UNDERSTANDING THE PROPERTIES AND DIVERSITY OF GALAXIES AND QUASARS THROUGH SPECTRAL DECOMPOSITION

by

Ching-Wa Yip

M.Sc. The Chinese University of Hong Kong 1996

M.Phil. The Chinese University of Hong Kong 1998

Submitted to the Graduate Faculty of

Physics and Astronomy department in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF PHYSICS AND ASTRONOMY

This dissertation was presented

by

Ching-Wa Yip

It was defended on

August 17, 2005

and approved by

Prof. A. J. Connolly, University of Pittsburgh

Prof. R. K. Sheth, University of Pennsylvannia

Prof. A. S. Szalay, Johns Hopkins University

Prof. D. A. Turnshek, University of Pittsburgh

Prof. X. L. Wu, University of Pittsburgh

Dissertation Director: Prof. A. J. Connolly, University of Pittsburgh

UNDERSTANDING THE PROPERTIES AND DIVERSITY OF GALAXIES AND QUASARS THROUGH SPECTRAL DECOMPOSITION

Ching-Wa Yip, PhD

University of Pittsburgh, 2005

In the first part of this thesis, we study spectral decompositions of galaxies and quasars (QSOs) by the Karhunen-Loève (KL) transform using the Sloan Digital Sky Survey (SDSS). Our goal is to understand the average properties and sample variances of the data, with an eye toward obtaining objective classifications of these objects.

The eigencoefficients describing the galaxies naturally place the spectra into several classes defined by the plane formed by the first three eigencoefficients of each spectrum. Spectral types, corresponding to different Hubble-types and galaxies with extreme emission lines, are identified for $\sim 170,000$ spectra and are shown to be complementary to existing spectral classifications. Bias in the spectral classifications due to the aperture spectroscopy in the SDSS is within the signal-to-noise limit for majority of galaxies.

We extend the analysis to the decomposition of $\sim 17,000$ QSO spectra in which the diversity is known to be larger. From a commonality analysis on the eigenspectra sets constructed using different QSO samples, we deduce that QSO spectral classification is redshift and luminosity dependent. The prominent redshift effect is found to be the evolution of the small bump. The luminosity effect is related to the Baldwin effect. We therefore perform the KL transforms in three cases: the local (in the vicinity of emission lines), the intermediate (redshift and luminosity binned) and the global (restframe $900 - 8000$ Å) spectral bandpasses. We find that the second order QSO eigenspectra, in both the global and the intermediate spectral bandpasses, represent features from the host galaxies of the QSOs. We discuss the insights the results provide toward classification.

In the second part of this thesis, we probe spectroscopic variability of galaxies and narrow line active galactic nuclei (AGNs) using multi-epoch observations in the SDSS. We study the galaxy variability per spectral type defined previously in the KL transform and in the Osterbrock diagram. The amplitude of galaxy variability is found to depend on spectral type. We show that the variability in the H II galaxies can be partly due to star formation; and that in the AGNs and the mean eClass type D are probably not related to starbursts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I thank my parents, Yuk-Ha Yip and John Yip, my sister Edka Tse, my brother-in-law Vincent Tse, my brother Hon-Lam Yip, my sister-in-law Siu-Wai Yip, and my brother Jeffery Yip, for their support, encouragement and phone calls. My sister has been selflessly showing me everything outside academea. My father has an unspoken wish of becoming a mathematician. I dedicate this thesis to him.

The experience of pursuing a Ph.D. and studying with Andrew Connolly to me will *never die, only fade away*. His unique appreciation of interesting problems has transformed my views on astronomy. I am grateful for his inspirations in his disguising way, and for his providing me opportunity and freedom to work on problems that I am interested in. For these, I feel fortunate.

I thank Daniel Vanden Berk for he has opened windows for me in quasar studies. I am grateful for his ideas and guidance.

I thank the SDSS collaboration for their immense efforts in producing the excellent data.

I thank Xiao-Lun Wu for his patience and guidance in my second year in Pittsburgh when I studied biophysics, whom has continued to concern about my studies.

I am thankful to the members in Andrew Connolly's group and Ravi Sheth's group: Andrew Hopkins, Alberto Conti, Jeffery Gardner, Simon Krughoff, Samuel Schmidt, Cameron McBride, Niraj Welikala, Daniel Vanden Berk, Ryan Scranton, Jorg Colberg and Mariangela Bernardi. Their diverse backgrounds in astrophysics and computational science broaden my views during the weekly meetings. Especially I thank Ryan for sharing his valuable knowledge in using the SDSS data; and Andrew and Simon who have always shown their kindness and support. I thank members of Larry Wasserman and Christopher Genovese's statistics group in Carnegie Mellon University for discussions.

1.0 CLASSIFICATION PROBLEMS IN ASTRONOMY

If I could remember the names of all these particles, I'd be a botanist.

Enrico Fermi

Classification is a basic step in the development of most fields in science and it provides a basic framework for discussion. As pointed out by Sandage et al. (1975), the hypotheses from a useful classification lead to predictions which, if verified, help to form the theoretical foundations of a subject. A well-known example in astronomy is the classification of stars leading to the theory of stellar evolution.

While astronomy has gone through its initial stage of development, discussing classification issues have recently became necessary because of the massive amount of new data from sky surveys. One of the most ambitious surveys, the Sloan Digital Sky Survey (York et al. 2000), will produce the largest spectroscopic samples of quasars (QSOs) and nearby galaxies to date. These sophisticated data provide us an unique opportunity to explore the classifications of both types of objects.

In the following we discuss the applications of classification for both galaxies and quasars. We discuss the current understandings of galaxy formation and evolution, and present the motivations and the outline of this thesis.

1

1.1 APPLICATIONS OF CLASSIFICATION IN ASTRONOMY

1.1.1 Galaxy types

The topic of galaxy formation and evolution is difficult because the related processes involve the physics of and the interplay between stellar atmospheres, the interstellar medium, the intergalactic medium together with other dynamical processes such as merging of galaxies, for which approaches from first principles do not always exist. The spectrum of a galaxy is probably the most accurate tool for probing its stellar composition and evolution. Galaxy spectra and their classification hence provide direct clues to how galaxies form and evolve.

An example of the power of galaxy classification is the morphological segregation of galaxies (Dressler 1980). Elliptical galaxies are more common in the centers of rich clusters than spiral galaxies, while in the field the opposite tendency is true. Subsequently there was the discovery of post-starburst galaxies (Dressler and Gunn 1983) (later also called the K+A galaxies[1]) in galaxy clusters, showing that an observed galaxy spectrum and its stellar content depend highly on the environment, and that a locally high density environment may suppress the formation of the disk component of spiral galaxies.

In Table 1.1 we summarize the timeline of spectroscopic classification of galaxies in the optical wavelengths. We note that the classifications are proposed for a variety of purposes, ranging from discriminating the stellar content (early and late types, post-starburst galaxies, eClass), the star formation rate (SFR; starburst and post-starburst galaxies), to the nuclear activity (Seyfert galaxies and emission-line galaxies).

1.1.2 Quasar types

Classification of QSOs is less well established compared with that of galaxies. As one type of active galactic nuclei (AGNs), the light-generating mechanism of QSOs are different from that in normal (or "inactive" by some researchers) galaxies, as such gravitational energy of the supermassive black hole (BH) at the center of the QSO is released in the form of radiation through the process of accretion. Always the radiations are modeled by the integrated flux

---

[1]The symbol "K+A" refers to a mixed stellar populations of K stars and A stars.

Table 1.1 Optical spectroscopic classification of major galaxy types.

Year	Galaxy types	Reference
1926	early and late types[a]	Hubble
1943	Seyfert galaxies[b]	Seyfert
1981	starburst galaxies[c]	Weedman et al.
1983	post-starburst galaxies	Dressler and Gunn
1987	emission-line galaxies[d]	Veilleux and Osterbrock
1995	eClass	Connolly et al.

[a]Originated from morphological classification, referring to Hubble's belief that elliptical galaxies are of "earlier" types of spiral galaxies. The notation persisted and in spectroscopy "early-types" refers to galaxies with spectra similar to *late-type* stars, and vice versa.

[b]Galaxies with powerful radiation from the nucleus, and show strong, high-excitation emission lines in their spectra.

[c]An extreme form of inactive galaxies which show a high rate of star formation.

[d]Comprises four types of galaxies – the Seyfert 2 galaxies, the narrow-line radio galaxies, the H II galaxies and the low ionization nuclear emission regions (LINERs).

of a locally thermal radiation, determined by a certain temperature profile on the accretion disk (Ross et al. 1992).

