. Very strong Balmer lines and [O III] $\lambda\lambda$4960,5008, weak [O II] $\lambda\lambda$3727,3730, as well as nitrogen deficiency can be observed. Even though our data are not reddening corrected, the small [O III]/[O II] ratio is a real strong effect. It implies a large ionization parameter of the emitting gas, which increases with c_1 at large c_1 values. In summary, the galaxies of the 'tail' have extremely strong emission, very low metallicities and high ionization parameter which indicates that they are young bursting objects.

Let us look at photometric characteristics. Figure 4.9 shows the color $u-r$ in the subspace of the first and the second PC. A red 'head' and a blue 'tail' can be seen, color is getting continuously bluer towards large c_1 values. This indicates a close relation of the distribution in emission line PC space to the bimodality seen in color. The histograms at the bottom of the diagram indicate the distribution of red and blue galaxies defined by the $u-r = 2.22$ cut of (Strateva *et al.* 2001). The transition between the two distributions is around $c_1 \approx -15$ where we have equal number of red and blue galaxies in our sample. By a cut at $c_1 = -15$ we can select 93%/86% of blue/red galaxies, with around 10%/11% contamination from the other red/blue group, respectively.

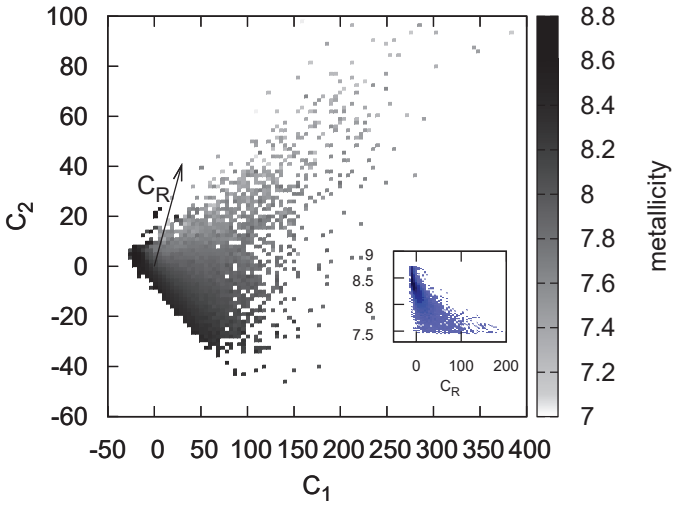

Figure 4.8: The variation of metallicity with the first two PCA coefficients. The quantity 12+log(O/H) estimated using equation (4.4) averaged over bins is plotted in grayscale. The metallicity decreases in the direction of vector $\mathbf{C}_R$, which is close to vector $\mathbf{u} - \mathbf{v}$ of Figure 4.6. The inset plot shows the distribution of metallicity along $\mathbf{C}_R$.

The color-c_1, c_2 correlation indicates that c_1 must be strongly correlated to the continuum shape of the SED. Figure 4.10 shows the relation between `eclass` and the first principal components. The c_1-`eclass` relation is very similar to the relation of c_1 and $-(u-r)$. This is is not surprising as they both measure the same effect: the difference between the intensity of the blue and the red end of the spectrum. Their relation is illustrated in Figure 4.11 and resembles two linear relations, one for the early type and one for the late type objects. The $u - r = 2.22$ cut in color corresponds to a cut at `eclass`≈ -0.05. As shown by the histogram on the bottom of Figure 4.11, it also roughly corresponds to the inflection point of the `eclass` distribution. The color distribution is also shown projected to the right margin of the diagram. In fact, the separator lines lie

somewhat bluewards from the inflection points in both `eclass` and color, which might be the effect of undersampling of early types by our selection. If we consider `eclass`$= -0.05$ as the separator of early and late spectral types, we can see that the separation is even slightly clearer than in the case of colors. The cut selects 93%/88% of late/early spectral types with a fraction of 8%/11% of early/late type objects misclassified by the cut. The correlation of `eclass` and the relative emission parameter μ is shown on the right panel of Figure 4.11. We will discuss the issue of the emission line PC's and spectral type in more detail later in Section 4.4.

Absolute magnitude in r-band is shown in Figure 4.12. Apart from the large scatter, the objects are generally fainter at larger c_1. At the largest c_1 values only low luminosity objects are present, with typical $M_r \approx 17$. This is in concordance with the earlier studies of SDSS galaxies by Tremonti *et al.* (2004) who found that the most metal deficient galaxies are faint. .

Figure 4.13 shows the connection of the first two PC's to the AGN/SF diagnostic diagram. The line ratios $N2$ versus $O3$ are plotted on the left panel, together with the AGN/SF separator line of equation (4.1). The points are colored by the first principal component, using a cut at $c_1 = -5$ which appears to separate the AGN from SF the most reliably. Red symbols mean $c_1 < -5$, blue $c_1 > -5$. This separator lies bluewards of the color or spectral type separator, as a significant fraction of AGN are bluer/later spectral type and have more contribution from emission lines than the average red galaxies. Up to the mixing at the lower edge of the BPT diagram, the two c_1 regions roughly agree with the AGN/SF separation: the c_1 cut selects 82%/84% of SF/AGN with 6.5%/37% contamination from the other group, respectively. The separation of the types is not well defined partly because of the mixing of SF and AGN activity in some low emission galaxies. On the right panel of Figure 4.13 AGN (red) and SF galaxies (blue), selected by equation (4.1) are plotted on the $c_1 : c_2$ diagram. The vertical line is at $c_1 = -5$. The plot shows the same subset of objects as on the left panel, which means all galaxies where equation (4.1) is not applicable are excluded. They are low emission objects having non-positive Hα, Hβ, [N II] λ6585 or [O III] λ5008. Nearly all of them (99.9%) are at $c_1 < -5$. Note that low nebular emission galaxies are underrepresented due to the selection criteria. If present, the missing galaxies would populate the 'head' too. We also note that AGN with broad Hα emission lines are excluded from our sample. With our present sample, the 'head' (at $c_1 < -5$) consists of low emission objects, 31% of which could be classified as AGN. The plot confirms that the 'tail' consists of SF galaxies. We can conclude that the PCA isolates two groups of objects: red, early spectral type, low emission, high

metallicity, bright galaxies, with a significant fraction of AGN in the 'head' and the rest which is blue, late spectral type, high emission, lower metallicity, fainter, star forming. In the main locus, there is a gradient of all quantities listed above. All these characteristics get continuously more prominent and reach extreme values towards the end of the 'tail'.

4.3.5 Reconstructing spectral lines

We studied the effect of truncation of the principal component basis on the restored equivalent widths. We examined the convergence of the truncated EW estimator (4.3) as a function of the number of eigenvectors used for the reconstruction. The error of the estimation can be characterized using the residuals added in quadrature over all lines

$$(\Delta y^{(m)})^2 = \sum_{i=1}^{11} (y_i^{(m)} - y_i)^2. \tag{4.5}$$

Figure 4.14 shows $\Delta y^{(m)}$ averaged over the whole sample as well as for three c_1 bins. Remember that larger c_1 values mean stronger emission and later spectral type. For the earliest bin $c_1 < 0$ the contribution of the emission lines is so small that the estimation by average (m=0 case) produces larger error than ignoring the emission lines, i.e. setting EW=0 for all lines. We demonstrate the error of the estimation by zero EW's for `eclass`<-0.05 galaxies by an arrow at the left margin of the diagram. For early spectral types, ignoring the emission would mean a 10Å error if summed over all lines.

If we drop all eigenvectors with eigenvalues less than 1%, we will have the first three PC's. Their total percent variance for the case $m = 3$ is 98.6%. Using the first three components we can reconstruct the total emission with 3Å average precision. The error is type dependent, its absolute value increases with c_1. For the strongest emission bin the average $\Delta y^{(3)}$ is 8Å. However, unlike the absolute error, the relative error defined as

$$\delta y^{(m)} = \Delta y^{(m)}/|y| \tag{4.6}$$

is smaller for stronger emission objects. The average relative error is 25% for the whole sample, dominated by the error of objects having small EW values. For the extremely strong emission bin it is only 5%.

We show the errors of the strongest, most important lines individually. In Figure 4.15 we showed the convergence of the residuals of the individual EW's of [O II] λ3730, [O III] λ5008, [N II] λ6585 and Hα λ6565

$$\Delta y_i^{(m)} = |y_i^{(m)} - y_i| \tag{4.7}$$

averaged over the sample. Similarly to the overall emission represented by the summed residuals, all the individual lines are reconstructed with an average error not larger than 2Å using three PC's (or not larger than 4Å using two PC's). For the early type objects at `eclass`< -0.05 we repeated the estimation by zero emission line flux similarly to Fig. 4.14. If we set all EW's to zero, the errors of the individual lines are still below 5Åfor this group, as shown by the errors at the left margin of the plot.

We checked the effect of the truncation on physical quantities such as metallicity and emission flux fraction if they are determined using 3-PC-reconstructed EW data. The relative emission line flux fraction μ is plotted in Figure 4.16, truncated to the first 3 PC's versus original. The rms error of the reconstruction is as small as 0.001 which means an estimation of emission line flux fraction within a precision of 0.1%. The lines reconstructed from the first three PC's obey the known ratios of the doublet lines, especially [O III] $\lambda\lambda$4960,5008 and [N II] $\lambda\lambda$6550,6585 with a relatively good precision. The line ratio [N II] λ6585/[N II] λ6550 in the reconstruction using the first three eigenvectors is 3.26, whereas the fitted ratio in the original data is 3.23. For [O III], the fitted [O III] λ5008/[O III] λ4960 ratio is 3.06 for the reconstructed data and 3.01 for the original data. These features are so strong that only higher PC's begin to violate them by including noise components. However, the flux ratios of lines that are not in the same doublet is not restored precisely. If we want to use the [N II] λ6585/Hα ratio for diagnostic purposes to distinguish between thermal emission of H II regions from the AGN-like emission, we need the first three and the fifth principal component as well. The fifth eigenvector is the one that makes possible an efficient fine-tuning of this ratio as it contains Hα and [N II] with opposite sign. Figure 4.17 shows that metallicity estimation with the first three PC's has a relatively large error, which can be suppressed by the inclusion of the fifth eigenvector. The rms errors of the reconstructed metallicity is 0.23 and 0.13 without and with $\mathbf{e}_5$, respectively. The reconstruction is less precise at high metallicities. This is because of the weak Hα and Hβ lines, since, as described in Section 4.3.4, these are typically early type galaxies.

4.4 Correlation of spectral lines and continuum features

The motivation of our current study is to explore the connection between the continuum spectral type and the emission pattern of the emission line galaxies. Although it is clear that there is no one-by-one relation, as a first step we disregard the variations and focus on

the systematic trends. For the practical applications, we would like to make predictions about emission lines based on continuum parameters. This can be then used e.g. to add emission lines to galaxy model SED's which only contain stellar populations.

We characterize the continuum spectrum with the three most informative coefficients of the spectral principal component expansion $\texttt{ecoeff}_0$, $\texttt{ecoeff}_1$ and $\texttt{ecoeff}_2$. We investigate their link to the first three emission line PCA coefficients c_1, c_2 and c_3 that proved to be essential in reconstructing the emission with a sufficient accuracy. Figure 4.18 illustrates how these parameters are linked to each other. The data points at all diagrams are colored by $\texttt{ecoeff}_0$, $\texttt{ecoeff}_1$ and $\texttt{ecoeff}_2$. For late type objects with significant emission lines a mapping from the first three `ecoeff`'s to the first three c_i appears to be possible.

The correlation of each of the first three c_i are plotted against `eclass` in the top panels of Figure 4.19. The first coefficient c_1 exhibits the strongest correlation with the continuum features. This coefficient has also the largest information content about the emission lines. There are apparent systematic trends in c_2 and c_3 as well. Given a continuum spectral type, we can determine the expectation values and variances of the emission line EW's based on these empirical relations. Continuum PC's $\texttt{ecoeff}_0$-$\texttt{ecoeff}_2$ carry even more information that can be used to establish an empirical relation. As indicated in the previous section and as one can see in the top panels of Figure 4.19, early type galaxies might be fitted by constant values which would yield nearly zero flux. However, now we choose to treat all data equally. We have checked that the two approaches do not not make a significant difference. We fit a second order polynomial of three variables $\texttt{ecoeff}_0$, $\texttt{ecoeff}_1$ and $\texttt{ecoeff}_2$ to each of c_1, c_2 and c_3,

$$c_i = \alpha_i + \sum_{k=0}^{2} \beta_i^k \texttt{ecoeff}_k + \sum_{k=0}^{2}\sum_{l=k}^{2} \gamma_i^{kl} \texttt{ecoeff}_k \, \texttt{ecoeff}_l. \tag{4.8}$$

The fitted coefficients are listed in Table 4.4. We can use this empirical relation to estimate emission properties solely from continuum features of the spectra. The residuals of the fit, $c_i(fit) - c_i$, which characterize the goodness of the estimation, are shown on the bottom panels of Figure 4.19. The rms error of the fit residuals is plotted as a function of spectral type too. The origin of the scatter is mainly the cosmic variance, which includes the effect of geometry (Yip *et al.* 2008) and other physical parameters not fully covered by the spectral classification parameter. The scatter gets large towards the largest `eclass` values. However, as the flux values themselves are large here, the resulting relative flux error is smaller than for earlier types.

We use the fitted $\tilde{c_i}$ values to reconstruct the emission lines analogously to equa-

	c_1	c_2	c_3
α	1657.808	674.528	508.194
β_0	-4390.826	-191.740	-639.650
β_1	488.087	108.530	-27.376
β_2	-31.382	-150.234	-16.725
γ_{00}	2717.923	-481.908	132.141
γ_{01}	-610.956	-59.598	25.392
γ_{02}	-153.612	226.695	-2.644
γ_{11}	546.841	-498.071	-245.997
γ_{12}	212.847	82.486	-36.546
γ_{22}	433.339	-617.698	-375.677

Table 4.3: Fitted polynom coefficients of equation (4.8) for c_1, c_2 and c_3 as a function of $\texttt{ecoeff}_i$.

tion (4.3):

$$\tilde{\mathbf{y}} = \sum_{k=1}^{3} \tilde{c}_k \, \mathbf{e}^k, \tag{4.9}$$

We compare the emission lines coming from this estimator with the measured values and calculate the errors described in equations (4.5 – 4.7), by substituting $\tilde{y}$ instead of $y^{(m)}$. The errors of this prediction are plotted in Figures 4.14 and 4.15 with small arrows at the right margin of each plot. We find that the strength of the total nebular emission can be predicted from the spectral continuum with an average accuracy of 5Å or 40% for the entire sample. However, for the objects of our greatest interest, those having significant emission, the relative precision is better. The average errors for the $0 < c_1 < 100$ bin is 10Å (20%), for the strongest emission bin ($c_1 > 100$) it is 25Å (≈10%). Estimating by the average only, without using the continuum dependence the error can be as large as ≈ 100%.

We can estimate, how well the reconstruction of emission lines works in the terms of photometry. We simulate photometry by convolving the SDSS filters with the spectra of the objects. We investigate the impact of emission lines by omitting them, convolving just the continuum and comparing these magnitudes with the values obtained from the entire spectrum (continuum+lines). The results for g, r, and i band are shown in the first and second row of Figure 4.20. The impact of the nebular lines is strongly type-dependent. For the strongest emission objects at high `eclass`, the magnitude difference due to the lack of emission lines can reach 0.5^m in g band. This is the effect of [O III] λλ4960,5008

and [O II] $\lambda\lambda$3727,3730. The largest difference in r and i band is $\approx 0.2^m$. Note the redshift dependence due to lines being redshifted into and out from the filters. Most apparent is the r band, the low redshift hump comes mainly from Hα the high redshift hump from [O III] $\lambda\lambda$4960,5008, which are then redshifted into i band. On the bottom three plots we estimated the emission lines from $\texttt{ecoeff}_i$, using the fitted $\tilde{c}_i$ values and Eq. (4.9). We carried out simulated photometry with continuum + predicted lines, and compared with photometry simulated with real emission lines. The results show that the prediction approximates the original values with a maximum error $\approx 0.1^m$ in g for the strongest emission objects and 0.05^m for r and i. The rms error is of order of 0.01^m for the extremely high emission bin $\texttt{eclass}> 0.6$ and $\approx 0.001^m$ for $\texttt{eclass}< 0.6$. (We note that the early types at $\texttt{eclass}< -0.05$ have errors not larger than this even if no emission line flux is added.) This precision is sufficient for the most photometric applications. For example, one can add spectral lines to model SED's of stellar populations or any empirical spectra with missing emission lines in a way which makes the continuum features and emission pattern consistent with the observations. As an example of practical applications a similar method has been successfully applied to improve the spectral templates used in SDSS photometric redshift estimation. We have shown how the relations shown in Figure 4.19 can be used to predict the expectation values of EW's based on the spectral type. The variance may serve as an additional information when a simulated distribution of emission pattern is generated.

4.5 Application to model spectra

The photometry check in the previous section (Figure 4.20) implies that the absence of emission lines in model spectra may yield large photometric errors, when they are used for comparison with observational data. For the latest type galaxies with strongest star formation the error can be $> 0.1^m$ in g, r, i. We also demonstrated that emission lines given in accordance with the empirical relation reduce this error significantly.

Now we show an example of application of these results to model SED's. The Bruzual & Charlot (2001) (BC01) model spectra are constructed by stellar population synthesis and thus have no emission lines (for a more detailed discussion see Section 5.2 or Bruzual & Charlot (1993)). We used the empirical relation between $\texttt{ecoeff}_i$ and c_k to generate lines for them. For red galaxies (roughly at $\texttt{eclass}< 0$) we found that the emission lines are not significant, therefore we did not modify the spectra. Here we note, that in the results presented in this section we used a previous version of the fit, which only

made use of the first two continuum parameters, with a somewhat different philosophy, including correlation analysis and a subsequent more simple fit, but this does not affect the numerical results significantly.

First $\texttt{ecoeff}_i$ are calculated for the line-free spectrum by expanding them in the spectral eigensbasis of Yip *et al.* (2004) which were also used to obtain $\texttt{ecoeff}_i$ for the SDSS galaxies. Then lines are added based on these values using the results of the emission lines vs. continuum empirical fit. This, however, is not the final result. The newly added emission lines will modify the `eclass`value. Although the first three eigenspectra are dominated by the continua and emission lines become significant only in the higher components, they do have some weak emission lines. Therefore, to make the lines consistent with the real spectral parameters, we recalculate the expansion and replace the lines. The recalculation of $\texttt{ecoeff}_i$ and subsequent addition of emission lines is repeated until the procedure converges. Figure 4.21 shows an illustration of this process. The solid line on the left panel is the original BC01 spectrum (parameters given in the figure caption). The inset plots show enlarged the regions around [O III]and Hα λ6565emisson lines, respectively. The dotted lines having emisson lines are the spectra reconstructed from the first five expansion coefficients ($\texttt{ecoeff}_i$, $i = 0..4$). It is approximately the spectrum of a typical SDSS galaxy having continuum shape similar to this model spectrum. The solid line in the right panel is the result of the line-adding procedure. The dotted line is the same as on the left panel. One can see that the resulting emission lines of the model spectrum are consistent with the reconstructed spectrum reflecting the statistics of the SDSS galaxies.

With this algorithm we simulate the main trend of the variation of emission lines with continuum. This is a one-to-one relation. However, in the observed data there is a scatter that can be seen in the fit residuals on the bottom panels of Figure 4.19. The effect of the not perfect fit may be present in the residuals, however, as we have argued, most of this scatter is 'real', originating in the cosmic variance. We can make use of the quantitative characteristics of the residuals to mimic the variability of the real galaxy population. We choose an approach where we only vary one parameter to generate a scatter to all emission lines. We perform PCA on the fit residuals and measure the variance of the data σ in the first component. We can randomize this 'scatter parameter' and then, with an inverse transformation with the first residuum eigenvector (analogous to equation 4.2) we get a scatter value for all emission lines. In Fig. 4.22 we show the predicted emission lines (red thick line) together with the $\pm 1\sigma$ variations (grey thin lines) for a BC01 model. The inset plots show the regions next to emission lines [O II], [O III]and Hα λ6565, respectively. As

can be seen, in the case of this model ($\tau = 1.54$ Gyr, $t = 6.17$ Gyr, which means only moderate present day star formation), the weakest possible emission pattern allowed by the 1σ variation has no emission in some of the lines.

4.6 Conclusions

Using principal component analysis of equivalent widths of the eleven selected emission lines we found that nearly 99% of information is included in the subspace generated by the first three eigenvectors. They reconstruct emission line fluxes within a precision of 5-25%, depending on spectral type. The separation of galaxy types by stellar population characteristics like color or spectral type `eclass` is clear in the eigenspace of emission lines too. Early type galaxies with old stellar populations (including some fraction of AGN with weak emission lines) can be clearly separated from the star forming, blue late type galaxies.

We found that based on a three-dimensional eigenspace representation of continuum spectra there is a simple way of estimating the most probable emission line pattern and its variation. If the real emission lines are replaced by lines generated using this empirical relation, the photometry of the real spectra can be reproduced with a precision $< 0.1^m$ in the investigated bands. We applied the empirical prescription for adding lines to model galaxy SED's to provide them with emisson lines consistent with the population statistics. The applications include the comparison of photometric observations with models, e.g. determining k-corrections and absolute magnitudes. A practical application in photometric redshift estimation will be presented in Section 5.3.5.

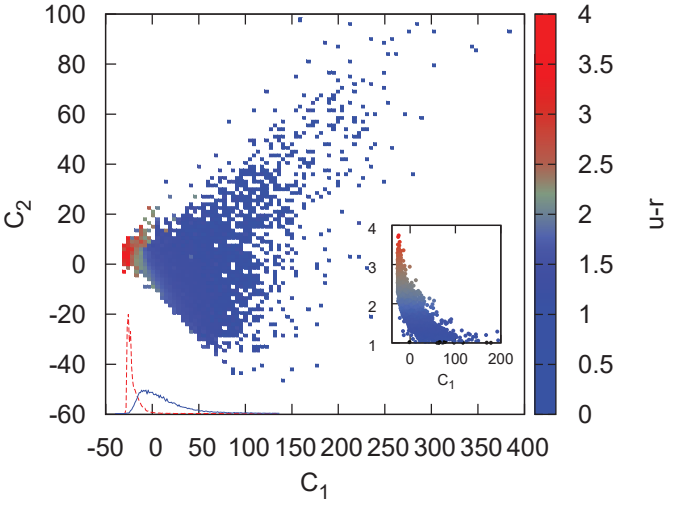

Figure 4.9: The distribution of $u-r$ color in the plane ($\mathbf{e}^1$,$\mathbf{e}^2$). Main plot: color averaged over pixels ($u-r$ color-coded, red means $u-r > 2.22$, blue $u-r < 2.22$, gray is the transition between them). The lowest c_1 values are dominated by red objects, blue becomes dominant at higher c_1 values. Transition from red to blue types ($u-r = 2.22$) is around $c_1 = -20$. Inset plot: $u-r$ vs. c_1 is a monotonic relation, the higher c_1, the bluer objects. However, low c_1 ranges have mixed colors, the relation gets tighter at higher c_1. Histograms at the bottom: c_1 distribution of the two color types. Red ($u-r > 2.22$) subset plotted with red dashed line, blue $u-r < 2.22$ subset with blue solid line.

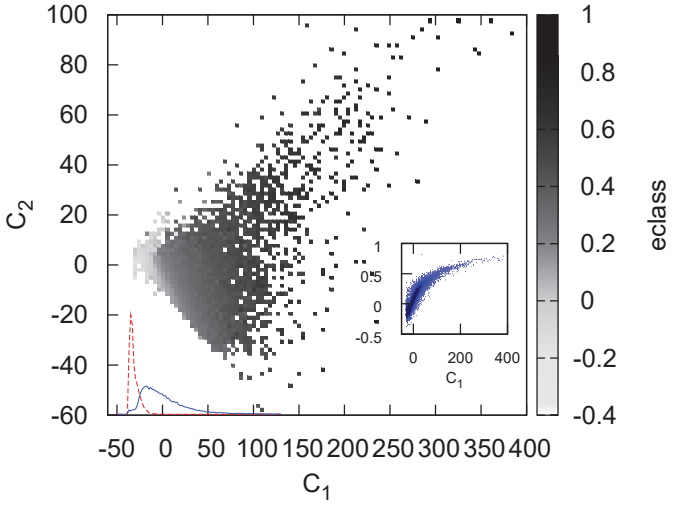

Figure 4.10: The distribution of the spectral type in the plane ($\mathbf{e}^1$,$\mathbf{e}^2$). Main plot: `eclass` averaged over pixels. Inset plot: `eclass` vs. c_1. We can see the same tendency for `eclass` as for color in Figure 4.9. Early types (negative `eclass`) are at negative c_1 values, late types (positive `eclass`) are at higher c_1. Histograms at the bottom: c_1 distribution of the two spectral type bins. Red dashed line: `eclass`< -0.05, blue solid line: `eclass`> -0.05.

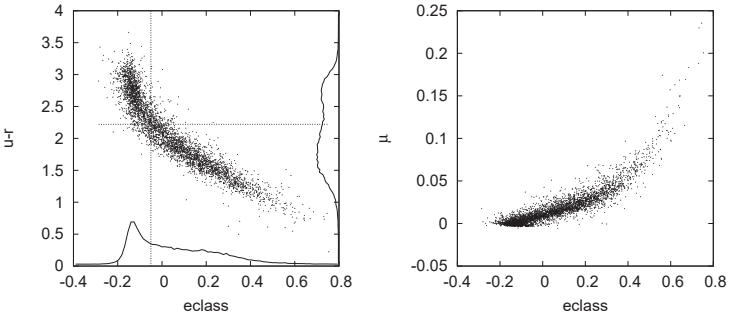

Figure 4.11: Left: color $u - r$ versus spectral type `eclass`. The relation is nearly linear, apparent bimodality. The $u - r = 2.22$ separator of blue and red types (vertical dotted line) corresponds to `eclass`$\simeq -0.05$ (horizontal dotted line). The histogram of the `eclass` distribution is plotted on the x axis. The distribution of $u - r$ is projected to the right margin. Right panel: connection between spectral type and relative emission parameter μ.

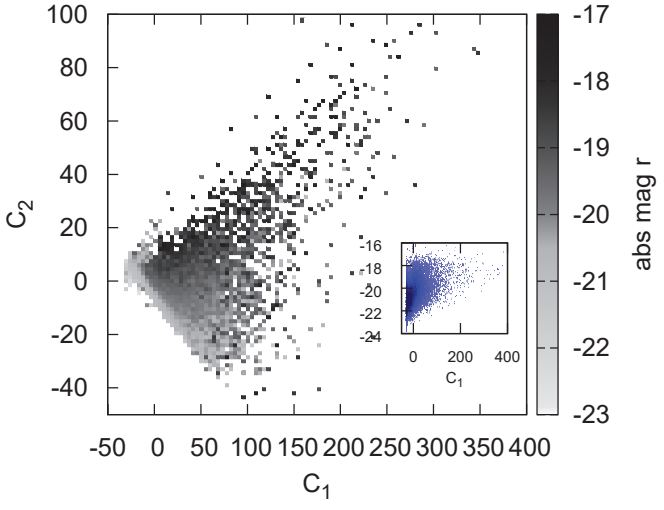

Figure 4.12: The r-band absolute in the plane of the first two PC's. Main plot: M_r averaged over bins (grayscale) in the plane ($\mathbf{e}^1$,$\mathbf{e}^2$). Inset plot: M_r versus c_1. On average, luminosity decreases with increasing c_1, however, the scatter is large. The strongest nebular emission objects are the faintest ones, with $M_r \approx -17$.

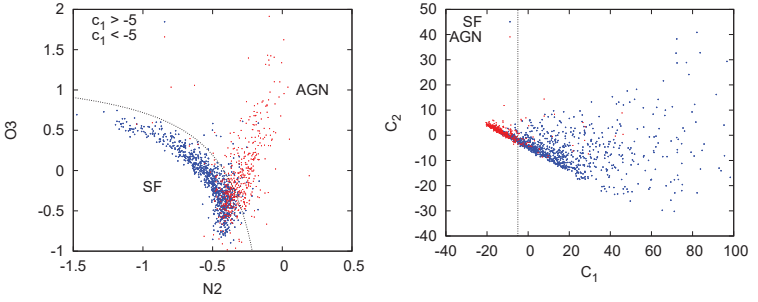

Figure 4.13: Left: N2:O3 diagnostic diagram for distinguishing between star forming galaxies and AGN. The dotted line shows the AGN separator of equation (4.1). The two types of symbols are selected by $c_1 = -5$ cut. Righ: AGN (red) and SF (blue) defined by the separator on the right panel. AGN are situated at $c_1 < 0$. A part of the 'head' is missing because of some negative values among the EW's.

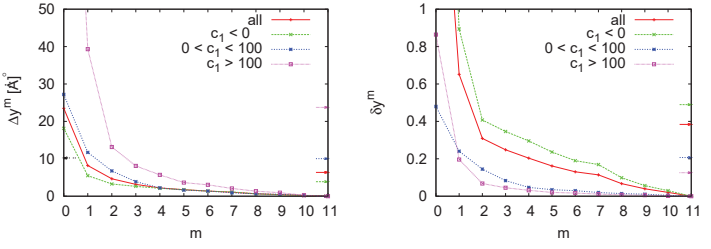

Figure 4.14: The convergence of the spectral line reconstruction. The error of the reconstruction as a function of truncation limit is plotted. m: number of eigencomponents kept; $m = 0$ represents the reconstruction using the mean EW for each line, without any principal components. Left: EW residuals summed over all lines. Right: relative error. The results are shown for all galaxies (solid thick line) and for three c_1 bins; larger c_1 values indicate stronger nebular emission. All EW's can be well reconstructed using the first 3 eigencomponents. The single arrow at the left margin of the left panel denotes the error of ignoring the emission line flux in `eclass`< -0.05 early type objects, see explanation in text. The arrows at the right margin show the same quantities for the prediction made from continuum expansion coefficients, see explanation in Section 4.4.