There are two major types of AGN: Type 1 and 2. In the optical, Type 1 AGNs are characterized by broad emission lines (full-width-at-half-maximum (FWHM) $\sim 500 - 30,000$ km s^{-1}, Peterson 1997), where the lines are Doppler broadened due to the high orbital velocity of a broad line region (BLR) close to the central black hole. Type 2 AGNs exhibit narrow emission lines (FWHM $\sim 200 - 900$ km s^{-1}) of a variety of ionization states. Under the scope of a phenomenological model, the unification model, the two types of AGN are actually the same type of object but with different inclination angle relative to the line of sight, with the Type 1 AGNs being objects having the accretion disk and the optically thick (dusty) torus closer to face-on such that the regions responsible for emitting the broad emission lines (broad line regions; BLRs) can be seen by an observer, and edge-on for that in the Type 2 AGNs. In this regard, the classification of AGNs gives a different purpose and favor from those of stars and galaxies.
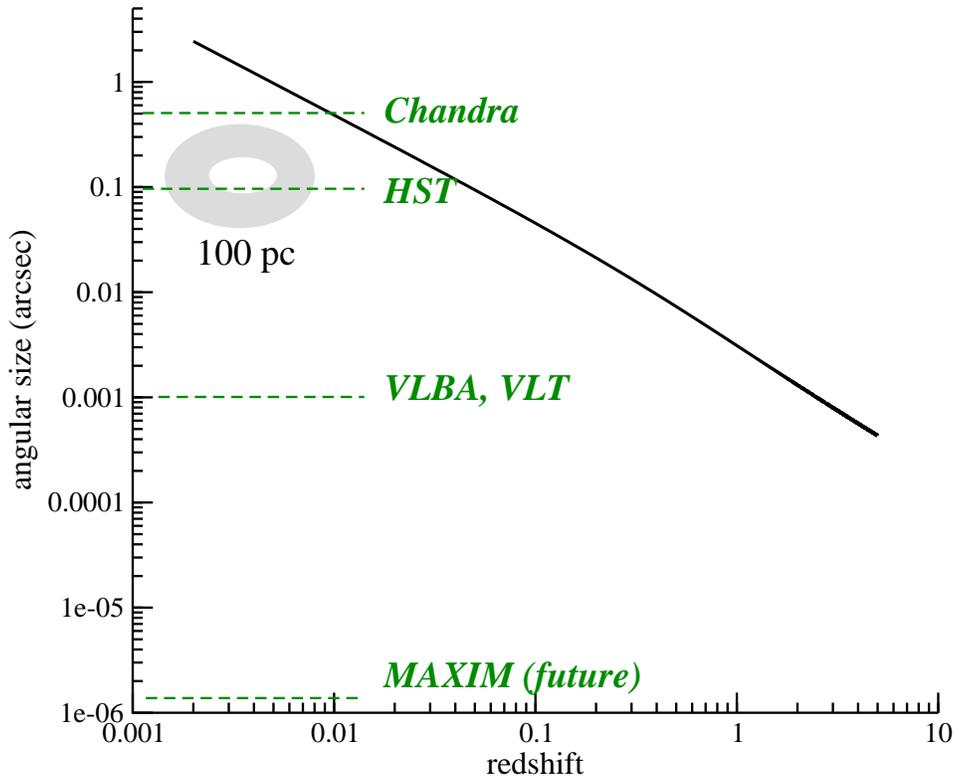
Figure 1.1 Angular size of an object as a function of redshift. Resolutions of some of the the largest telescopes in the world are marked. (Chandra: the Chandra X-ray telescope; HST: the Hubble Space Telescope (UV-optical-infrared); VLBA: the Very Long Baseline Array (radio); VLT: the Very Large Telescope (UV-optical-infrared); MAXIM: the Micro-Arcsecond X-ray Imaging Mission)

Table 1.2 Distance scales of structures in AGNs and their host galaxies (from various literatures).

accretion disk	10^{-3} pc
broad line region (BLR)	0.1 pc
optically thick torus	$10 - 100$ pc
narrow line region (NLR)	10 pc -10 kpc
host galaxy	10 pc -100 kpc
radio jets and lobes	$10 - 1000$ kpc

One of the reasons to study QSOs by their spectra is because we can yet see directly their central structures. In Fig. 1.1 we show the angular size (in arcsecond, $''$) vs. redshift for a dusty torus of size 100 pc[2]. At the redshift of 6, the dusty torus is just within the resolution of the VLBA and VLT. For smaller structures like the accretion disks (10^{-3} pc), they are beyond the reach of the largest telescopes in the world at present. The MAXIM in the future is a promising tool to probe the event horizons of the central BHs of AGNs. In Table 1.2 we summarize the current understandings of the scales of various structures found in AGNs.

With bolometric luminosities that can reach $\sim 10^{48}$ erg s^{-1} (Krolik 1999; Vestergaard 2004), and redshifts in excess of ~ 6 (Fan et al. 2003), QSOs serve as tools, in conjunction with studies of the intergalactic medium, for probing conditions in the early universe. It is thus very important to understand the properties of QSOs.

1.2   CONTEMPORARY TECHNIQUES

Classification of spectra were traditionally carried out through visual inspection, often accomplished by a small number of experts. New methods are inevitable for studying large sets of data, which should be able to consider relevant information objectively and automatically,

---

[2]By assuming $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ and $H_0 = 71$.

and to reduce the error rate. In Fig. 1.2 we show the number of objects in major redshift surveys. The increase of the number of redshifts for both galaxies and QSOs (after the year 1995) follows nearly the Moore's Law. While this coincidence is probably by the designs of the surveys, the speed of increase of the data is high by the common standard in computer science.

One of the first works is by Connolly and his co-workers (Connolly et al. 1995) in which a linear combination of an orthogonal set eigen-functions (called "eigenspectra" by the authors) was adopted to represent each galaxy spectrum. The method used to construct the eigenspectra, namely the Karhunen-Loève (KL) transform, is a powerful and popular technique used in classification and dimensional reduction of massive data sets. The eigenspectra were shown by the authors to be applicable to classifying galaxy spectra into a sequence ranging from the old to the younger stellar populations (later known as the "eClass"). The KL transform has been applied in the spectroscopic classifications of stars (Bailer-Jones et al. 1998; Singh et al. 1998), galaxies (Connolly et al. 1995; Sodre and Cuevas 1997; Bromley et al. 1998; Galaz and de Lapparent 1998; Ronen et al. 1999; Folkes et al. 1999; Yip et al. 2004a) and QSOs (Boroson and Green 1992; Francis et al. 1992; Shang et al. 2003; Yip et al. 2004b).

The other notable method used in classification is the artificial neural network. Also called "supervised learning", the idea behind is the users manually classify a subset of the data, and from this the neural network is "trained" to classify the remaining of the data. This method has been applied in the spectroscopic classifications of stars (von Hippel et al. 1994; Weaver and Torres-Dodgen 1997; Bailer-Jones 1997; Singh et al. 1998; Snider et al. 2001), galaxies (Folkes et al. 1996) and AGNs (Rawson et al. 1996).

There are many works applying other machine learning techniques to classification problems in astronomy. The decision tree induction (returns a boolean function "yes" or "no" by finding logical patterns within the input data) has been used in star-galaxy separation (Weir et al. 1995; Odewahn et al. 2004) and in morphological classification of galaxies (Owens et al. 1996). The information bottleneck method (finds the best trade-off between accuracy and compression when clustering a random variable x, given the joint probability distribution between x and an observed variable y) has been applied in galaxy spectral classification
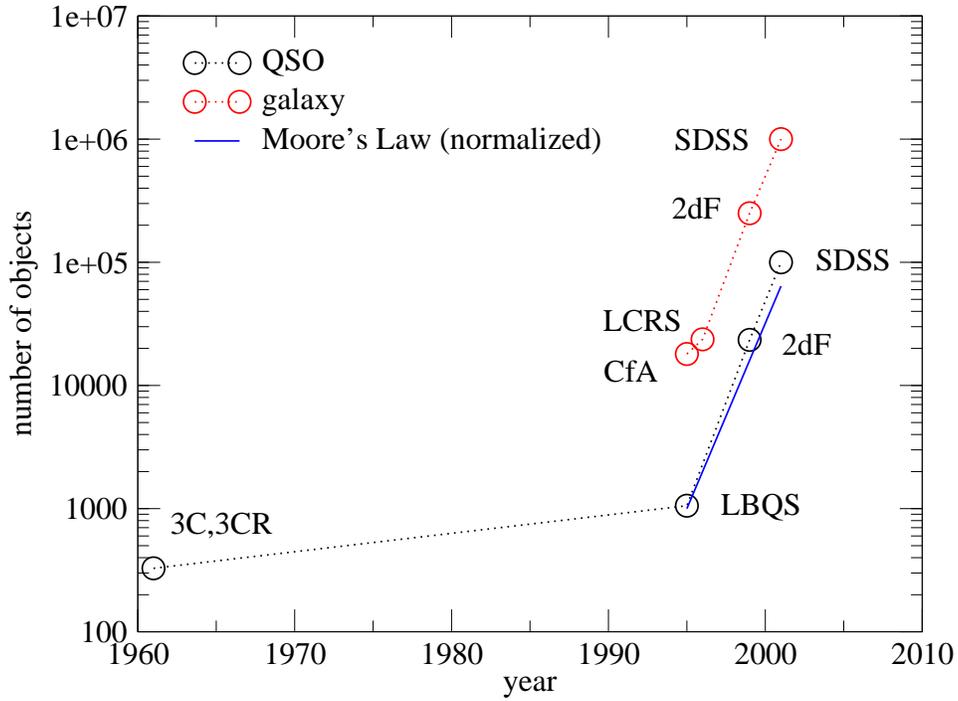
Figure 1.2 Number of galaxies and QSOs in redshift surveys. (3C: the third Cambridge catalog (redshifts measured after the survey); LBQS: the Large Bright Quasar Survey; 2dF: the two degree field survey; CfA: the Center for Astrophysics redshift survey; LCRS: the Las Campanas redshift survey)

Table 1.3 Comparisons between stellar population synthesis models.