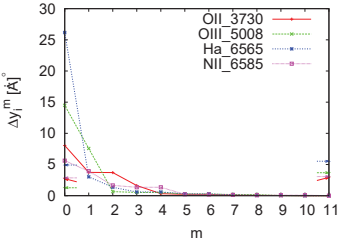

Figure 4.15: Reconstruction of four selected emission lines using m eigencomponents. Δy_i: sample-averaged absolute EW error, m: same as in Figure 4.14. The arrows at the left margin show the error of estimation by zero emission line flux for the `eclass`< -0.05 subset. The arrows at the right margin show the same quantity for the prediction made from continuum expansion coefficients, see explanation in Section 4.4.

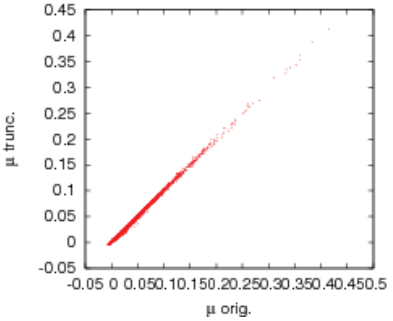

Figure 4.16: Reconstruction of the relative emission strength using truncated eigenbasis. The rms error is 0.001.

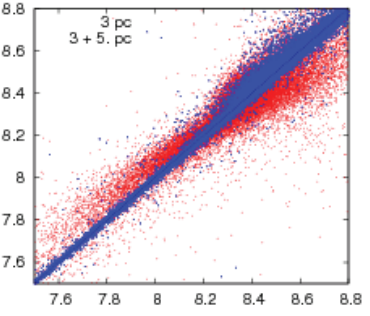

Figure 4.17: Reconstruction of metallicity using truncated eigenbasis. With the first three eigenvectors, the error is relatively large. Adding the fifth eigenvector suppresses the error significantly.

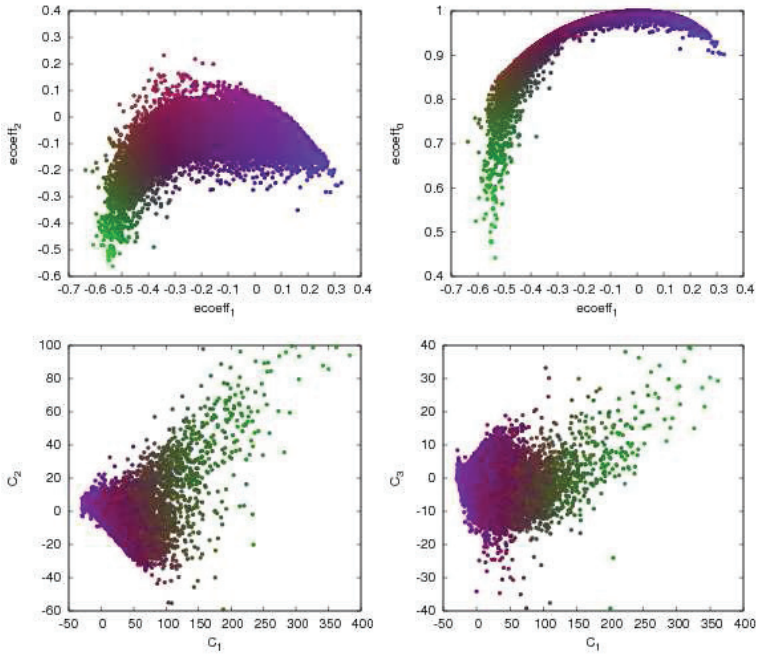

Figure 4.18: Data points plotted in the planes $\texttt{ecoeff}_1$:$\texttt{ecoeff}_2$ (top left), $\texttt{ecoeff}_1$:$\texttt{ecoeff}_3$ (top right), $c_1 : c_2$ (bottom left) and $c_1 : c_3$ (bottom right). The coloring is made by rgb-coding of $\texttt{ecoeff}_1$ (green), $\texttt{ecoeff}_2$ (blue), $\texttt{ecoeff}_3$ (red). The same coloring is applied for the emission line PCA subspaces. By matching the points of same colors in the various plots we can see how the $\texttt{ecoeff}_i$ regions are mapped into the c_i space. The picture shows, that (at least for $\texttt{ecoeff}_1 < 0.1$, which means positive `eclass`, later type objects) there *is* a mapping. Early types are not resolved, however, they are located in the 'head' of the distribution, having weak emission lines.

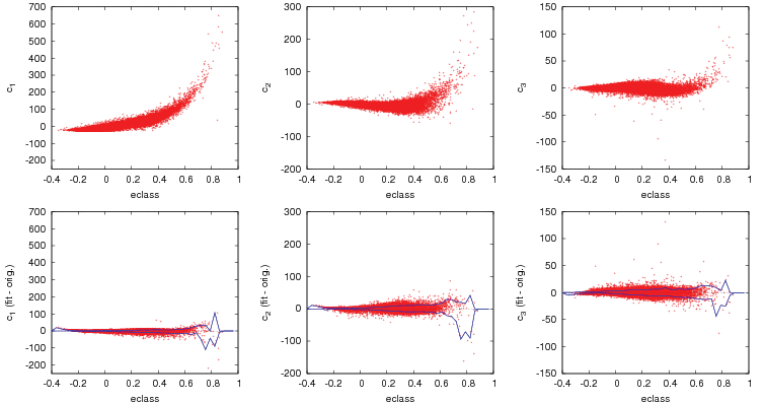

Figure 4.19: Top: PC coefficients as a function of `eclass`. Bottom: The residuals of the first three emission line PC coefficients after subtracting the continuum fit, as a function of the continuum type `ecoeff`. An empirical connection between the continuum spectral shape and the nebular emission pattern can be established from the observed correlations. The average residual scatter after subtracting the fitted estimator is plotted with blue dotted lines as a function of the spectral type.

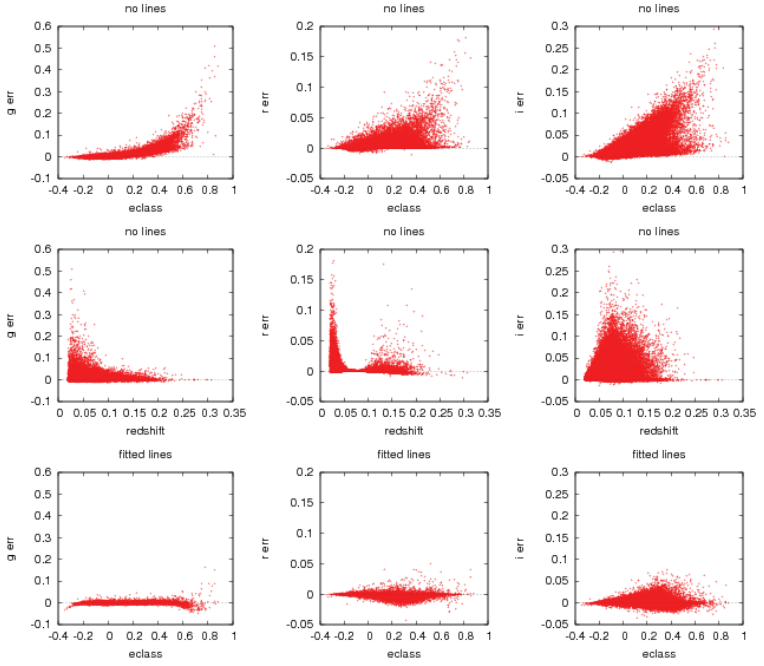

Figure 4.20: Difference in g, r and i magnitude, if the photometry is made without emission lines, as a function of spectral type (first row) and redshift (second row). Late spectral types at larger `eclass` having stronger nebular emission will have larger errors. Redshift dependence shows pattern caused by lines being redshifted into and out from the filter. Third row: magnitude errors of the simulated photometry with emission line from the continuum fit, as a function of spectral type.

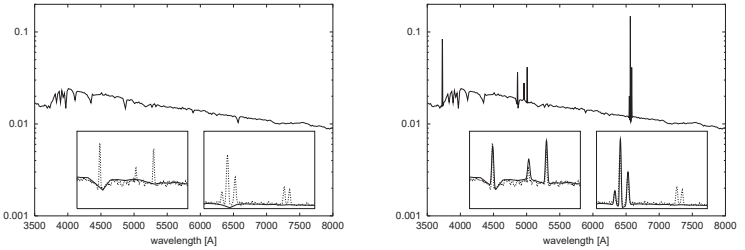

Figure 4.21: Reconstructed emission lines from the stellar continuum. Left: The $Z = 0.02$, $\tau = 19.44$ Gyr , $t = 3.20$ Gyr BC01 model (solid line). The inset plots show the emission line regions. Dotted line: the reconstructed spectrum of the model SED using the first five eigenspectra. Right: dotted line same as the left panel; solid line: BC01 spectrum with emission lines added from the continuum fit. (In this vesion the weak S[II] lines visible in the right end of the right inset were ingored.)

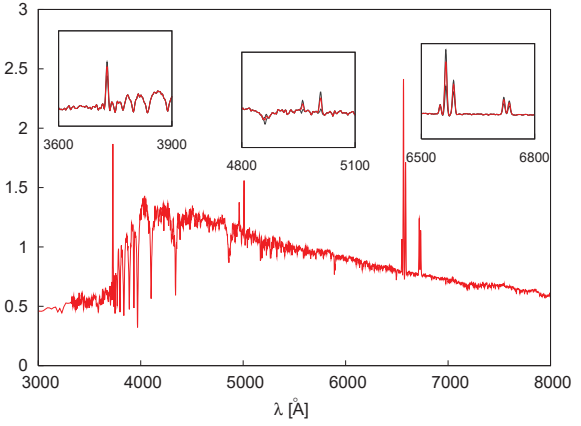

Figure 4.22: Variability of the reconstructed emission lines. Reconstructed emission lines from the stellar continuum for a BC01 model ($Z = 0.02$, $\tau = 1.54$ Gyr, $t = 6.17$ Gyr). The plot illustrates the the variability in the type dependence of the emission lines. Inset plots show the emission line regions zoomed in. The red solid line shows the expectation value, the grey lines the emission pattern deviating by $\pm 1\sigma$ from the mean.

Chapter 5

Photometric redshift estimation

5.1 Introduction

In the previous sections we have shown several examples of how studies of the color distribution of galaxies reveil fundamental information on their stellar populations, structure and evolution. In this chapter we discuss another important field of application. Detailed studies of the dimensionality of the color-space distribution of galaxies (Connolly *et al.* 1995a) has shown that the observed objects are localized to a two or three dimensional narrow region (Figure 2.1 shows a two-dimensional projection). The surface is span by two major trends: variation in stellar population properties and redshift. The tight mapping between redshift and the color locus enables the estimation of the redshift of a galaxy from its photometric data only.

The reason originates in the nature of the galaxy spectra. In multicolor photometric measurements the spectra are sampled with broad-band filters. (The SDSS filter set is shown in Fig. 1.1.) The integrated fluxes (in SDSS: five magnitudes for each object) represent a low resolution, "rough" spectrum. Thanks to the characteristic, slowly varying, non scale-invariant shape of the galaxy spectra (see several examples in Figure 4.3.1), redshifting has a specific effect on the multi-color magnitudes. The shift in wavelength can be distinguished from other effects, like differences in apparent magnitude or stellar population.

The methods where redshift is determined from colors or magnitudes are generally called photometric redshift estimation. To distinguish from spectroscopic redshift, they are often referred to as photo-z. The advantage of the photometric estimation of redshifts is its speed. Spectroscopy is a time consuming measurement. To obtain a high quality spectrum with a wavelength resolution of $\sim 1\,\text{Å}$, long exposure time is needed. The

photometric measurement with the resolution given by the typical width of the filters, $\sim$ 2000 Å is orders of magnitudes faster. Furthermore, photometric information is available even for distant and faint objects, where spectroscopic measurements are not feasible. As mentioned in Section 1.2, the SDSS photometric catalog contains over hundred times more galaxies than the spectroscopic sample. Albeit the accuracy of the photometric redshift is worse than the spectroscopically determined one, the larger statistic and deeper sample can have an advantage over the high precision especially in statistical studies of galaxy properties, distribution, large scale structure and evolution.

The idea of photometric redshift estimation is not new. Baum (1962) used multi-color photometric data to determine redshifts of elliptical galaxies by comparing them to each other. This was possible because of the strong 4000 Å break in early types (see left panel of Figure 4.3.1). The early photo-z works include developments by Butchins (1981, 1983), who used magnitude prior to break some of the degeneracies. Puschell *et al.* (1982) used a χ^2 fit of model and empirical spectra to the photometric data. Koo (1981, 1985, 1986) used synthetic spectra (Bruzual, 1983) to identify the redshifting tracks in colors and used the iso-redshift surfaces to estimate the redshift. In the work of (Connolly *et al.* 1995a) which has already been mentioned, a simple (linear or polynomial) relation between colors and redshift is assumed and determined using a training set with known photometry and redshifts. The results can be then applied to samples with similar photometric characteristics.

These methods can be basically divided into empirical and template fitting methods. In the empirical approach, a training set of galaxies with known redshifts and photometry is used to establish an empirical relation between colors and redshift. These methods include the already mentioned local fits in color or magnitude space, nearest neighbour estimator or neural networks. The advantage of this approach is that all physical effects, like dust of galaxy evolution that are present in the training set are implicitly included in the empirical relation. The drawback is that this redshift estimation only works for objects similar to those in the training set, the empirical relation cannot be extrapolated for example to larger redshifts.

The other approach, the template fitting techniques is based on matching template spectra with the photometric data (Koo 1985; Connolly *et al.* 1995b; Gwyn & Hartwick 1996; Fernández-Soto *et al.* 1999; Benítez 2000). The templates have to be similar to the spectra of the measured sample. One option is to use model SED's, like Bruzual & Charlot (2001) or PEGASE (Fioc & Rocca-Volmerange 1997). The use of models is strongly physically motivated. Being defined by a set of physical parameters, they

allow, in principle, the estimation of these physical parameters of the galaxies. In the next Section 5.2, we will discuss the application of Bruzual-Charlot synthetic spectra in redshift estimation of SDSS galaxies. Widely used are also empirical template sets, for example the four SED's of Coleman, Wu & Weedman (1980) (CWW).

The photometric redshift estimation of SDSS is derived from the tempate fitting technique. Budavari *et al.* (1999) showed that the galaxy spectra can be well represented as a one dimensional sequence. Thus besides the redshift, a single spectral type parameter can be estimated from photometry. Another important improvement was achieved by the hybrid method that holds the advantages of both template fitting and empirical approach (Budavari *et al.* 2000). Using a training set of galaxies, the spectral templates are iteratively repaired to match better the photometry of those objects of the training set which are similar to them. In SDSS, this hybrid method is used for photo-z. The templates are derived from the CWW spectra, which are then trained with this technique. Figure 5.1 shows the repaired templates with blue dotted line. These are then interpolated to form a continuous set (red solid lines). Using χ^2 minimization, the best fitting redshift and the single spectral type parameter (photo-z type) is assigned to each galaxy. For a complete overview of SDSS photometric redshift estimation see Csabai *et al.* (2003).