Parameters	Bruzual & Charlot[a](2003)	PEGASE (1997)	Starburst (1999)
λ_{rest} (Å)	$3,200 - 9,500$	$220 - 20,000$	$91 - 160,000$
resolution	3 Å linear	≥ 3 Å uneven	≥ 3 Å uneven
age (Myr)	$0.1 - 10,000$	$1 - 2,000$	$1 - 1,000$
metallicity, Z	$0.0004 - 0.03$	$0.0001 - 0.1$	$0.04 - 0.001$
star formation	single burst/continuous	single burst/continuous/$\propto$ gas mass	single burst/continuous
note		with gas component	optimized for star forming galaxies

[a]We only list data for the highest resolution models.

(Slonim et al. 2001).

Furthermore, to extract physical parameters from galaxy spectra, a promising method is the Multiple Optimised Parameter Estimation and Data compression (MOPED, Heavens et al. 2000), a maximum likelihood method which reduces the number of wavelength bins in a spectra without increasing the uncertainty in the estimated parameters. It has been applied to the fast estimation of the star formation history, metallicity and dust content of galaxies (a few seconds for 17 parameters, Reichardt et al. 2001).

Parameter estimations of galaxy spectra requires stellar population synthesis models, which represent the stellar contents by considering coherently the properties of, for example, the stellar atmosphere, the star formation, the stellar evolution and the supernova explosion. The most popular ones at present are the Bruzual and Charlot models (Bruzual and Charlot 2003), the Projet d'Etude des GAlaxies par Synthése Evolutive models (PEGASE, Fioc and Rocca-Volmerange 1997), and the Starburst 99 synthesis models (Leitherer et al. 1999). In Table 1.3 we compare the major parameters between them.

These models are based on the evolutionary population synthesis technique (Tinsley 1978), in which the main parameters are the initial mass function (IMF, the number of stars within a given mass interval), the SFR, and the rate of chemical enrichment (e.g., by the supernova explosion or the stellar wind) to the interstellar medium of the galaxy. The time evolution modeling of these parameters results in a distribution of stars of different ages in the Hertzsprung-Russell diagram, which in turn gives the integrated spectral evolution of the stellar population in the galaxy.

1.3 MOTIVATIONS

Under the scenario of hierarchical galaxy formation (Press and Schechter 1974), galaxies form in the locally high density regions in the universe (the dark matter halos) through baryon cooling (the free-free transitions). Many galaxies in the local universe have undergone mergings with the other galaxies during their lifetime by the interactions between the dark matter halos. Stars form and die (by forming compact objects) along the process, and occasionally materials are recycled throughout the galaxies for the formations of next-generation stars through mainly supernova explosions. The stellar populations in galaxies and their clustering properties are thus the "fossils" resulted from the above processes.

On the other hand, the discovery of supermassive BHs ($10^6 - 10^9$ M$_\odot$) in many inactive galaxies in the local universe (Kormendy et al. 2001), and the fact that the values of BH masses coincide with those in QSOs (Vestergaard 2004; McLure and Dunlop 2004), suggest that BH plays an important role in the evolution of galaxies. The current understanding is that QSOs were once common in most massive galaxies (Kormendy et al. 2001), since then the AGN activity has been dying out, so QSOs are quite rare in the local universe. Furthermore, a recent CO radio imaging (Walter 2005) gives a dynamical mass estimates of $\sim 10^9$ M$_\odot$ of a $z = 6.42$ QSO, inconsistent with a 10^{12} M$_\odot$ stellar bulge normally found in the present-day galaxies, suggesting that supermassive BHs formed before galaxy bulges during the formations of galaxies. The nuclear activity is therefore an important physical parameter to describe galaxy evolution.

The long term goals in studies of stellar populations in galaxies are hence to at minimal understand the star formation rate, the metallicity, the gas content, the extinction characteristics and the nuclear activity of each galaxy and how are they related to its evolutionary path. To probe the above processes is, and will continued to be, a big challenge for different areas in astrophysics, as the tracking of the evolutionary history of individual galaxy by their appearance in the local universe is both a multivariate and non-linear problem.

Many of these important questions will be answered by the SDSS, a multi-frequency survey that will map a quarter of the entire sky. At its completion, the SDSS will determine the positions and magnitudes of $100,000,000$ celestial objects, and measure the spectra of

$1,000,000$ galaxies and $100,000$ quasars[3]. These will result in the largest map of our local universe with complete samples of spectroscopic data.

Progresses have already been made toward understanding galaxy formation by using the SDSS data. For example, Gómez et al. (2003) have shown that the SFR of galaxies in $0.05 \leq z \leq 0.095$ is correlated strongly with the local density of the environment and that the decrease of SFR in galaxies in dense environments is universal over a wide range of densities. Kauffmann et al. (2003) have constrained the star formation histories, dust extinction and stellar masses of 10^5 galaxies and show that there is a bimodal distribution in the 4000 Å break, indicating a division between galaxies dominated by old stellar populations and those with more recent star formation.

The goals of this thesis are twofold. Firstly, we study the properties and the diversity of galaxies and quasars using the SDSS spectra through spectral decompositions, using the KL transform, in order to improve our understanding on the populations of these objects, and to explore the spectral features represented by the eigenspectra. Secondly, we study the spectral variability of inactive galaxies and Type 2 AGNs in order to understand and discriminate their nuclear activity. The results will gain us insights on galaxy formation and evolution, especially in combination with works on spectral parameter estimations and clustering analysis in the future.

1.4 OUTLINE OF THESIS

In this thesis, we firstly apply the KL transform to decompose large samples of galaxy and quasar spectra in the SDSS. We make physical interpretations of the eigenspectra and study sample variations for both types of objects. We discuss the applications of the eigenspectra. We hence probe nuclear activity in inactive galaxies and narrow line AGNs in the SDSS by applying the spectral classification from the KL transform.

In chapter 2, we apply the KL transform to a sample of galaxy spectra in the SDSS. We investigate the convergence behavior of the galaxy eigenspectra as a function of gap

---

[3]http://www.sdss.org

correction step and sample size. We study the effects of sample size and aperture bias on the classification.

In chapter 3, we apply the KL transform to a sample of QSO spectra in the SDSS. We establish the importance of studying the eigenspectra in wavelength bandpasses of different sizes. We exploit the quasar eigencoefficients as tools to probe redshift and luminosity effects in the spectra, and argue the importance of both effects in QSO classification.

In chapter 4, we probe the temporal variability of spectra of galaxies and narrow line active galactic nuclei. We explore galaxy variability and difference spectrum per spectral type as defined in chapter 2 and in the Osterbrock diagram. We investigate the physical origins of the variabilities.

In chapter 5, we conclude the results and discuss future directions.

Following the convention in the SDSS, every spectrum in this thesis is expressed in vacuum wavelengths and in the unit of Å. Unless otherwise specified, spectral flux densities f_λ are expressed in 10^{-17} ergs s^{-1} cm^{-2} Å^{-1}. In all the relevant figures, the restframe wavelengths of emission lines are indicated by dotted green lines in order to guide the eye, so that they do not necessarily imply the presence of the lines in the spectrum.

2.0   DECOMPOSITION OF GALAXY SPECTRA

Of every goodness and badness, solely our hearts know the best.

Chinese Proverb

In this chapter we apply the Karhunen-Loève transform to classify a sample of galaxy spectra from the SDSS. We present results of a convergence analysis of the Gap-Correction Formalism (Connolly and Szalay 1999) in the KL Transform. With the gap-correction implemented, eigenspectra for a large galaxy sample from the SDSS are constructed and their physical interpretations are made.

2.1   DATA

As part of the Sloan Digital Sky Survey (York et al. 2000) spectra are taken with fibers of $3''$ diameter (corresponding to 0.18mm at the focal plane for the 2.5m, f/5 telescope). All sources are selected from an initial imaging survey using the SDSS camera described in Gunn et al. (1998) with the filter response curves as described in Fukugita et al. (1996), and using the imaging processing pipeline of Lupton et al. (2001). The astrometric calibration is described in Pier et al. (2003). The photometric system and calibration are described in (Smith et al. 2002; Hogg et al. 2001; Ivezić et al. 2004b). We select the Main Galaxy sample (Strauss et al. 2002) for our analysis and use only those galaxies defined as being of survey quality: a signal-to-noise lower-limit of approximately 16. The galaxies in this sample have r-band Petrosian magnitudes $r_p < 17.77$ and Petrosian half-light surface brightnesses $\mu_{50} < 24.5$ mag arcsec^{-2}, defined to be the mean surface brightness within a circular aperture

containing half of the Petrosian flux (called the Petrosian half-light radius). The spectral reductions used are the standard SDSS 2D analysis pipeline (idlspec2D v4.9.8, as of 18th of April, 2002) and the 1D SpecBS pipeline in the SDSS. The resultant spectra are flux- and wavelength-calibrated, and sky-subtracted. From these data approximately two hundred galaxies are removed, as they have zero flux in all pixels. This results in a final sample of 176,956 galaxy spectra. The median redshift is about 0.1, and we find that about 6.5% of the sample have redshifts $cz < 10,000$ km s^{-1}, so that their Petrosian half-light radii can be substantially larger than the $3''$ aperture of the fiber (Strauss et al. 2002). 1,854 spectra of the final sample are found to be duplicated observations; identified as being within a search radius of $2''$ and a redshift tolerance of 0.01. All spectra are shifted to a common rest frame, and rebinned to a vacuum wavelength coverage of $3450 - 8350$ Å. The binning of the spectra is logarithmic, with a velocity dispersion of 69 km s^{-1}. This procedure emphasizes the blue end of the optical spectrum, enabling our analysis to focus on the Ca H and Ca K lines, and the Balmer break. The resultant spectra cover rest-wavelength range $3450 - 8350$ Å over 3839 pixels.

2.2 CONVERGENCE ANALYSIS OF GAP CORRECTION

An assumption has always been made when researchers apply the KL trasform in spectral classification, that is, the spectra are without any gaps. In reality, however, there are several reasons for gaps to exist: different rest-wavelength coverage, the removal of skylines, bad pixels on the CCD chips all leave gaps at different rest frame wavelengths for each spectrum. All can contribute to incomplete spectra. The idea behind the gap-correction process is to reconstruct the missing regions in the spectrum using its principal components. The first application of this method to analyze galaxy spectra is due to Connolly and Szalay (1999), which expands on a formalism developed by Everson and Sirovich (1994) for dealing with two-dimensional images. Initially, we fix the missing data by some means, for example, linear-interpolation. A set of eigenspectra are then constructed from the gap-repaired quasar spectra. Afterward, the gaps in the original spectra are corrected with the linear combination

of the KL eigenspectra. The whole process is iterated until the set of eigenspectra converges. The mathematics concerning the KL transform is presented in appendix A.

There are some questions to be solved in our analysis before the KL eigenspectra and hence the classification itself become robust and meaningful. These questions are: do the resultant eigenspectra converge and, if so, how many iteration steps are required, what is the dependency of the quality of the KL-repaired spectra on how the gaps are initially corrected, how much information is contained in the eigenspectra and most importantly, how many galaxy spectra are needed in order to derive a convergent set of eigenspectra?

Several authors have tried to assess the performance of a KL analysis in a number of different quantitative ways. An example of this is the χ^2 assessment (Francis et al. 1992) in which the authors calculated the difference between the observed spectrum and the spectrum reconstructed with the principal components in order to determine the number of components needed for reconstructing a quasar spectrum. With the implementation of gap-corrections in our analysis, this comparison of only *one* spectrum to another may not suffice. We are more interested in how well the *set* of eigenspectra describes the distribution of spectra rather than a one-to-one comparison. For example, how does the set of eigenspectra differ as the gap-correction procedure progresses? Given two subspaces, each formed by a set of eigenspectra obtained with different conditions (e.g., at different points in the iterative gap correction or computed with different numbers of observed spectra), we require a method that will quantitatively compare one set of eigenspectra with another. In other words, instead of just comparing two spectra, a mechanism is desired to compare two subspaces, which are spanned by a finite number of spectra respectively. Mathematically it can be stated, as in Everson and Sirovich (1994), that two spaces, E and F, are in common if

$$Tr(\mathbf{EFE}) \;=\; D \;,\tag{2.1}$$

where $\mathbf{E}$ and $\mathbf{F}$ are the sum of projection operators of space E and F respectively, and D is equal to the dimensionality of each space. We assume that these eigenbases have the same dimensions for a meaningful comparison. The sum of the projection operators, $\mathbf{E}$, of a space is given by the sum of the outer products

$$\mathbf{E} = \sum_{\epsilon} |\epsilon><\epsilon| \;,\tag{2.2}$$

where $|\epsilon >$ are the basis vectors which span the space E (e.g., Merzbacher 1970). A basis vector is an eigenspectrum if E is considered to be a set of eigenspectra. The two spaces are disjoint if the trace quantity is zero and are identical if the quantity is equal to the dimension of the subspace. This provides a quantitative way of measuring the commonality of two subspaces (i.e., how similar the two subspaces are).

In investigating the convergence behavior of eigenspectra as a function of the number of iterations, we define one of the spaces to be that formed by a finite number of eigenspectra obtained after initially interpolating over the gap regions, but *without* gap correction. The other spaces are defined to be those formed by the same finite number of eigenspectra but at different iterations in the gap correction. The sum of projection operators in the first case is named e^0 and in the latter, e^i. The subspaces are named E^0 and E^i respectively, where i denotes the i-th iteration. The dimension of the space is D_e, which is the number of eigenspectra forming the subspace.

The trace quantity $Tr(e^i e^0 e^i)$ as a function of iteration is plotted in Fig. 2.1a, in which the KL transform is applied to $N = 4003$ randomly-chosen SDSS galaxy spectra, where D_e is set to be 5 and 10, meaning that the subspaces are spanned by the the first five and the first ten eigenspectra respectively. Repairing of the galaxy spectra in the iteration procedure is performed with $m = 10$ eigenspectra. It should be noted that m is independent from D_e and that D_e is always smaller or equal to m. The traces are normalized by the corresponding D_e in each curve to simplify comparison. Initially, let us concentrate on the curves with open symbols in Fig. 2.1. These curves denote that the initial gappy regions are approximated by linear-interpolation. In this linear-interpolation method the flux of a pixel, $f^g_{\lambda_k}$, in the gap is simply approximated by the average of its neighbors, so that

$$f^g_{\lambda_k} = (f_{\lambda_{k-1}} + f_{\lambda_{k+1}})/2 \ . \tag{2.3}$$

The trace quantity decreases gradually as the iteration step increases, indicating that the space E^i is less and less in common with E^0. This implies that the KL gap-correction and eigenspectra construction are changing the spectrum of a galaxy within the gappy regions. As such the eigenspectra from the KL-repaired spectra differ progressively more from those formed from the original spectra. The above is true for both $D_e = 5$ and 10. The above is
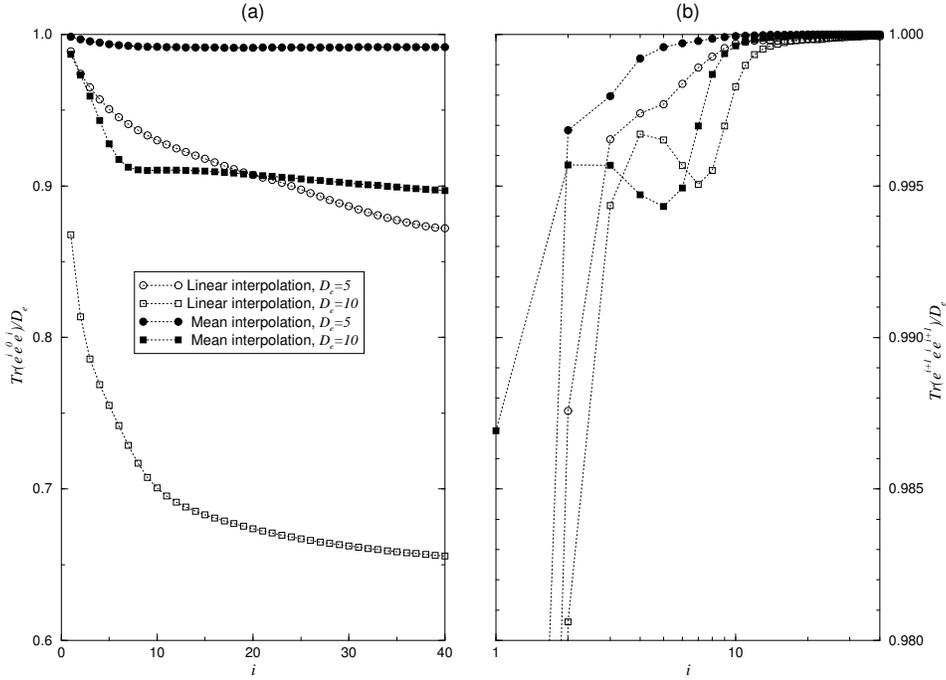
Figure 2.1 Convergence of eigenspectra (a) $Tr(e^i e^0 e^i)$ and (b) $Tr(e^{i+1} e^i e^{i+1})$ as a function of iteration step in the KL gap-correction formalism. Curves with open symbols are linearly-interpolated across the gaps while filled symbols represent mean-interpolated data. The circles and squares denote $D_e = 5$ and 10 respectively, where D_e is the dimension of the subspace formed by the eigenspectra.

generally valid for D_e from $1-10$. As the iteration increases, the slope of the curve decreases, which implies that a we have converging set of eigenspectra.