The accuracy of the photometric redshift estimator depends on a number of conditions. It is obvious, that the quality of the photometry, the measurement errors will largely influence the precision of photo-z. Photometric redshifts are therefore naturally less reliable for faint objects. Type dependence is also an important issue. Redshift estimation is robust for early type red galaxies because of their prominent 4000 Å break. Blue galaxies tend to have more uncertainties in photo-z. Their spectrum is more power-law like (see middle and right panel of Figure 4.3.1), which cannot be distinguished from the redshifted form. This brings in a degeneracy into the color-type-magnitude parameter space. The differences in distribution of the two types detailed in Chapters 2 and 3 also acts in favor of the early type objects. The faint end cutoff in the blue luminosity function means that there are no bright blue galaxies for calibration. The distribution of blue objects being shifted towards faint magnitudes may result in statistically larger photometric errors, or a smaller fraction of blue objects in our samples, depending on the selection criteria. All these effects make the accuracy of the photometric redshifts generally non-uniform in type.

SDSS photometric redshifts (photo-z) have been used in different studies related to distances, clustering properties, evolution of galaxies, weak lensing, cosmological tests (Budavari *et al.* 2003; Sheldon *et al.* 2004; Stanford *et al.* 2004; Riess *et al.* 2001). The

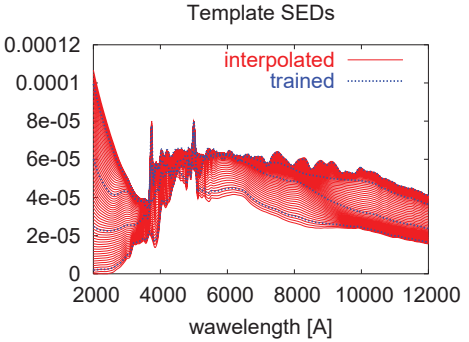

Figure 5.1: Repaired CWW spectra (thick blue line) and interpolated template set (thin red lines) used in SDSS photo-z.

reliability of the photometric redshifts is essential for scientific applications, therefore it is worth putting efforts into the tuning of this method. In Section 5.3 we overview several improvements we made in order to achieve a more accurate redshift estimator, especially for blue and faint galaxies. These changes have been implemented in SDSS photo-z since DR4.

5.2 Template fitting with the Bruzual-Charlot model galaxy spectra

5.2.1 Why to use models

When the photometric redshift estimation program searches for the best match between the measured and simulated photometry, usually a fine redshift grid is used, but only few spectral templates. This limits the precision of the redshift estimation. In order to reduce this problem, the present version of the SDSS photo-z code interpolates between template spectra when guessing the spectral type. We attempt to find a more physically motivated approach of templating and create a template library from model galaxy spectra. The idea of using galaxy models for photometric redshift estimation is not new. Koo (1985) used synthetic spectra to trace iso-redshift lines in color-shape diagrams and estimated the redshifts for a hundred of faint galaxies. Recent surveys use models for photo-z too, for example COMBO-17 (Wolf *et al.* 2001) uses a set of PEGASE (Fioc & Rocca-Volmerange

1997) templates.

There are several reasons that make the usage of model SED's as photo-z templates desirable. Model spectra have a physically defined set of parameters, like age, star formation rate, metallicity, dust content, etc. These can help us with the physical interpretation of the photometric data. When fitted to magnitudes or colors (as it is in the photo-z procedure), besides the estimated redshift we get back the parameters of the best fitting model. Ideally, this would give us a guess of the underlying parameters of the galaxy – its age, star formation history, chemical composition. Another advantage of the spectral models is their large wavelength coverage. This is an essential requirement for extrapolation, i.e. if we want redshift estimation to work over redshift ranges that are larger than that of our training/test set. When extrapolating to larger redshifts, one must be careful and keep in mind the possible evolution of spectral types. In the recent years it has become clear that the color-magnitude distribution of both red sequence and blue cloud galaxies evolves in time, they redden (and fade) towards lower redshifts. The usage of a locally selected empirical template-set not covering the full range of the populations of the earlier evolutionary stages present at larger redshifts may give misleading results. In these regimes a physically well-funded set of models, if used with precaution, could be a powerful tool.

We studied the suitability of the Bruzual-Charlot galaxy isochrone synthesis spectral evolution models (Bruzual & Charlot 2001) for being used as spectral templates in SDSS photo-z. Our aim was to find a subset of model spectra that is optimal for this purpose. The finer the type-resolution is, it means, the more spectral templates of different types are used, the more exact estimates one may get. Naturally, the number of used templates is limited by the computational time of the photometric redshift estimation. We have to find a set of models that is not too large but at the same time gives a good enough resolution and cover of all physical spectral types. It is important to keep the template set free from unphysical spectra, as they could easily mislead the estimation. This requirement can be fulfilled if we use the observations as our starting point and require our selected templates to be consistent with them.

5.2.2 Synthetic spectra matching spectroscopic observations

SDSS composite galaxy spectra

To proceed from a small number template set to a better spectral type resolution we created composite spectra from about 20 000 galaxy spectra measured by the SDSS as

follows. We devided the 4-dimensional color-space (u–g, g–r, r–i, i–z) into 64 cells which roughly correspond to galaxies of a common spectral class (even if colors are influenced by redshift as well). In each cell, using the redshift values determined by spectroscopy, we shifted the spectra back to rest-frame, normalized them to a common amplitude and computed the median value at each wavelength. In this way we can hope to get a fair resolution spectral library which covers most of the galaxy spectral types.

However, these composite spectra cannot be used directly as templates for redshift estimation. The reason is that they do not cover the whole possible wavelength range of the investigated objects. Their redshifts vary in a large range and the spectrogram measures different rest-frame wavelength ranges for the objects at different redshifts. Since redshift correlates with color, the redshift coverage of the different color bins will systematically deviate from each other. As extrapolation without any extra physical input is not reliable, we would like to combine the benefits of our results obtained from observations and galaxy models described with several physical parameters.

Isochrone synthesis spectral evolution

The Bruzual-Charlot model uses isochrones (evolutionary stages at a given time) of an instantaneous-burst stellar populations to calculate spectra of galaxies with given star formation histories (Bruzual & Charlot 1993). The spectrum of a composite stellar population with a more complicated star formation history is calculated by adding up the SED's of bursts that happened at different times in the past and evolved until their present age. The isochrones of the instantaneous-burst are interpolated from a refined set of stellar evolutionary tracks, so the distribution of stars of various masses and ages evolves smoothly in time. The evolutionary tracks of stars of different masses and metallicities are constructed from measured stellar spectra when available or are supplied by model spectra at some evolutionary stages.

Using these libraries one can compute the evolution of the spectral energy distribution $f_\lambda(t)$ of an instantaneous-burst stellar population. The spectrum $F_\lambda(t)$ of a stellar population with an arbitrary star formation rate (SFR) $\Psi(t)$ can be obtained using the convolution integral

$$F_\lambda(t) = \int_0^t \Psi(t-t') f_\lambda(t') dt'. \tag{5.1}$$

As input for the galaxy model, we have the freedom of selecting an initial mass function (IMF) and metallicity, as well as to define an arbitrary star formation rate.

Photo-z with composite-like synthetic templates

We performed our analyses on a simple 2-dimensional subspace of models using stellar populations with solar metallicity and the Kroupa (2001) universal IMF. We constrained the SFR to a one-parameter exponentially decreasing function of time t,

$$\Psi(t) \propto e^{-t/\tau} \tag{5.2}$$

. A set of synthetic spectra was generated on a 20×20 grid in both galaxy age t and the parameter τ, each ranging from 1 Gyr to 20 Gyr. We selected the synthetic spectra most similar to our SDSS composites by χ^2 minimization.

As can be seen in Figure 5.2, even with this simplistic model assumptions we can approxiate our composites representing the typical galaxy spectra with a remarkable precision.

In accordance with our expectations, bluer spiral galaxies are best fit by models with recent star formation, it means their age is comparable to the characteristic scale τ of star formation. Spectra of red elliptical galaxies do not show features implying presence of young stars, they mostly host old stellar populations. In our exponentially decreasing SFR model this can be achieved by small value of τ compared to the age.

An apparent feature of the model fitting is that blue galaxies, being young, are not very sensitive to the value of τ. This results in a degeneracy in χ^2. Sometimes it produces multiple minima which differ only slightly in χ^2 but belong to completely different physical parameters. This is illustrated in Figure 5.3 which shows the χ^2 surface of this minimization problem in the case of a red and a blue spectrum. In the case of the blue SED (bottom panel) there is an apparent strong degeneracy in τ, with two minima within the allowed parameter range. The two best fits are shown overplotted on the bottom panel of Figure 5.2, they are practically indistinguishable. Consequently, the star formation decay parameter remains unconstrained from such a fit. There is even a larger number of such degeneracies in the spectra if more model parameters are used. Age, metallicity and dust have a very similar effect, all of them make galaxies redder. Even though this ambiguity has no influence on the template library we construct, we have to pay attention to this effect when trying to identify the most relevant physical parameter combinations.

Figure 5.4 shows the models most preferred by the different main spectral types. Models with age >15 Gyr clearly cannot be physical. Also fits with minimal χ^2 at the border of the parameter region are not reliable. We might have received these χ^2 minimizing parameters because of our constraints to the model.

The fact that in some cases our composite spectra cannot be fit by physically reason-

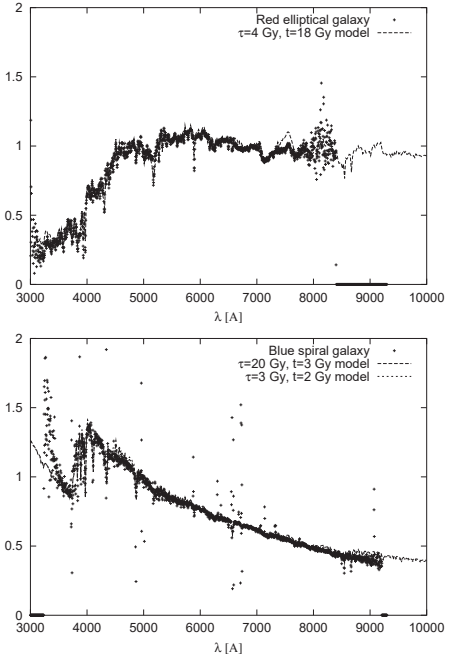

Figure 5.2: Red elliptical (top) and blue spiral (bottom) composite spectrum with the best fitting synthetic spectra. Illustration of degeneracy: the two best fitting models on the right panel, also having very different parameters, are practically indistinguishable.

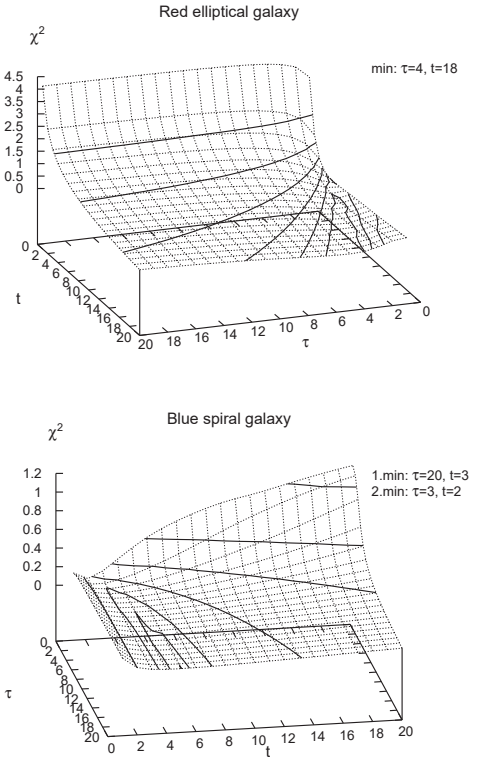

Figure 5.3: χ^2 of the model fit to a red (upper panel) and blue (lower panel) composite spectrum. The blue SED has two minima within the minimization region, the constant χ^2 contours are nearly parallel to the τ axis.

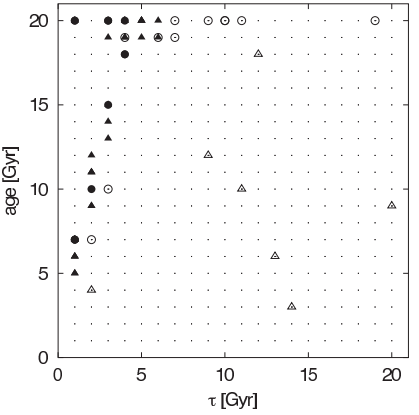

Figure 5.4: Model parameters best describing our SDSS composites. Galaxy types from bluest to the reddest: open triangle, open circle, solid triangle, solid circle.

able models implies that we have to broaden our models. First of all, metallicity should be treated as a free parameter as it has a large impact on color.

However, if we construct a library from the most preferred synthetic spectra and use them as templates for photometric redshift estimation, we can get a relatively good agreement with the actual redshift values determined by spectroscopy. Figure 5.5 shows the photometric redshifts versus the spectroscopic values in the case of two different methods.

In the first case we used a set of four optimized templates generated by template reconstruction described in Budavari *et al.* (2000) and Budavari *et al.* (2001). In the second case we used 40 synthetic spectra most similar to the composites. The better resolution apparently reduces the error of estimation in some redshift ranges. However, there appear to be systematic deviations in the photometric redshift relation with the BC templates. The galaxy redshifts tend to be systematically underestimated, with this effect becoming more pronounced as a function of redshift out to $z = 0.3$.

From this simple test, it has become clear that in spite of the selection of model templates based on observations, there remain significant systematic deviations. They migth be induced by the insufficient coverage of spectral types or the photometric calibration (or both). Note that the selection was made using spectra, while redshift estimation uses photometric data. The connection between them is never straightforward, partly because

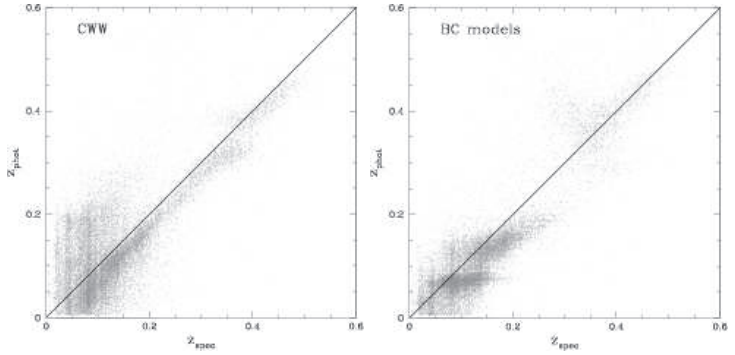

Figure 5.5: Comparing photometric redshifts to actual values from photometric redshift estimation made with four optimized templates (left) and with the 40 best fitting model spectra (right).

the problems in calibration, partly because the aperture bias (the spectroscopy only samples the central 3 arcsec of the galaxy, which is normaly redder than the integrated color measured by photometry).