The choice of linear-interpolation in the initial correction for the gappy regions is arbitrary. In fact, if the gap formalism is robust, the quality of the KL-repaired spectrum and the eigenspectra should be independent of the way the observed spectra are initially repaired. We test an alternative method of correcting for gaps, where the flux at each wavelength bin in the gap region is approximated by the mean of all other spectra within that region, i.e., the flux of a pixel in a gap is approximated by,

$$\hat{f}^g_{\lambda_k} = \left\langle \sum_{all\ spec} \hat{f}_{\lambda_k} \right\rangle_{all\ spec} , \qquad (2.4)$$

where $\hat{f}_{\lambda_k}$ is the normalized flux at λ_k. We call this method mean-interpolation. With this alternate method the trace quantity also converges as we can see from Fig. 2.1a, but at a higher value than those in the case of linear-interpolation. This behavior shows that, in the case of mean-interpolation, the eigenspectra constructed after the gap correction deviate *less* from the initial interpolated spectra than in the case of linear-interpolation. The rate of convergence is faster when using the mean-interpolation method. This suggests that the mean-interpolation provides a better initial estimate of the true spectra within the gap regions. We do not, therefore, require as many iterations as in the case of linear-interpolation. This is important as each step in repairing the spectra and constructing the eigenspectra is computationally expensive when large amounts of data are under consideration.

Fig. 2.1b shows the convergence behavior of the sets of eigenspectra given in Fig. 2.1a except that the trace quantity now compares the subspace from one iteration to the subspace from the next iteration, i.e., $Tr(e^{i+1}e^i e^{i+1})$. As expected, the convergence with the number of iteration steps can be seen in both methods, but with this more sensitive measurement the convergence is now no longer found to be monotonic. This implies that we may need more iterations, than it first appeared from our previous example in order to obtain a convergent set of eigenspectra. Again, the mean-interpolation method is shown to converge to a consistent set of eigenspectra faster than for linear-interpolation. Consequently, in the following all gaps in the spectra will be fixed using the mean-interpolation method, unless otherwise specified.
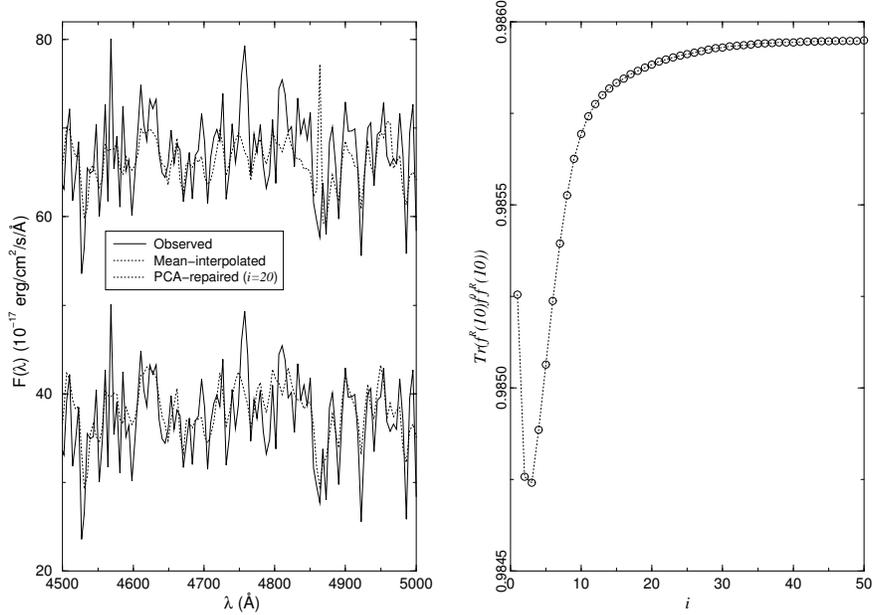
Figure 2.2 (a) The mean-interpolated and KL-repaired (20 iterations) of an artificially masked spectra, overlaid on the original unmasked spectrum. (b) KL-repaired spectra converge as a function of iteration.

An example of the actual performance in the mean-interpolated and repaired spectra is shown in Fig. 2.2. The data set is the same 4003 galaxies as before, except that one randomly chosen spectrum is artificially masked in the region [4500, 5000]Å. The upper panel of Fig. 2.2a shows that the mean-interpolated region is already close to that of the original spectrum before masking. The lower panel shows the KL-repaired spectrum at $i = 20$ overlaid on the original spectrum, using all 10 eigenspectra. The spectra are offset by an arbitrary amount for illustration. There is a substantial improvement in retrieving the original spectrum as the iterations proceed. To compare the KL-repaired and unmasked spectrum quantitatively, we apply a similar convergence measure as described previously. The convergence measure in this case is defined to be

$$Tr(f^R(m)f^0 f^R(m)) \ , \tag{2.5}$$

where $f^R(m)$ is the projector of the KL-repaired spectrum with m eigenspectra in the gaps, and f^0 is that of the unmasked spectrum. In Fig. 2.2b, the trace quantity versus the number

18
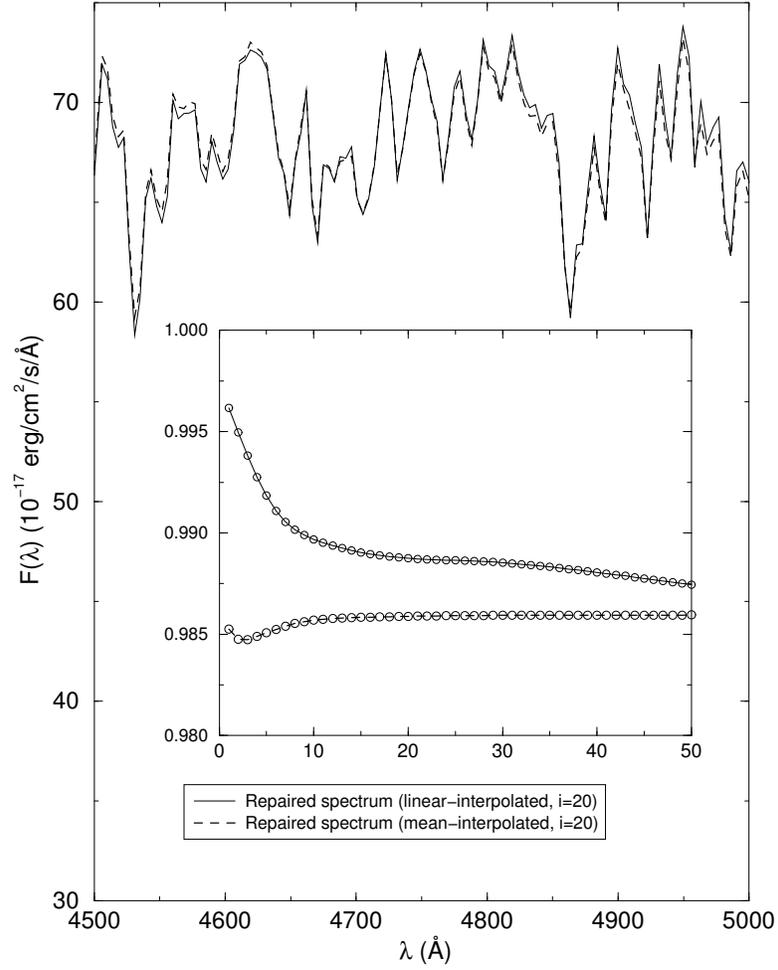
Figure 2.3 The KL-repaired spectra, linear-interpolated (solid line) and mean-interpolated (broken line), in the gap regions of the unmasked spectrum as in Fig. 2.2 (after 20 iterations). The insert is $Tr(f^R(m)f^0 f^R(m))$ as a function of iteration step for both cases.

19

of iterations is plotted for the case corresponding to Fig. 2.2a. $Tr(f^R(m)f^0 f^R(m)) = 1$ means that the repaired spectrum is identical with the original unmasked one. We see that after the initial few iterations, the two become more similar. After 30 iterations, the KL-repaired spectrum converges to that of the original spectrum with a high degree of accuracy, the difference in the trace quantities of 8×10^{-4} %.

Combining the results of the convergence measures in both KL-repaired spectra and the previously discussed eigenspectra, it can be concluded that the convergence of the eigenspectra implies the convergence of the repaired spectrum, and vise versa. Furthermore, the quality of a repaired spectrum should not depend on the initial gap-interpolation technique. Fig. 2.3 shows the KL-repaired spectra, using linear- and mean-interpolations for the initial gap approximation, $m = 10$ and at $i = 20$. The two are shown to be very similar to each other. The insert shows the corresponding $Tr(f^R(m)f^0 f^R(m))$ as a function of iteration. The convergence behavior seems different in both cases, nevertheless they are approaching each other with a difference in the actual value about 0.3% which is small as can be seen in the plots of the spectra in the main graph. This is a desired result, because if the whole formalism is robust, the repaired spectrum should not be different due to different procedures used in the initial gap-fixing.

2.3   EFFECT OF SAMPLE SIZE

Another important aspect of the KL-eigenspectra construction is the number of observed spectra necessary as the input. In principal, as we increase the number of galaxy spectra, a more representative and general set of eigenspectra should result. The question remains, however, how much more generality would be gained by including more observed spectra in the analysis? Fundamentally, does there exist a minimum number of input galaxy spectra such that the eigenspectra set start to converge? This is important because we can thus use a minimum number of randomly-chosen observed spectra in the survey to derive a set of eigenspectra which nevertheless contain all the necessary information within the full data set. Fig. 2.4 shows an attempt to answer this question. In these figures, the commonality

percentages of two subspaces spanned by (a) 2 (b) 3 and (c) 10 eigenspectra are plotted versus the number of galaxy spectra $N_{(h)}$ used in the sample. The commonality is similar to that previously discussed for the trace quantities except here we compare the set of eigenspectra derived from $N_{(h)}$ input galaxy spectra with that from a smaller number of spectra $N_{(h-1)}$. This is defined as follows

$$commonality(\%) \equiv \frac{Tr(e(N_{(h-1)})e(N_{(h)})e(N_{(h-1)}))}{D_e} \times 100\% , \qquad (2.6)$$

where $e(N_{(h)})$ is the sum of projectors of the subspace spanned by D_e eigenspectra, derived from $N_{(h)}$ galaxy spectra, using $m = D_e$ eigenspectra for gap-repairing. The number of iterations for the gap-correction is 20. The smallest number of spectra we consider is 139 ($= N_{(0)}$) and the largest number 40044. The galaxies in each case are randomly selected from the full SDSS sample.

For all cases, convergent trends are present as more spectra are included. For the case of two eigenspectra (Fig. 2.4a) we can see that only about 500 galaxy spectra are needed in order to construct the first two eigenspectra to an 0.5% accuracy when compared to the final converged set. The inclusion of the third mode (Fig. 2.4b) requires more spectra to obtain a similar accuracy though the convergence behavior is very similar to (a) (with slightly larger fluctuations). Nevertheless, only about 1000 galaxy spectra are needed for 90% commonality. These results are consistent with the fact that most types of galaxy spectra can be described with 2 to 3 colors (Connolly et al. 1995) and therefore a random sampling of a few thousand galaxies can be expected to cover the full color distribution for these galaxies.

In Fig. (c), it is interesting that the convergent behavior is different from that in (a) and (b). With a sample size smaller than about 3000 to 4000, the improvement in the set of 10 eigenspectra is small. However, once the number of spectra used exceeds that threshold, the convergent rate dramatically increases. This finding suggests that there exists a minimum number of galaxy spectra that we need to include in our KL analysis in order to fully sample the true distribution of galaxy spectra. Combining this with the fact that the higher-order modes in the eigenspectra tend to correspond to spectral features in galaxies with prominent emission lines (this will be discussed in detail in §2.4) and the fact that those galaxies only comprise about 0.1% of the whole sample, the behavior in (c) can be understood as the effect
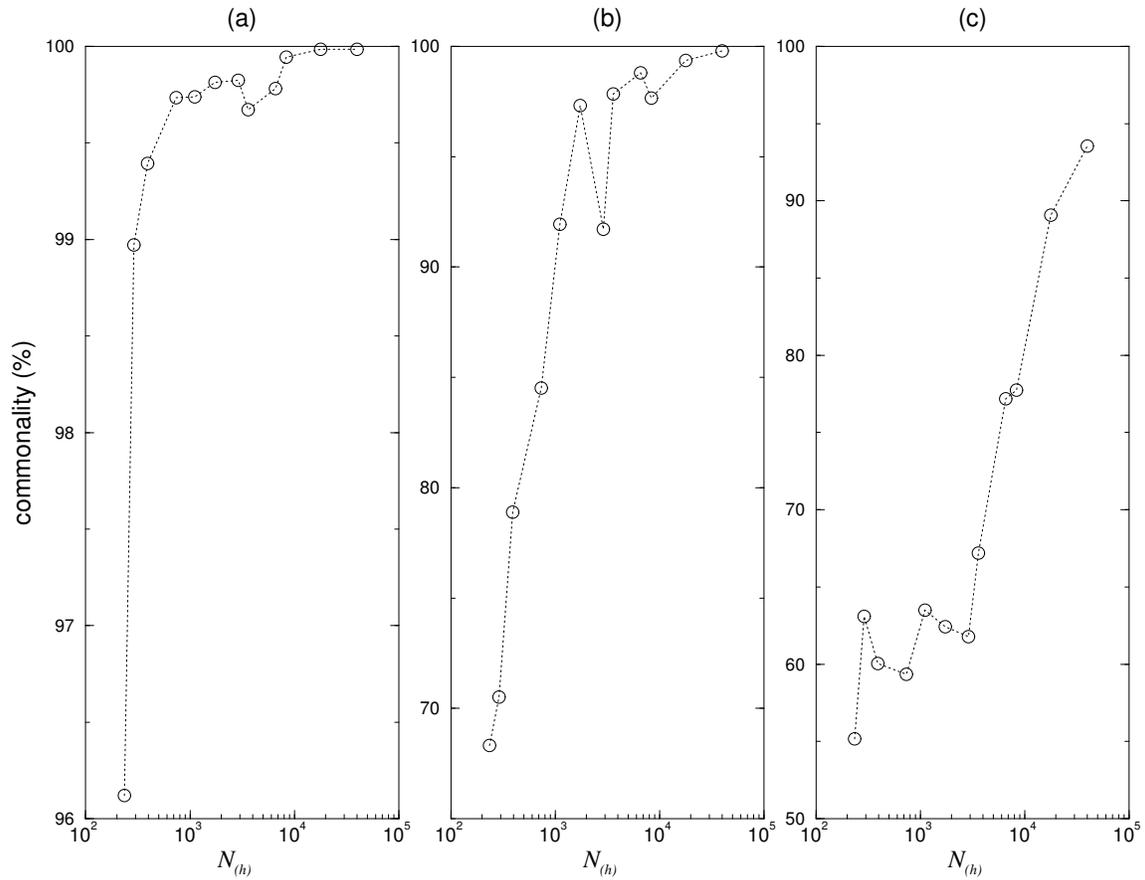
Figure 2.4 The commonality measurement of the subspaces formed by the set of eigenspectra derived using different numbers of observed galaxy spectra where the subspaces are spanned by the first (a) 2 (b) 3 and (c) 10 eigenspectra respectively. The results show that the eigenspectra set converge as a function of learning-set size in the KL gap-correction formalism.

of including galaxies with relatively extreme spectral properties. When we randomly pick about 1000 galaxies from the sample there are a few emission line galaxies. As more spectra are included we eventually reach a threshold where we begin to sample the extreme emission line galaxies. Once we have included a small number of these emission line galaxies (with a sample size of $3000 - 4000$ galaxies we would expect three to four emission line galaxies) the information they contain is now incorporated within the KL eigenbases. The dramatic increase in the convergence rate come from the fact that, while rare, these extreme emission line galaxies can still be described by a handful of spectral components (i.e., once we have a small number of them in the sample we can map out their full distribution).

To conclude, there exist a minimum number of galaxy spectra we need to observe in order to derive a convergent set of eigenspectra. Approximately 10^4 spectra are sufficient for a 90% convergence level with ten eigen-components (Fig. 2.4c). This is sufficient to characterize the spectral types of 99.9% of galaxies within the local universe. These results are, however, purely empirical, based on randomly selecting spectra from the current data set. Thus, there is no concrete evidence to support the present result that 10^4 galaxy spectra are all we need in deriving the most complete set of eigenspectra. There may exist populations of galaxies that comprise much less that 0.1% of the full galaxy sample that our current analysis is not sensitive to. In general, for a larger data set (e.g., at the completion of the survey), new galaxy types, if any, may call for more spectra to be included when constructing the eigenspectra.