We conclude that the studied model family is a step forward towards a physically motivated fine spectral resolution photo-z, but is clearly still not an optimal choice for redshift estimation. An extension of the model family, as well as a more photometrically driven approach might lead to better precision of redshift estimation.

5.2.3 Optimizing for photo-z error

In the prevoius section we showed that the matching of our model family with the measured spectra does not yield an unbiased redshift estimator as it induces severe systematic errors. As a next step, we choose a direct approach and try to select model spectral templates for photo-z by photo-z itself. We are looking for a subset of the BC01 templates which minimizes the photo-z error in a training set of galaxies. We also extend the allowed region of the considered models to aim a better type completeness.

This approach is closer to empirical, so we have to make sure we avoid its handicaps while making profit from both the possibility of calibration in the empirical methods and the "universality", ability to extrapolate in template fitting. Having a large set of free parameters there is a danger of "overtraining", i.e. selecting templates that give a

Z	0.0001	0.0004	0.004	0.008	0.02	0.05	0.10
τ [Gyr]	0.01	0.14	0.32	0.54	0.80	1.13	1.54
	2.05	2.71	3.61	4.88	6.85	10.43	19.44
t [Gyr]	0.01	0.28	0.74	1.36	2.17	3.20	4.51
	6.17	8.31	11.09	14.79	19.85		

Table 5.1: Metallicity (Z), τ and age (t) parameters of the model SED's on the grid

nearly perfect match for the training set but are completely unphysical, which makes extrapolation to a different set of galaxies unreliable. We can avoid this by testing the results of the minimization on an independent set of galaxies.

The model family

As suggested by the results of the previous check, we extended the model family and set the metallicity parameter Z free for the further studies. We generated a set of synthetic spectra on a 3-dimensional grid (Z, τ, t). Each model is characterized by its metallicity Z, star formation time scale τ and age t. The metallicities are the seven distinct values available in the BC01 library. They range from 0.0001 to 0.1, Z=0.02 corresponding to solar metallicity. (See Table 5.2.3 for a complete list.) Starting out from the solar metallicity equidistant grid We studied the χ^2 distances of the spectra in this subspace. The equal spacing of the model parameters results in large inequalities in the differences of spectra corresponding to the neighboring points. For example, a $\tau = 3$ Gyr spectrum changes rapidly between $t = 1$ and 2 Gyr but only slightly at later epochs, say between $t = 10$ and 11 Gyr. We changed the spacing of the model parameters τ and t in such a way that the χ^2 distances of the neighbouring points on the resampled grid is possibly constant. There was a constraint from the isochrone synthesis algorithm that made us to make a compromise: the points in τ had to be same for all t values in order to avoid an excessive increase in computational time. We decided to use an orthogonal grid. The values for τ and t are listed in Table 5.2.3 We have a $7 \times 14 \times 12$ grid with 1176 synthetic spectra.

Minimization with simulated annealing

We used the model set described above to choose a subset of spectra which yields the best photometric redshift estimator. As a trainig set we used a sample of 6000 galaxies with known redshifts and colors from the Eary Data Release (EDR) spectroscopic sample.

We simulated photometric redshift estimation of the trainig set galaxies using the model SED's as spectral templates. The rms-z minimization was carried out by simulated annealing. The algorithm also implements steps resembling the learning vector quantizatioin (LVQ) algorithm (Kohonen 1995) which is used in the SDSS photo-z template repair procedure in use (Budavari *et al.* 2000). The main idea is as follows: each object is classified by assigning it to the template which is closest to it according to a certain metric. In the present case this is the σ_z, the error of the photometric redshift estimator for the galaxy with that particular template. The redshift estimator of the best template is then considered to be the photo-z vale for that object. The cost function σ_z (summation made over the training sample) is then minimized by changing the configuration, i. e. the set of the used model spectra.

The minimization starts from an initial configuration (certain templates 'switched on', others 'off'), with an assigned 'temperature' T. The cost function (or 'energy') is evaluated. A new, slightly changed configuration is suggested and accepted if the new value of the cost function cf_2 is smaller than the original cf_1; if the new value is larger, than the new configuration is accepted with a probability $exp(-(cf_2 - cf_1)/T)$. . The temperature parameter is T reduced, the system anneals and converges to a final configuration with an (at least local) minimum of the cost function, where it has no longer enough 'energy' to change.

Our algorithm was constructed to be automated but at the same time to allow different interactive interventions. These enable for example driving the system out of local minima by increasing the temperature, fixing or setting free the number of templates, prescribing the allowed modes of changes in configuration (e.g. 'migration' to nearest neighbour gridpoints only).

Photo-z with the optimized synthetic spectra

First we searched for the optimal template set in the $Z = 0.02$ (solar) metallicity plain only. From this configuration we allowed the templates to migrate through the whole grid. By this optimization we have found a subset of 40 BC01 synthetic spectra that minimizes the cost function. The selected model parameters are shown in Figure 5.7. Again it is clear that some of the models (largest ages, metallicities) are unphysical, even though the model family has been significantly extended. If we believe that the models are correct, this fact suggests that the main problem with the spectrally selected template SED's is due to photometric calibration.

For the training set, the optimized template set yields an $\sigma_z = 0.0376$ (0.0330 if 3σ

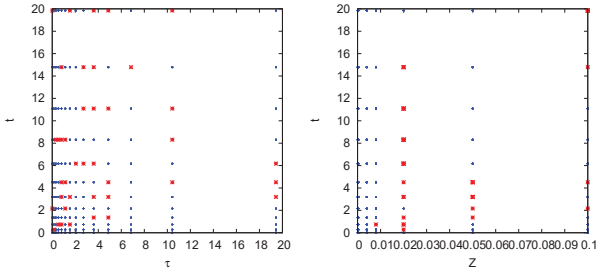

Figure 5.6: Distribution of the models minimizing the error of the photometric redshift estimator in (τ, t) (left) and (Z, t) (right).

outliers removed, in 98.1% of the sample). This is a significant quantitative improvement if compared with the results of the spectral selection (σ_z=0.051) presented in Section 5.2.2. The resulting $z_{\texttt{photo}} : z_{\texttt{spec}}$ diagram is plotted in Figure 5.7. It is apparent that the strong systematic deviations also disappeared. We have to make a check to rule out the possibility of overtraining. An independent set of 6000 galaxies was used to test the selected templates. We got an $\sigma_z = 0.0348$ (0.0308 98.2%) which reassures us that the templates are generally applicable in photo-z as far as the statistical distribution of the estimated galaxies (magnitude, color) is similar to the SDSS spectroscopic sample.

Conclusions and outlook

We can conclude that it is possible to select a template set of synthetic SED's that can yield a photometric redshift estimate of reasonable precision. The optimized model template fitting described above yields an accuracy of the photometric redshifts which is superior to the hybrid method used in the data release version of SDSS photo-z. However, the estimation of the physical parameters as age, metallicity, star formation history in a naive way by reading the parameters of the best fitting model is not feasible for several reasons. The models matching photometry do not necessarily have parameters lying in the physical range. This might be due to discrepancies between photometric and spectroscopic calibration, or because the models are insufficient. The models can be improved in several ways, for example adding emission lines or including dust to models. As described in Section 4.5, the BC spectra with empirical emission lines is now available. The detailed photo-z oriented study of these templates is a challenge to be met. However, even in the case of a perfect calibration and perfect spectral models the parameter space

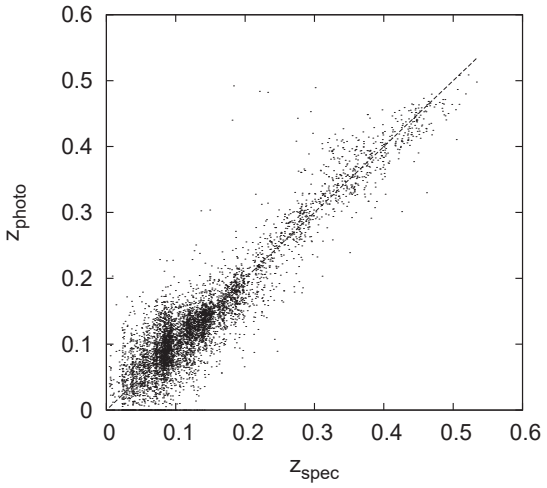

Figure 5.7: Photo-z versus spectro z with the selected BC templates.

is strongly degenerate. With broad-band photometry like that of SDSS the degeneracy is even stronger. This makes the disentangling of the effect of age, metallicity and dust, or star formation history and age hard or even impossible. These limitations originate on the one hand in the nature of the stellar populations, on the other hand in the survey itself. Although we cannot change any of these conditions, we can study their impact on information carried by our observatons. A detailed study of degeneracies can tell us what is the maximal information that can be retrieved from photo-z. We can for example study the Fischer information matrix (see e.g. Lahav (2000)) specifically for our photo-z algorithm and thus get a map of the errors and the correlations of the errors of the estimated parameters in the parameter space. Thus we can isolate the most important combinations of parameters that lead the changes in the observables. If we restrain ourselves to estimating of these parameters, using a subset of models not wider than the degeneracies make reasonable, we can approach a well-funded physical interpretation of our data.

5.3 Tuning the photometric redshifts of SDSS

5.3.1 Why we need new reference sets

In the previous section we discussed the possible applications of synthetic galaxy SED's as template spectra in photometric redshift estimation. From certain aspects, this would be an ideal choice. This approach could enable us to mine from our measurements the maximal amount of information about the underlying physical properties of the observed objects. However, to make maximal use of the advantages of this method requires further developments in both the applied models and the method of use. Until this is done, we use empirical spectra adapted to fit the SDSS photometry in the data releases. As described in Section 5.1, the templates are derived from CWW spectra, repaired on a training set of galaxies from the SDSS spectroscopic sample and interpolated to form a continuous type. The iterative repair of templates requires a training set of spectroscopically observed galaxies of all possible types.

As these photometric redshifts have been widely used for science purposes, it is essential to test the performance of the estimation. After the first data releases (EDR, DR1) there was a strong demand for such tests and also for an improvement of the precision in the furhter data releases that were to contain much larger samples of galaxies. In this section we present a test of the DR1 photo-z version used in the first SDSS data releases up to DR3. The analysis helped us in developing a new version which has been implemented since DR4 until the most recent release DR6.

Both testing and calibration of the photometric redshift relation requires reference samples with known redshifts and SDSS photometry. A training set has to explore the same range of apparent magnitudes and colors as the objects for which photometric redshifts will eventually be derived.

As described in Section 1.2, the SDSS spectroscopic galaxy sample consists of two subsets. The main sample with a magnitude limit $r < 17.8$ contains galaxies of all types, with redshifts $z < 0.3$. The SDSS LRG sample obtains spectra for luminous red galaxies with faint apparent magnitudes up to $r < 19.5$, $z < 0.5$. The redshift histogram is shown in Fig. 5.8. Fig.5.9 shows the detailed redshift-type distribution. One can see that all types are only present at low redshifts. The DR1 version of photo-z was calibrated for this dataset (we refer to it as DR sample hereafter).

Photometric versus spectroscopic redshift of the DR sample is plotted in Fig. 5.3.1. The dispersion in the photometric redshift relation is σ_z= 0.045 (0.035, if 3σ outliers rejected). But at redshifts larger than $z > 0.3$ the sample practically consists only of red

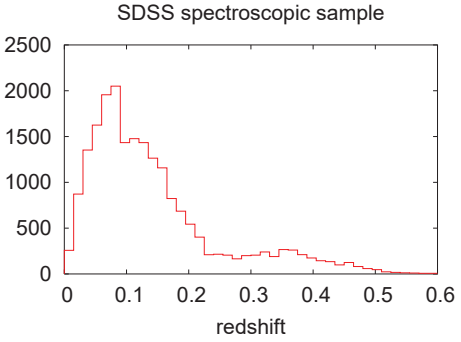

Figure 5.8: Redshift distribution of the SDSS spectroscopic sample

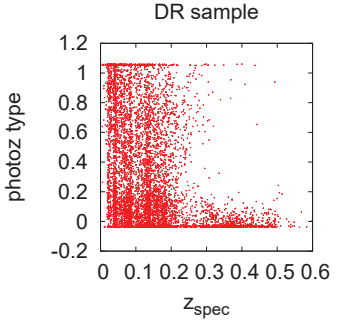

Figure 5.9: Distribution of the DR sample in redshift and spectral type, shown with a subset of ~ 8000 galaxies.

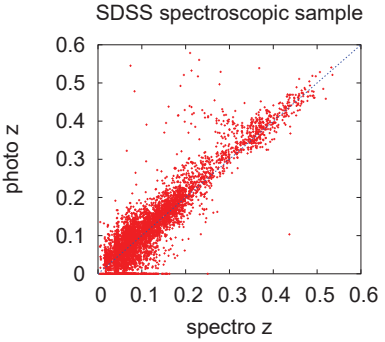

Figure 5.10: Photometric versus spectroscopic redshift for the EDR spectroscopic galaxy sample.

galaxies of the LRG sample. Photometric redshifts of this relatively uniform population (Eisenstein *et al.* 2003) are fairly robust (e.g., Padmanabhan *et al.* (2005)). The DR sample does not have a corresponding sample of faint blue galaxies for the calibration of photometric redshifts. Therefore, a series of spectroscopic plates was designed to obtain redshifts for the blue end of the galaxy color distribution at the faint end (photo-z Plates), aiming to targeting non-red objects above $z = 0.3$.

Unfortunately, it is not easy to obtain samples with similar apparent magnitude and redshift properties for both early and late types. For reasons originating in their different evolution (see Chapter 3), the cosmic distribution of red and blue populations are different (Baldry *et al.* 2004). The blue luminosity function has a sharp faint-end cutoff at $M_r \approx -22$, whereas the most luminous red galaxies extend out to $M_r \approx -23$. Thus, with a uniform apparent magnitude cut we would probe a lower redshift range for blue galaxies than for red types. Or, if aiming the same redshift limit, the apparent magnitudes of the blue galaxies will be in general fainter.

Another important aspect is the depth of the test sets in redshift and apparent magnitude. The photometric sample contains a large fraction of objects that are much fainter and at higher redshifts than the SDSS spectroscopic survey measures. It is essential to have the quality of the photo-z under control even at these these faint and distant ranges, to ensure a correct understanding of photometric redshifts and their errors and consequently to avoid misinterpretation. For such tests we need data from deep external redshift surveys. The survey strategy of COMBO-17 survey proved suitable to probe

SDSS photometric redshifts. In the next subsections we give a description of the new reference sets we used to test and improve our photo-z.

5.3.2 Photo-z plates in the Southern Survey

Special spectroscopic plates were targeted in the SDSS Southern Survey for photometric redshift calibration purposes (pzPlate sample). The aim was to obtain spectra of non-red galaxies above the normal spectroscopic sample ($r > 17.8$) with redshifts and r-band magnitudes ranging up to the upper limit of the LRG sample.

The targets were selected as follows. Using the photo-z code, photometric redshifts and types were calculated for galaxies in the target area. To minimize the overlap with the main and LRG sample, objects bluer than photo-z type 5 having photometric redshifts $z_{\texttt{photo}} > 0.3$ were separated. (Note, that as already indicated in Sec. 5.2, there are two versions of photo-z type: one ranging from 0 to 55, the other between $\sim 0 - 1$, small values corresponding to red colors in both. Here the former was used.) However, the photometric redshift estimation is a complex algorithm, so it is hard to give a concise description of target selection. Furthermore the versions change hence the reproducibility of such a selection is hard to guarantee. Therefore a different approach was choosen: to find a tranlsation of the above criteria in the terms of colors and then to make the apropriate selection in color space. Plotting the color of galaxies in our high redshift blue sample and the rest (see Fig. 5.11), one can see that the two sets can be relatively well separated by two simple color cuts. One to select higher redshifts:

$$g - r > 0.40 + 0.6 * (u - g), \tag{5.3}$$

and one to separate the targets from early types:

$$g - r < 1.7 - 0.1 * (u - g). \tag{5.4}$$

Some more cuts were added to remove outliers:

$$-0.5 < u - g < 3.0$$
$$0 < g - r < 1.8$$
$$-0.5 < r - i < 1.5$$
$$-1 < i - z < 1.5$$
$$18 < u < 24$$
$$18 < g < 21.5$$

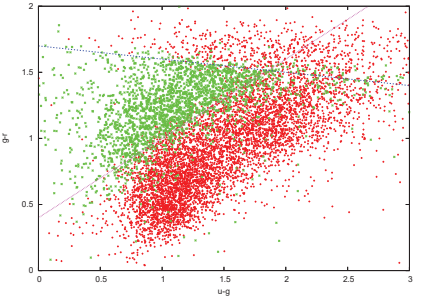

Figure 5.11: The color cuts applied in the photometric redshift plate selection process. The green labels show the distribution of galaxies in SDSS color space for blue spectral type above redshift of 0.3. The rest of the galaxies in the target area is plotted in red. The blue dashed line indicates the applied type cut, the magenta dotted line the redshift separator. The two lines separate the two sets well.

$$17.8 < r_{petro} \;\; ; \;\; r < 19.5$$
$$16.5 < i < 20.5$$
$$16 < z < 20$$
$$\sigma_u < 0.6; \; \sigma_{g,r,i,z} < 0.25,$$

σ denoting the photometric errors, r_{petro} the r band Petrosian magnitude. All criteria are understood in dereddened model magnitudes, as photo-z works best with them. The number of objects selected with the above cuts naturally did not match the number of available fibers on each spectroscopic plate exactly. To resolve this, a soft boundary was applied on the redshift side, namely using a random criterium $\exp(c[(g-r)-0.40-0.6*(u-g)]) > rnd$ with rnd a random number between 0 and 1 and adjusting the value of c to match the number of fibers. With this targeting strategy the so-called pzPlate spectroscopic sample has been recorded.

The sample consists of over 7000 galaxies of the following spectroscopic plates of the SDSS Southern Survey: 0669, 0670, 0671, 0672, 0807, 0808, 0809, 0810, 1629, 1632, 1633, 1635, 1636. Besides galaxies, a small fraction of other objects were also caught by the targeting: QSO, stars, and several unknown type objects. The data contain ∼400 objects with double spectroscopy.

We used the observed pzPlate data, constructed a unique sample and matched the

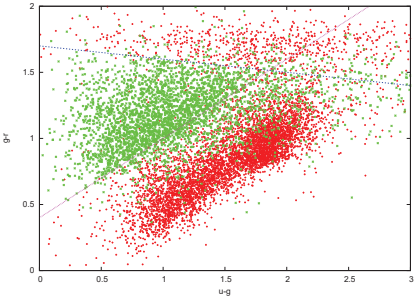

Figure 5.12: The observed pzPlate galaxies (green) and a random subset of the DR sample from the target area (red), plotted with DR6 colors. The lines are same as in Fig. 5.11. The colors of the pzPlate galaxies extend towards blue (smaller g-r values) from the intended lower edge cut.

objects therein with the photometric catalogue.

Some of the characteristics of the sample are presented below.

The observed galaxies have a color distribution comparable to the selection cuts (see Fig.5.12). One can see that the pzPlate galaxies fill the proposed region. They extend under the lower edge cut, towards smaller $g - r$ colors (lower redshifts). This is because of the soft boundary in the cut on the redshift side.

To check how succesful the targeting was, we plotted the redshift - spectral type distribution of the pzPlate galaxies in Fig. 5.13. If compared with 5.9, one can see that the pzPlate sample is indeed complementary to the DR dataset. The effect of the soft redshift bias can be seen here explicitly: a large fraction of $z < 0.3$ galaxies is present.

We compiled a reference set from the DR and pzPlate datasets. This $r < 19.5$ sample was constructed to be closely uniform in redshift and photo-z type up to $z = 0.4$. To achieve this, we found that a combination ratio 2/3 DR sample, 1/3 pzPlate is the most effective. When using reference sets for calibration, it is also important to control the quality of the spectroscopic data. The SDSS spectroscopic catalog has fields `zStatus` and `zWarning` which can indicate if there are possible problems with the spectroscopy. We reject a number of objects from the calibration set based on this information. We only allow the following values of `zStatus` and `zWarning`:

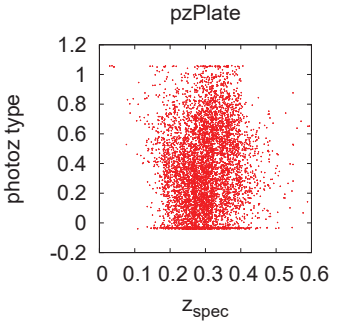

Figure 5.13: Distribution of the pzPlate sample in redshift and spectral type.

```
--------
zStatus
--------
name            value   description
XCORR_EMLINE    3       Redshift determined from cross-correlation
                        and emz are consistent.
XCORR_HIC       4       Redshift determined from x-corr with high
                        confidence.
EMLINE_XCORR    6       Redshift from emz plus consistent xcorr
                        redshift measurement.
EMLINE_HIC      7       Redshift determined from em-lines with
                        high confidence.

--------
zWarning
--------
name            value           description
OK              0x00000000      No warnings.
NOT_QSO         0x00000020      Classification does not match quasar target.
NOT_STAR        0x00000040      Classification does not match star target.
HIZ             0x00000800      Redshift is high.
```

Two different combined samples with the same compositions were used for training

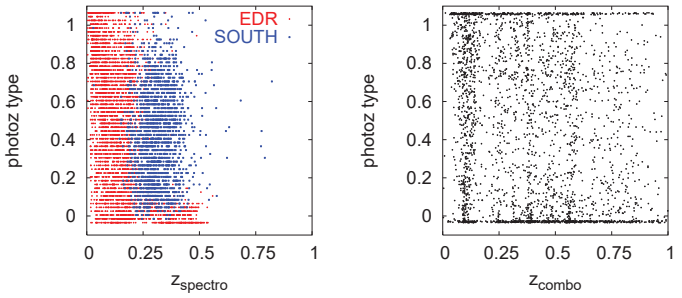

Figure 5.14: Left: Distribution of the combined reference set constructed DR sample (red) and the Southern Survey pzPlates (blue, labeled 'SOUTH') in redshift and photo-z type (the smaller the redder). Right: the same shown for the identified COMBO-17 galaxies.

and testing, each containing about 8000 galaxies. As their properties are practically same, in the following description we do not distinguish between them and refer to either of them as the combined sample, unless explicitly stated otherwise.

The distribution of the combined sample is shown on the left panel of Fig. 5.14. For the reason of better comparability, the right panel shows the same distribution for the COMBO-17 sample, which will be discussed in the next Section 5.3.3. The pzPlate galaxies filled in the previously missing blue and faint regions. The combined sample is nearly homogeneous in type and redshift for $z < 0.4$. The left panel of 5.15 shows the redshift-magnitude distribution. The newly targetted objects extend the reference set to larger redshifts and magnitudes.

5.3.3 COMBO-17

COMBO-17 uses a 17-band photometry to determine photometric redshifts. Redshifts go up to $z \sim 1$, r-band magnitudes up to ~ 23. In this study we compared SDSS DR1 photo-z with COMBO redshifts. SDSS DR1 overlaps with the S11 field of COMBO-17. The overlaping region is within the range ra$>$ 175.4 and ra$<$ 176 and dec$>$ -2 and dec$<$ -1.45. The objects were identified as follows. To each galaxy of the SDSS sample was selected the closest object of the S11 database with

1. consistent position - within a circle of r = 0.0007 degree (2.52 arcsec) and

2. consistent colors. This condition was checked using an empirical relation between $(R-B)_{\texttt{COMBO}} : (r-g)_{\texttt{SDSS}}$ and $(I-R)_{\texttt{COMBO}} : (i-r)_{\texttt{SDSS}}$ colors, taking into account the SDSS

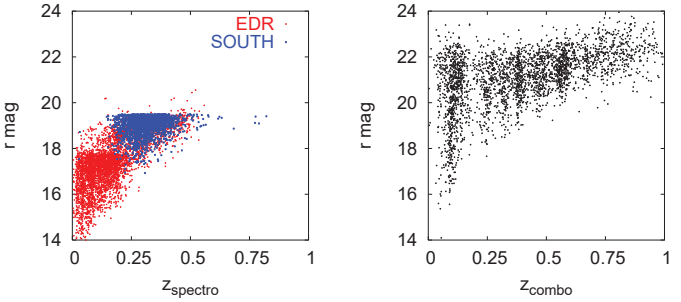

Figure 5.15: Left: Distribution of the combined sample in redshift and r-band magnitude. Red: DR sample, blue: pzPlates. Right: the same shown for the identified COMBO-17galaxies.

photometric errors and allowing 0.5 magnitude deviation from the empirical relation. With these constraints an estimated upper limit on the probability of misidentification is $\sim$3%.

Approximately 3000 objects were identified. Their redshift, spectral type and magnitude distributions are shown in the right panels of Figures 5.14, 5.15. (For comparison see the left panels where the same distributions are plotted for the sample compiled from the DR sample and PzPlate set.) There are several clustering features in COMBO redshift. The condensations near z=0.4, 0.6 are COMBO artefacts. At z=0.1 which is a real cluster the clustering seems to be type-dependent. This is an artifact due to the type dependent error of the COMBO-17 redshift estimator. The COMBO-17 reference set is deeper about a factor of two in redshift and 3 magnitudes in r-band than the DR and pzPlate samples together. The overlap of the two samples is small, hence the COMBO-17 sample probes photo-z in a different region.

Fig. 5.16 shows type and redshift histograms of the COMBO-17sample with different magnitude limits. The sample has a good type coverage at all magnitudes. These magnitudes are much fainter than the typical DR magnitudes. LRG's having with a limit $r < 19.5$ are in the brightest bin, and below the magnitude limit of the main spectroscopic sample we found only 80 galaxies. In redshift, the sample seems to be more complete at faint magnitudes.

The deviation of SDSS photometric redshift $z_{\texttt{photo}}$ from the COMBO redshift $z_{\texttt{COMBO}}$ correlates with magnitude and other quantities affected by magnitude, e.g. photometric error, combo-z error, color difference of the two objects supposed to be the same, the

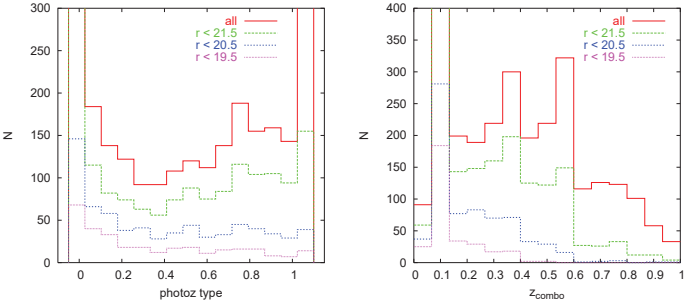

Figure 5.16: Spectral type (left) and redshift (right) histograms for different magnitude cuts. $r < 19.5$ corresponds to the magnitude limit of the DR+pzPlate sample. At $r < 17.8$ where the photo-z was calibrated (only for red LRGs above this limit) there are only 80 objects in the sample (not shown).

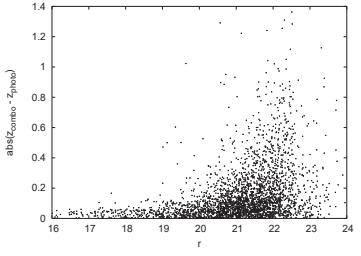

Figure 5.17: The absolute difference of $z_{\texttt{photo}}$ and $z_{\texttt{COMBO}}$.

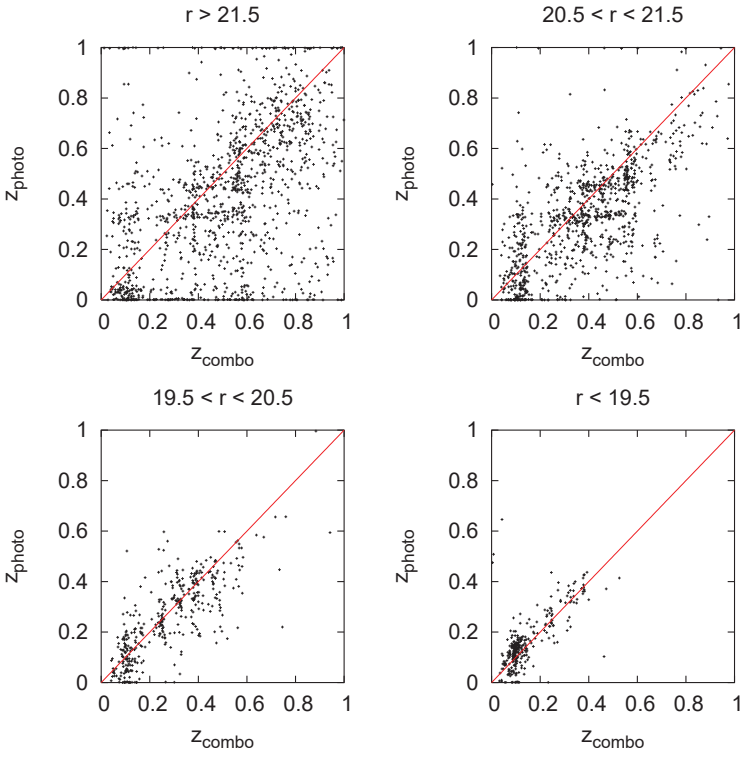

Figure 5.18: SDSS versus COMBO-17 photometric redshifts in r magnitude bins

probability that an object is misidentified, etc. However, the strongest correlation can be observed with the SDSS r magnitude. Fig. 5.19 shows the difference of the two redshifts, $|z_{\texttt{photo}} - z_{\texttt{COMBO}}|$. The error increases strongly with r.