2.4 ECLASS AND GALAXY PROPERTIES

The first 10 KL eigenspectra of the 170,000 SDSS galaxies are shown in Fig. 2.5, derived from 20 iterations and using 10 eigenspectra for gap repairing. The eigenspectra are publicly available (from the website http://www.sdss.org). The first eigenspectrum is the mean of all galaxy spectra in our sample. The continuum is similar to a Sb-type [1]. As we would expect

---

[1]In our work, the red- and blue-types are determined from the spectral information in the galaxies. Thus, the conventional morphological-type nomenclatures "early", "late", and E, S0, types etc. used in this thesis are referring to *spectral features* which usually would have be seen in the corresponding morphological types,
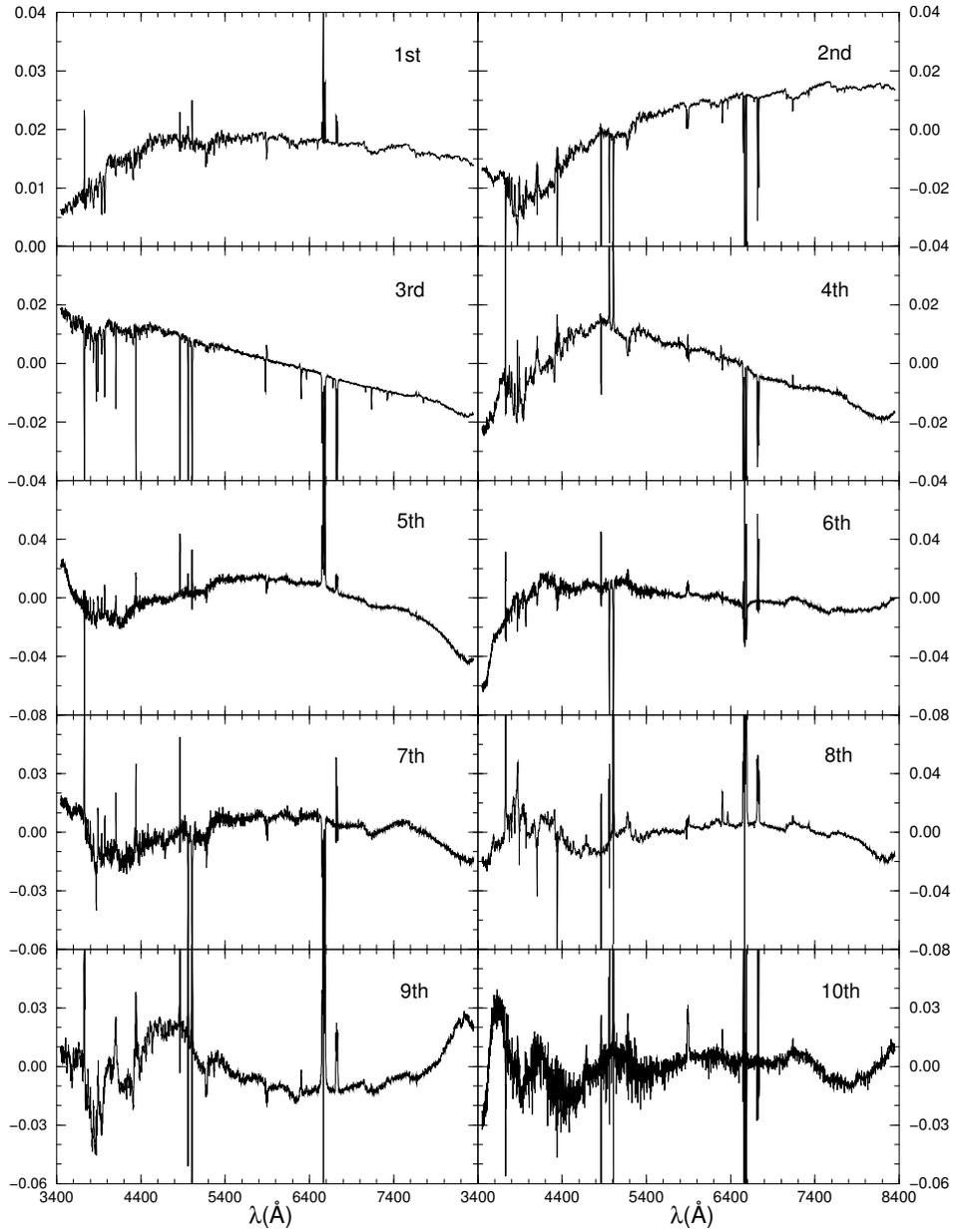
Figure 2.5 The first 10 KL eigenspectra of $\sim$ 170,000 galaxy spectra in the SDSS. Gap correction is implemented for 20 iterations.

Table 2.1 The relative weights (normalized by total weight) of the eigenspectra. The first 3 eigenspectra comprise about 98% of the total sample variance.

m (number of modes)	weight
1	0.9594
2	0.9784
3	0.9815
5	0.9837
8	0.9849
10	0.9852
20	0.9855
50	0.9860
100	0.9867
500	0.9908
1000	0.9940

from the mean of all spectra, nebular lines and other emission lines, as well as absorption lines such as Ca H and Ca K, are present within this spectrum. The second eigenspectrum has one zero crossing, positive toward longer wavelengths, at around 5200Å, which marks the wavelength at which the modulation in the continuum level relative to the 1st eigenspectrum occurs. In the third component, there is a zero crossing in the continuum, negative toward longer wavelengths, at around 6000Å. The higher the order of the eigenspectrum, the larger the number of zero crossings which in turn adds high-frequency features to the final spectrum as these higher order modes are added or subtracted. In the higher-order modes, the eigenspectra are dominated by emission and absorption lines because each of these eigenspectra comprises emission and absorption lines plus a small fluctuation of the continuum level around zero. We illustrate this point later.

Statistically, the amount of information contained in each eigenspectrum is given by the eigenvalue of the correlation matrix of that particular mode. Table 2.1 lists the weights of the first m-modes of eigenspectra, where the weights are normalized to unity. We find that the first three eigenspectra contain more than 98% of the total variance of the data set.

It is known that there is a one-parameter description of galaxy spectra which correlates with the spectral type of a galaxy (Connolly et al. 1995). This parameter, ϕ_{KL}, is the mixing

---

as suggested in Kennicutt's Atlas and other studies.

angle of the first and second eigencoefficients. Explicitly,

$$\phi_{KL} = \tan^{-1}(a_2/a_1) \tag{2.7}$$

where a_1 and a_2 are the eigencoefficients of the first and second modes of a galaxy respectively. Furthermore, the inclusion of the third component discriminates the post-starburst activity (Connolly et al. 1995; Castander et al. 2001). To follow this classification scheme, we define here

$$\theta_{KL} = \cos^{-1}(a_3) \tag{2.8}$$

where a_3 is the third eigencoefficient. Here we adopt the normalization

$$\sum_{k=1}^{10} a_k^2 = 1 \ . \tag{2.9}$$

The first three eigencoefficients of the whole sample are plotted in Fig. 2.6 and 2.7 in the forms of a_2 versus a_1 and a_3 versus a_2. More than 99% of the total galaxy population is located on the locus in Fig. 2.6, in which the second eigencoefficients have values from ~ 0.25 to -0.75. The appearance of this locus is very similar to previous works (Connolly et al. 1995). Red galaxies have positive, and relatively large second eigencoefficients, while blue galaxies have smaller, or in some cases negative values. From Fig. 2.7 we clearly see that by introducing the third eigen-component, there is a group of galaxies being separated out from the main group. These galaxies, with negative a_3 and negative a_2 values, exhibit post-starburst activity in their spectra. A much smaller group (about 0.1%) with positive a_3 and negative a_2 values is also seen. These are outliers and will be discussed later. The resulting ϕ_{KL} versus θ_{KL} is plotted in Fig. 2.8 for all galaxies in the sample excluding those galaxies with $a_1 < 0$ (118 objects). These $a_1 < 0$ sources tend to be either artifacts within the data (Mark SubbaRao, private communication) or spectra that are not visually confirmed as a galaxy spectrum. In Fig. 2.8a, the sequence from red to blue to extreme emission line galaxies is illustrated. The boxes drawn show the regions from which a set of spectral types are identified. They range from the early-type at the top of the plot to emission line galaxies at the bottom. The spectra for these subsamples are shown in Fig. 2.9(a-f), ranging from red to emission line galaxies. The spectral energy distributions shown are the mean

of all the *observed* spectra classified to be in the range (ϕ_{KL}^s, ϕ_{KL}^e, θ_{KL}^s, θ_{KL}^e), where the superscripts ^{s} and ^{e} denote the starting and ending values bounding the range. The actual values are chosen such that the resultant mean spectra agree with the galaxy spectra of each type in Kennicutt's atlas of nearby galaxy spectra (Kennicutt 1992) by visual inspections for features described in the atlas. The flux levels are in good agreement with those in the atlas, which leads us to believe that the classification is physically sound as well as having statistical rigor. Spectra with similar spectral features are therefore seen to be clustered by the KL procedure. Due to the smoothing of spectral inhomogeneities with a large number of galaxies, the resultant mean spectra have very high signal-to-noise levels. This result demonstrates the power of the KL transform for calculating mean (or composite) spectra (e.g., Eisenstein et al. (2003) for the mean spectrum of the SDSS massive galaxies).