We plotted the redshift relation in four r magnitude bins in Fig. 5.18. A striking feature is the almost complete lack of signal in the faintest bin. The usual rms scatter of the $z_{\texttt{photo}}$ - $z_{\texttt{COMBO}}$ pairs for the whole sample as well as for the four magnitude ranges are given in Table 5.2. The differences are dominated by the actual errors of SDSS photo-z, as the mean COMBO redshift error is less than 0.03 for the sample and <0.04 even for the faintest bin. However, since some individual objects whith $R > 23$ (or $r > 21.5$) have a large redshift error ($\Delta z_{\texttt{COMBO}} > 0.05$), as well as because of the possibility

sample	rms	N	rms_iter	p_iter
all	0.2433	2895	0.1636	93.5%
r:[21.5, :]	0.3212	1233	0.2874	97.9%
r:[20.5, 21.5]	0.1957	954	0.1503	96.2%
r:[19.5, 20.5]	0.1191	394	0.0850	95.4%
r:[:, 19.5]	0.0771	311	0.0494	97.4%

Table 5.2: Rms errors for the COMBO-17 sample and for four magnitude bins. (rms_iter: 3 sigma outliers removed in an iterative way, p_iter: fraction of objects left after the iteration)

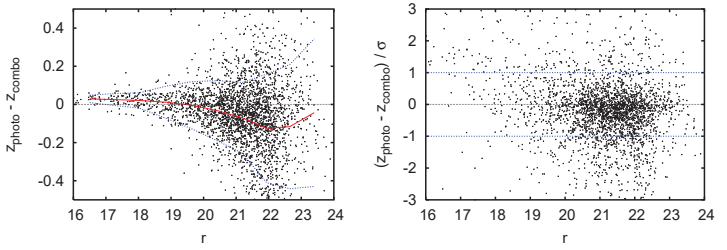

Figure 5.19: Left: magnitude dependence of the $\Delta z = z_{\texttt{photo}} - z_{\texttt{COMBO}}$. Red line: mean, blue lines: mean $\pm\sigma$ calculated in bins. Right: redshift difference normalized by σ.

of misidentification, $r > 21.5$ data should not be used for photo-z calibration.

Figure 5.19 helps to summarize the magnitude dependence of the deviations and their relation to the nominal photo-z error. On the left panel we plotted the difference of the two redshifts. The right panel shows the above quantity normalized by the resultant error: $(z_{\texttt{photo}} - z_{\texttt{COMBO}})/\sigma$, with $(\sigma = \sqrt{\sigma^2_{\texttt{COMBO}} + \sigma^2_{\texttt{SDSS}}})$. At small magnitudes there is a systematic overestimation by about 0.03 (see left panel) inconsistent with the nominal errors (see right panel). An indication off this effect can also be found if SDSS spectroscopic redshifts are used. For fainter galaxies underestimation is typical and as expected, the scatter grows with magnitude but photo-z error estimation seems to work better.

The performance of the SDSS photo-z is different for different spectral types. More than from the statistics is the effect of type apparent from the $z_{\texttt{photo}}$:$z_{\texttt{COMBO}}$ plots (see Figure 5.20) as many of the type dependent deviations are systematic.

Most outliers and systematic errors can be found in the extreme types, particularly in the bluest bin. An improvement of the latest type templates (e.g. including/improving

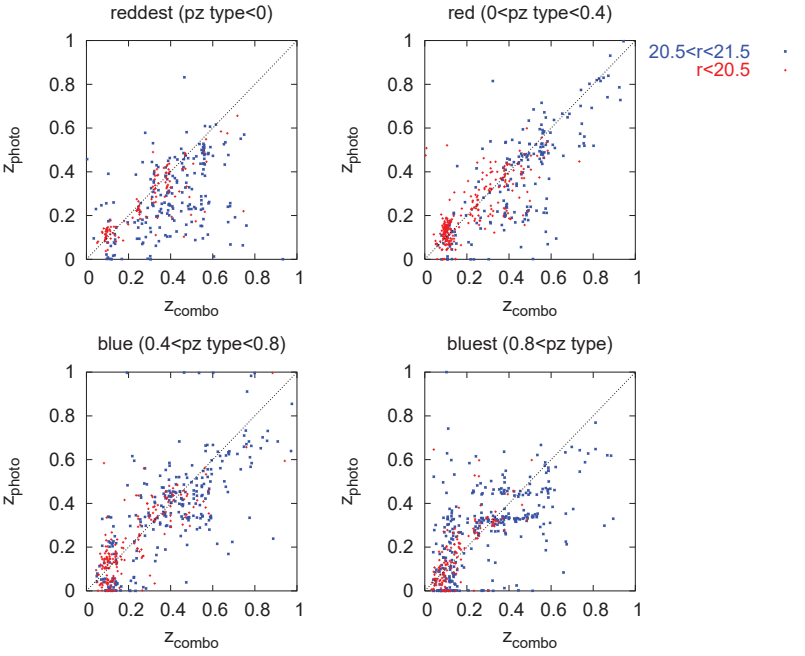

Figure 5.20: The redshift relation in four different type bins. Magnitude ranges are color coded: $20.5 < r < 21.5$ blue, $r < 20.5$ red.

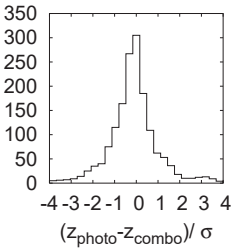

Figure 5.21: Normalized variance histogram for $r < 21.5$ objects

bursting templates, emission lines) may be useful.

For intermediate type (0 < photo-z type< 0.8) galaxies with $r < 21.5$ photo-z works relatively well up to z=0.9-1. The $\sigma_z$values for this group are 0.1370 (0.1110 without 3σ outliers). This suggests that SDSS photometric redshifts may be useful for purposes like detemining abundances of massive red objects at high redshifts, weak lensing, and similar high redshift statistical studies.

There remains an important check: the quality of the redshift errors calculated by the photo-z algorithm. The distribution of the normalized errors $\Delta z/\sigma$ plotted in Figure 5.21 gives a check of SDSS photo-z error estimation. As histograms show, the distribution of the normalized deviation is non-Gaussian, asymmetric and affected by systematic errors. The small bump at $\Delta z/\sigma = 3$ are due to the slight overestimation at small magnitudes that was mentioned above.) However, on average 92% of the $r < 21.5$ objects are within 2σ (for the four type-bins of Figure 5.20: 87, 89, 96 and 95 precents respectively), which suggests the photo-z error estimation works well for most galaxies up to $r = 21.5$.

5.3.4 Recalibration

At that time when these developments were carried out, DR1 photo-z was the latest version of SDSS photometric redshifts. DR2 had a new version of photometry, but no new version of photo-z. Our efforts aimed to construct an improved version for the later data releases. We made use of the newly obtained, faint, blue pzPlate observations and used the combined (DR + pzPlate) sample defined in Section 5.3.2 for calibration of the hybrid photometric redshifts. We note that although `zStatus`=7 is in principle acceptable, we excluded these objects when training the templates. It only affects a small number of objects, they are usually faint blue galaxies with strong emission lines and low signal/noise continua. We also note that in the present (DR6) version there are almost

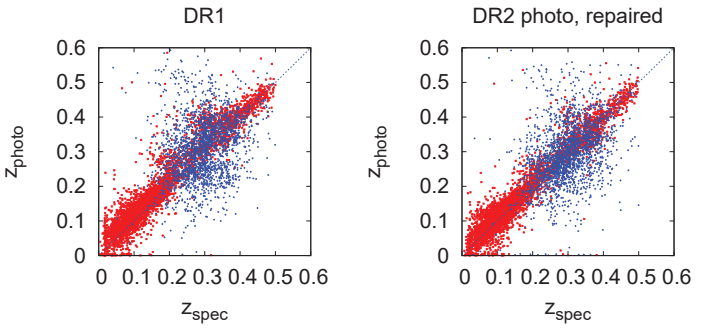

Figure 5.22: $z_{\texttt{photo}}$:$z_{\texttt{spec}}$ relation for the DR1 photo-z (left) and the recalibrated version (right). Red: DR sample, blue: pzPlates.

no objects having the allowed `zWarning` flags and `zStatus`=7 at the same time, this is probably the result of the updated spectroscopic pipeline.

The photometric vs. spectroscopic redshift relation of the original DR1 photo-z is plotted on the left panel of Fig. 5.22. The rms precision of the redshift estimation for the DR sample is $\sigma_z \sim 0.04$. For the combined sample $z = 0.5$ the error turned to be $\sigma_z = 0.08$ (0.054 if 3σ outliers removed). There are sever systematics in the pzPlate galaxies plotted with blue labels.

We performed template calibration described in Section 5.1 using the combined sample with the updated, DR2 photometry. The recalibrated photometric redshift relation is shown on the right panel of Fig. 5.22. The resulting error for the combined combined sample is $\sigma_z = 0.06(0.036)$, which means an improvement in precision by about 35% for $z < 0.5$, $r < 20$ galaxies with respect to the DR1 photo-z. The systematic deviations seem to be reduced too. A more detailed insight into the type dependence of the systematic deviations can be gained from Fig. 5.23 (note that photo-z type is small for red, large for blue galaxies). The results show that the template training suppressed the scatter in red galaxies near photo-z types ~ 0.2, as well as the overestimation in blue types at photo-z type > 0.2, thus smoothing out the type dependence.

Figure 5.24 shows the rms error σ_z averaged over r magnitude and photo-z type bins in three different photo-z versions. It demonstrates the effect of using DR2 photometry instead of DR1 and that of the template training. The left panel shows the variation of the photometric redshift with magnitude and type errors for the original, DR1 version of

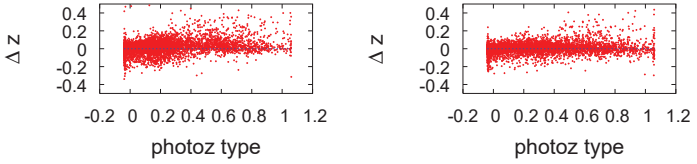

Figure 5.23: The redshift estimate error $\Delta z = z_{\texttt{photo}} - z_{\texttt{spec}}$ as a function of spectral type for the DR1 photo-z (left) and the recalibrated version (right).

photo-z. While the average error is less than 0.05 for the most bins of all types at $r < 18$, $\sigma_z \leq 0.1$ for blue galaxies with $r > 18$.

The new DR2 photometry itself improves the redshift estimate for the DR sample by about 10% (see middle panel of Fig. 5.24) but does not influence the results for the pzPlate galaxies not included in the original spectroscopic sample. For these objects the template repair plays an essential role. Due to the poor photometry quality of SDSS at $r > 20$ the training does not change the results for the COMBO-17 sample significantly, even when COMBO objects are included in the training set.

The most remarkable improvement is, however, the removal of some systematic errors and the reduction of the type and magnitude dependence of the quality of the photometric redshift estimation.

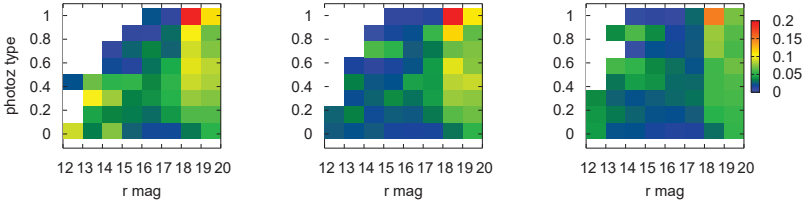

Figure 5.24: Magnitude and spectral type dependence of σ_z for the combined DR + pzPlate(South) spectroscopic sample (`photo-z type`: the smaller the redder). Left: DR1 photometric redshifts; Middle: DR2 photometry, old template set used; Right: photo-z after training.

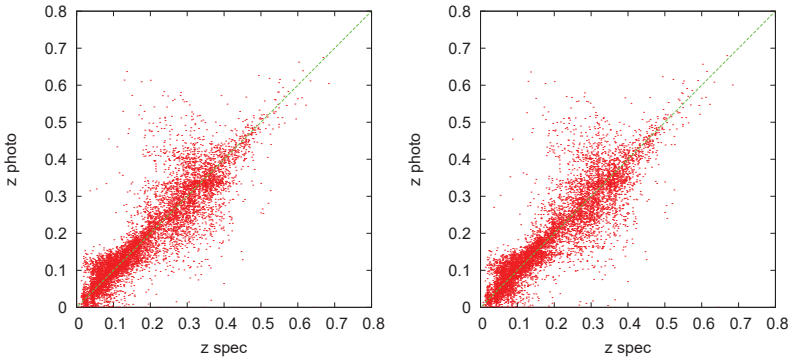

Figure 5.25: Left: photo-z with repaired CWW without lines, right: photo-z using CWW spectra repaired with added emission lines.

5.3.5 Emission line correction

In this subsection we present the application of the empirical relation between the emission lines and continuum spectra described in detail in Section 4.4 in photometric redshift estimation. The CWW spectra most widely used in photo-z are empirical spectra derived from a small number of local galaxies. Because of the insufficient statistics they cannot cover the entire range of galaxy populations up to the redshifts where photo-z is used. When compared to statistical studies of SDSS galaxies it turns out that they do not contain all emission lines that were found to have a possible strong impact on photometry, discussed in section 4.4 (See left panel of Figure 5.26).

In order to keep the advantage of being derived from galaxy spectra, and at the same time to be consistent with the SDSS emission line galaxy distribution, we modified the emission line pattern of the CWW spectra. We removed their original emission lines and added new lines based on the continuum shape, as derived in our statistical study of SDSS galaxies in section 4.5. Then we applied the usual template repair, using a training set of galaxies containing main sample+LRG and pzPlate galaxies in mixing ratio 2:1.

Figure 5.25 shows $z_{\texttt{photo}}$ versus $z_{\texttt{spec}}$ in the case of trained original CWW templates and if emission lines are added previously. Table 5.3.5 summarizes the changes in σ_z due to the addition of lines. As first line shows, there is a slight improvement in the overall σ_z. In agreement with our the expectations, the effect comes mainly from blue objects, the emission line correction does not influence the red galaxies. The type-dependence of

	t	original	added lines
all	0-55	0.0694 (0.0403 94.7%)	0.0680 (0.0398 94.4%)
red	0-20	0.0557 (0.0377 94.8%)	0.0556 (0.0375 94.8%)
blue	21-55	0.0876 (0.0440 94.2%)	0.0843 (0.0425 93.5%)
bluest	50-55	0.1512 (0.0324 88.2%)	0.1374 (0.0314 88.2%)

Table 5.3: rms z error of photo-z for different galaxy sets if CWW spectra without and with addition of emission lines are repaired and used as templates. The type parameter t goes from 0 to 55, 0 being the reddest, 55 the bluest type. The photo-z rms error is given if the original CWW templates are trained or if emission lines are added before the training. The values in round brackets denote the error with the 3σ outliers removed, together with the percentage of remaining objects.

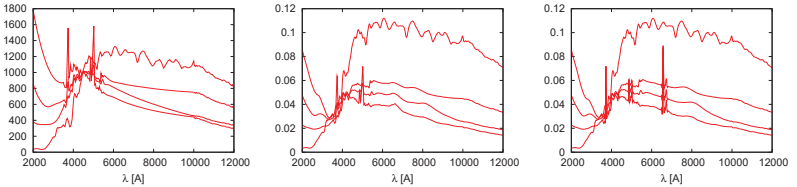

Figure 5.26: Left panel: CWW spectra. Middle: CWW spectra after template repair. Right: emission lines added from the continuum fit, then trained. The emission lines prevent forming of bumps near 6500 Å.

the improvement is strong. (The type parameter t goes from 0 to 55.) While there is no significant improvement in the red types ($t <= 20$), rms z decreases by $\approx 4\%$ in the blue types ($t > 20$). For the bluest objects ($t > 50$) the emission lines reduce rms z from 0.1512 to 0.1374, which is a 10% improvement, mainly due to outliers.

Besides the directly measurable effect on the photo-z precision, there is another important benefit. Emission lines consistent with the spectra of the measured objects make the training more stable. Left panel of 5.26 shows the original CWW scpectra. It is apparent that they miss the strong emission lines, especially Hα at 6565Å. However, the galaxies we measure do have these lines. When repairing templates to match better the photometry this makes the algorithm build bumps where the emission lines should be. Correcting for emission lines stabilizes the training by avoiding the generation of bumps.

5.3.6 Summary

Here we summarize the ingredients from which DR4 and later versions of SDSS photo-z was created:
- template fitting
- CWW spectra
- template repair
- 1-D continuum type
- reconstructed emission lines
- using DR+pzPlate sample.

But this is not the final stage. We have data with deeper, coadded photometry for a significant area of the sky that have to be used for photometric redshift estimation. We are just before the last data release of SDSS. Pan-STARRS survey is coming soon, with a promise of a good quality photometry. In the future, there are still challenges for both the developers and the users of the photometric redshift estimation.

I would like to thank my advisor, István Csabai and my senior colleagues Alex S. Szalay, Eric Bell and Tamás Budavári for many inspiring ideas and for their help in my research. I thank Tamás Budavári for the LaTeX template of his thesis which helped me a lot. I am grateful to my family – my husband, children, parents and parents-in-law. Without their support and patience this work would have never be done.
I also acknowledge support from the following grants: OTKA-T037548, OTKA-T047244, MSRC-2005-038, MRTN-CT-2004-503929, MTA97-OTKA049957-NSF, NKTH: RET14/2005, KCKHA005, Polányi.

Bibliography

Abazajian, K., *et al.*2005, AJ, 129, 1755

Adelman-McCarthy, J.K. *et al.*2006, ApJS 162, 38

Adelman-McCarthy, J.K. *et al.*2008, ApJS 175, 297

Baldry, I. K.; Glazebrook, K.; Brinkmann, J.; Ivezić, Z.; Lupton, R. H.; Nichol, R. C.; Szalay, A. S., 2004, ApJ, 600, 681

Baldwin, J.A, Phillips, M.M., & Terlevich, R., 1981, PASP, 93, 5

Barnes, J. E., & Hernquist, L. 1992, ARA&A, 30, 705

Baum, W.A., 1962, IAU Symp. 15, 390

Bell, E. F. et al., 2003, ApJS, 149, 289

Bell, E. F. et al., 2004a, ApJ, 608, 752

Bell, E. F. et al. 2004b, ApJL, 600, 11

Bell, E. F. et al. 2006, ApJ, 640, 241

Bell, E. F. et al., 2007, ApJ,663,834

Bell, E. F., 2008 ApJ, 682, 355

Benítez, N. 2000, ApJ, 536, 571

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Birnboim, Y., Dekel, A., & Neistein, E. 2007, MNRAS, 380, 339

Blanton, M. R., et al. 2003, ApJ, 594, 186

Borch, A., Meisenheimer, K., Bell, E. F., Rix, H.-W., Wolf, C., Dye, S., Kleinheinrich, M., Kovacs, Z., & Wisotzki, L. 2006, A&A, 453, 869

Bower, R. G., Lucey, J. R., & Ellis, R. S. 1992, MNRAS, 254, 601

Bower, R. G., Benson, A. J., Malbon, R., Helly, J. C., Frenk, C. S., Baugh, C. M., Cole, S., & Lacey, C. G. 2006, MNRAS, 370, 645

Brinchmann, *et al.*, 2004, MNRAS, 351, 1151

Brown, M. J. I., Dey, A., Jannuzi, B. T., Brand, K., Benson, A. J., Brodwin, M., Croton, D. J., & Eisenhardt, P. R. 2007, ApJ, 654, 858

Bruzual G. & Charlot S., 1993, ApJ, 405, 538

Bruzual, G. & Charlot, S., 2001, http://www.sdss.mpg.de/sdssMPA/Spectral Tools/

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Budavári, T., Szalay, A.S., Connolly, A.J., Csabai, I., & Dickinson, M.E.: in *"Photometric Redshifts and High Redshift Galaxies"*, eds. R.J. Weymann, L.J. Storrie–Lombardi, M. Sawicki, & R. Brunner, (San Francisco: ASP), 19 (1999)

Budavári, T., Szalay, A.S., Conolly, A.J, Csabai, I. & Dickonson, M.E., 2000, AJ, 120,1588

Budavári, T. et al., 2001, AJ, 122, 1163

Budavári, T. et al., 2003, ApJ,595,59

Butchins, S.A., 1981, A&A, 97, 407

Butchins, S.A., 1983, MNRAS, 203, 1239

Calzetti, D., Kinney, A.L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582

Cattaneo, A., *et al.*, 2006, MNRAS, 370, 1651

Chan, B.H.P., Mitchell, D.A., & Cram L.E. 2003, MNRAS, 338, 790

Charlot, S. & Fall, S.M. 2000, ApJ, 539, 718

Charlot, S. & Longhetti, M. 2001, MNRAS, 323, 887

Cole, S., Lacey, C. G., Baugh, C. M., & Frenk, C. S. 2000, MNRAS, 319, 168

Chen, H.-W., *et al.*2003, ApJ, 586, 745

Coleman, G.D., Wu., C.-C., & Weedman, D.W., 1980, ApJS, 43, 393

Connolly, A. J. & Szalay, A. S. 1999, AJ, 117, 2052

Connolly, A.J., Csabai, I., Szalay, A.S., Koo, D.C., Kron, R.G., & Munn, J.A., 1995, AJ, 110, 2655

Connolly, A.J, Szalay, A.S., Bershady, M.A., Kinney, A.L., & Calzetti, D. 1995, AJ, 110, 1071

Conselice, Christopher J., Bershady, Matthew A., Jangren, Anna, 2000, ApJ, 529, 886

Croton, D.J. *et al.*, 2006, MNRAS 365, 11

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

Csabai, I. *et al.*2008, in preparation

Dasyra, K. M., et al. 2006, ApJ, 651, 835

Dekel, A., & Birnboim, Y. 2006, MNRAS, 368, 2

—. 2008, MNRAS, 383, 119

Di Matteo, T., Springel, V., & Hernquist, L. 2005, Nature, 433, 604

Eisenstein, D. *et al.* 2001, AJ, 122, 2267-2280

Eisenstein, D.J. et al. 2003

Emsellem, E., *et al.*, 2007, MNRAS, 379, 401

Faber et al.,ApJ, 2007, 665, 265

Ferland, G. J. 1996, Hazy, a brief introduction to CLOUDY, University of Kentucky, Deptartment of Physics and Astronomy Internal report

Fernández-Soto, A. Lanzetta, K.M., & Yahil, A., 1999, ApJ, 513, 34

Fioc, M. & Rocca-Volmerange, B. 1997, A&A, 20, 306

Fukugita, M., Shimasaku, K., & Ichikawa, T. 1995, PASP, 107, 945

Fukugita, M., Ichikawa, T., Gunn, J.E., Doi, M., Shimasaku, K. & Schneider, D.P. 1996, AJ, 111, 1748

Gallazzi, A., Charlot, S., Brinchmann, J., White, S. D. M., & Tremonti, C. A. 2005, MNRAS, 362, 41

Gallazzi, A., Charlot, S., Brinchmann, J., & White, S. D. M. 2006, MNRAS, 370, 1106

Guo, F., & Oh, S. P. 2008, MNRAS, 384, 251

Gwyn, S.D.J. & Hartwick, F.D.A., 1996, ApJ, 468, L77

Häring, N., & Rix, H.-W. 2004, ApJL, 604, L89

Häußler, B. 2007, PhD Thesis, University of Heidelberg (http://www.ub.uni-heidelberg.de/archiv/7190)

Hopkins, Philip F. *et al.*, 2008, ApJS, 175, 356

Hopkins, Philip F. *et al.*, 2008, ApJS, 175, 390

Hubble, E. 1936, *The Realm of the Nebulae*, Oxford University Press

Humason, M. L. 1936, ApJ, 83, 10

Kauffmann, G., & Haehnelt, M. 2000, MNRAS, 311, 576

Kauffmann, G., *et al.*, 2003, MNRAS, 341,54

Kauffmann, G. *et al.*2003, MNRAS, 346, 1055

Kauffmann, G., Heckman, T. M., De Lucia, G., Brinchmann, J., Charlot, S., Tremonti, C., White, S. D. M., & Brinkmann, J. 2006, MNRAS, 367, 1394

Kauffmann, G., White, S. D. M., & Guideroni, B. 1993, MNRAS, 264, 201

Kennicutt, R. C., 1992a, ApJS, 79, 255

Khochfar, S., & Ostriker, J. P. 2008, ApJ, 680, 54

Kohonen, T., 1995, Self-Organizing Maps, Springer Series in Information Sciences

Koo, D.C., 1981, ApJ, 252, L75

Koo, D.C., 1985, AJ, 90, 148

Koo, D.C., 1986, ApJ, 311, 651

Koo, D.C., 1999, in

Kroupa P., 2001, MNRAS, 322, 231

Lahav, O., 2001, Compression and Classification Methods for Galaxy Spectra in Large Redshift Surveys, Mining the Sky, Proceedings of the MPA/ESO/MPE Workshop held at Garching, Germany, 31 July-4 August, 2000. Edited by A. J. Banday, S. Zaroubi, and M. Bartelmann. Heidelberg: Springer-Verlag, 2001., p.33 (astro-ph/0012407)

Lotz, J. M., *et al.*2008, ApJ, 672, 177

Lupton, R.H., *et al.*, 2001, ASPC, 238, 269

Magorrian, J., et al. 1998, AJ, 115, 2285

Malin, D. F., & Carter, D. 1983, ApJ, 274, 534

Masjedi, M., *et al.*2006, ApJ, 644, 54

McIntosh, D. H., Rix, H.-W., & Caldwell, N. 2004, ApJ, 610, 161

McIntosh, D. H., Bell, E. F., Weinberg, M. D., & Katz, N. 2006, MNRAS 373, 1321

McIntosh, D. H., Guo, Y., Hertzberg, J., Katz, N., Mo, H. J., van den Bosch, F. C., & Yang, X. 2008, MNRAS in press (arXiv:0710.2157)

Morgan, W. W. & Mayall, N. U. 1957, PASP, 69, 409

Moy, E., Rocca-Volmerange, B., & Fioc, M. 2001, A&A, 365, 347

Naab, T., & Burkert, A. 2003, ApJ, 597, 893

Naab, T., Johansson, P. H., Ostriker, J. P., & Efstathiou, G. 2007, ApJ, 658, 710

Naab, T., Khochfar, S., & Burkert, A. 2006, ApJL, 636, 81

[Padmanabhan, N. *et al.*, 2005 MNRAS, 359, 237-250

Panuzzo, P., Bressan, A., Granato, G. L., Silva, L., & Danese, L. 2003, A&A, 409, 99

Peng, C. Y. *et al.*, 2002, AJ, 124, 266

Pettini, M. & Pagel, B.E.J. 2004, MNRAS, 348, 59

Puschell, J.J., Owen, F.N., & Laing, R.A., 1982, ApJ, 275, L57

Riess, A. G. et al., 2001, ApJ,560,49

Ruhland, C., et al., 2008, submitted to ApJ

Sandage, A., & Visvanathan, N. 1978, ApJ, 225, 742

Sanders, D. B., & Mirabel, I. F. 1996, ARA&A, 34, 749

Scarlata, C., *et al.*2007, ApJS, 172, 494

Schawinski, K., Thomas, D., Sarzi, M., Maraston, C., Kaviraj, S., Joo, S.-J., Yi, S. K., & Silk, J. 2007, MNRAS, 382, 1415

Schiminovich, D., *et al.*2007, ApJS, 173, 315

Schlegel, D., Finkbeiner, D. P. & Davis, M. 1998, ApJ 500, 525

Schweizer, F., & Seitzer, P. 1988, ApJ, 328, 88

Schweizer, F., & Seitzer, P. 1992, AJ, 104, 1039

Schweizer, F., & Seitzer, P., Faber, S. M., Burstein, D., Dalle Ore, C. M., & Gonzalez, J. J. 1990, ApJL, 364, 33

Sheldon, E. S., et al., 2004, AJ, 127,2544

Skrutskie, M. F., *et al.*2006, AJ, 131, 1163

Sodré, L. & Stasinska, G. 1999, A& A, 345, 391

Somerville, R. S., Hopkins, P. F., Cox, T. J., Robertson, B. E., & Hernquist, L. 2008, MNRAS, in press (arXiv/0808.1227)

Stanford, S. A., et al. 2004, AJ,127,131

Stasinska, G. & Leitherer, C. 1996, ApJS, 107, 66

Stasinska, G. & Sodré, L. 2001, A& A, 374, 919

Stasinska, G. 1990, A&AS, 83, 501

Stoughton, C. *et al.* 2002, AJ, 123, 485-548

Strateva, I. *et al.* 2001, AJ, 122, 1861

Strauss, M. *et al.* 2002, AJ, 124, 1810-1824

Szalay, A.S., 1999, The Sloan Digital Sky Survey, CiSE, Vol. 1, No. 2, 54

Sérsic, J. L., 1968, Atlas de galaxias australes (Cordoba, Argentina: Observatorio Astronomico)

Toomre, A., & Toomre, J., 1972, ApJ, 178, 623

Trager, S. C., Faber, S. M., Worthey, G., González, J. J. 2000, AJ, 120, 165

Tremonti, C. A., Moustakas, J., & Diamond-Stanic, A. M. 2007, ApJL, 663, L77

Tremonti, C. *et al.* 2004,ApJ, 613, 898

Wolf, C., Meisenheimer, K., Röser, H.-J., 2001, A&A, 365, 660

Yasuda, N., *et al.*, 2001, AJ, 122, 1104

Yip, Ching-Wa *et al.* 2004, AJ, 128, 585

Yip, Ching-Wa *et al.* 2008, submitted to AJ

York, D. *et al.* 2000, AJ, 120, 1579

van Dokkum, P. G. 2005, AJ, 130, 2674