Table 2.2 shows the number of observed galaxy spectra in each of the regions described previously. We stress that the sum of all galaxies listed in the table is not equal to the total number of galaxies in the data set because the ranges chosen comprise a subset of the full (ϕ_{KL}, θ_{KL})-plane. The early late- to intermediate late-types galaxies (with $-12^o < \phi_{KL} < 5^o$, $80^o < \theta_{KL} < 100^o$) dominate within the whole data set which agrees with the well-known fact that late-type galaxies dominate the field populations in terms of number counts.

Apart from the main locus in Fig. 2.8 the region marked "(b)" identifies a group of outliers, forming approximately 0.1% of the full sample (190 sources). These are unusual sources that arise due to artifacts within the reduction pipelines, errors within the spectra themselves or possibly due to new classes of astrophysical sources. In later processing runs (idlspec2D v4.9.8, as of 13th of August, 2002) only 68 of these sources remain in the main galaxy sample. Approximately half of them have the ZWARNING flag set to 4, which indicates that there are large errors in the redshift estimations. This results in less than 0.02% of the spectroscopic sample having spectra that can be considered unphysical (a testament to the remarkable accuracy and performance of the current spectroscopic reduction pipelines). Considering all of these sources as a whole 90% have signal-to-noise ratios (S/N) higher than the mean survey quality (the $\langle S/N \rangle$ is 15.9 in the data set).

Of the remaining 30 galaxies within this outlier class, most have relatively high redshifts ($z \sim 0.2 - 0.5$) as assigned by the pipelines. Some of these sources classed as galaxies by the

pipelines do not appear to be galaxy spectra when inspected visually. On the other hand, for those that *are* galaxies as inspected by us, we found that the pipelines have assigned incorrect redshifts to some of these spectra. As expected, the gap-repairing procedure fails in those objects and the resulting expansion coefficients have unphysical values. An example is shown in Fig. 2.10, according to the assigned redshift, this object has a redshift 0.5394, which is obviously incorrect from the locations of the [N II]+Hα lines, as shown in the insert (this galaxy should have a redshift of 0.0236). The outcome is that the magnitudes of the 2nd and the 3rd eigencoefficients obtained by a KL of all the objects in this group are roughly the same but with different *signs*, meaning that no lines that are representative of typical spectral types are found. This result suggests that KL technique is a powerful tool for identifying artifacts within any spectral reduction procedure.

The above results show that the classification is successful in allowing the galaxy types to be identified using the first three eigencoefficients and that it may serve as a way for error checking. How do the eigenspectra actually perform in reconstructing the spectra? Fig. 2.11(a-f) shows, for the same range of $(\phi_{KL}^s, \phi_{KL}^e, \theta_{KL}^s, \theta_{KL}^e)$ as above, the means of all KL-reconstructed spectra. A KL-reconstructed spectrum, using m-eigenspectra, is given by

$$f^R(m;\lambda) = \sum_{k=1}^{m} a_k e_k(\lambda) \ . \tag{2.10}$$

It should be noted that the KL-reconstructed spectrum is different from the KL-repaired spectrum we mentioned previously (in that case the repairing is in the gap regions only). For convenience, the *mean* of all KL-reconstructed spectra in a given range is abbreviated as "KL-reconstructed spectrum" in the following sections unless otherwise specified. Comparing the mean spectra in Fig. 2.9(a-f) with the reconstructed ones in Fig. 2.11(a-f) (3 modes are used), the continuum levels and most emission lines are in excellent agreement with the mean spectra (except for the galaxies with extreme emissions, which we will discuss later). These results are consistent with previous claims that two eigenspectra are enough to describe most of the spectral types in galaxies (Connolly et al. 1995). Our present classification scheme of using two mixing angles of the first three eigencoefficients is also justified.

The 3-mode KL-reconstructed spectra shown in Fig. 2.11(a-f) also suggest that to reconstruct some of the lines and line ratios, more eigenspectra are necessary. Fig. 2.12 shows in

detail the emission lines that require more than 3 modes for reconstruction. These figures show early-type to late-type galaxies (from top to bottom). With the first eight eigenspectra, the amplitude of the [N II] line for galaxies with classification angles in the range (7.5, 20, 86, 92) can be correctly recovered. Similarly, the first four modes are sufficient for the [N II] and Hα reconstruction for galaxies with classification angles in the range (5, 6, 80, 100). Progressing to bluer galaxies with classification angles in the range (0, 2, 80, 100), the first eight modes are enough for the [O III] reconstruction. Similarly, Fig. 2.13 shows the cases for galaxies with prominent emission features. We find that the first four modes are enough to reconstruct the amplitudes and line-ratios [O III]λ5008/[O III]λ4960 and H$\beta\lambda$4863/[O III]λ4960 for those with classification angles in the range (-12, -8, 80, 100). For galaxies in the range (-60, -40, 120, 135), the line-ratio [N II]λ6585/[N II]λ6550 is correct with three eigenspectra, while the first eight modes are enough to further retrieve the amplitudes of the two [N II] lines. The maximum differences between the amplitudes of the mean and the reconstructed lines (which we define as the error of the reconstruction of a particular line) in the above-mentioned cases are about 10%.

Therefore, for all but the most extreme emission line galaxies, eight spectral components, or modes, are sufficient to reconstruct the spectral line ratios to an accuracy of about 10% (a factor of 500 in compression of information within the galaxy spectra). For the reconstruction of galaxies with extreme emissions, however, the performance is not satisfactory when using a small number of eigenspectra. Nevertheless, ten eigenspectra are sufficient to recover the continuum level (see the mean and KL-reconstructed spectra in the enlarged continuum region). The residuals of the mean spectra and the KL-reconstructed spectra are shown in Fig. 2.14, where (a-d) correspond to the reconstructions with 3, 4, 5 and 10 eigenspectra respectively. There are substantial improvements in using ten eigenspectra, especially in nebular lines and S II lines, and various line ratios. The typical errors in the fluxes of lines remains around $15 - 25\%$.

This is not a surprising result. On one hand, the result follows because of the increasingly dominant role of lines in the higher-order modes compared with the continuum. On the other hand, statistics also play a factor. The early and intermediate-type galaxies dominate the population of galaxies while emission and extreme emission line galaxies comprise just a

few percent of the total population. Thus, galaxies with significant emission call for more eigenspectra and higher-order modes in their reconstructions. Besides the statistical reasons, the inevitable variations in line-widths of emission lines make it comparatively difficult to reconstruct them accurately using linear combinations of eigenspectra.

Due to the fact that the spectral features in extreme emission line galaxies are distinct from other types of galaxies, they still reveal themselves in the plane (ϕ_{KL}, θ_{KL}). Thus, for the main purpose of this work, which is obtaining a robust and objective classification of galaxies, the less satisfactory performance of reconstructing some emission lines in galaxies with extreme emission lines does not have a significant effect. However, if detailed diagnosis of lines (for example, the flux-ratio of two lines) in those galaxies are of interest, then more modes are needed. Better yet, a separate analysis using KL with those emission line galaxies is suggested.

2.4.1 Kennicutt's atlas of nearby galaxies

We perform least square fittings of the mean galaxy spectra to the spectral atlas of nearby galaxies by Kennicutt (1992). The atlas is chosen because it covers a range of spectra (54 in total) from the early to the late types, each with a morphological classification.

Best fit spectrum from the atlas is found by minimizing $\chi^2 \equiv \sum_\lambda w_\lambda (f_\lambda^g - f_\lambda^k)^2$, where the superscripts g and k represent that the flux density f is from the mean galaxy spectrum and from the spectrum in Kennicutt's atlas respectively. The fitting is done with the inclusion of absorption or/and emission lines. The weight w_λ is set to unity for each pixel. Each spectrum in the atlas is rebinned to vacuum wavelengths to comply with our wavelength convention.

In Fig. 2.15 to 2.20 we overlay the best-fit spectrum onto the mean galaxy spectra. The object identity is printed on each figure.
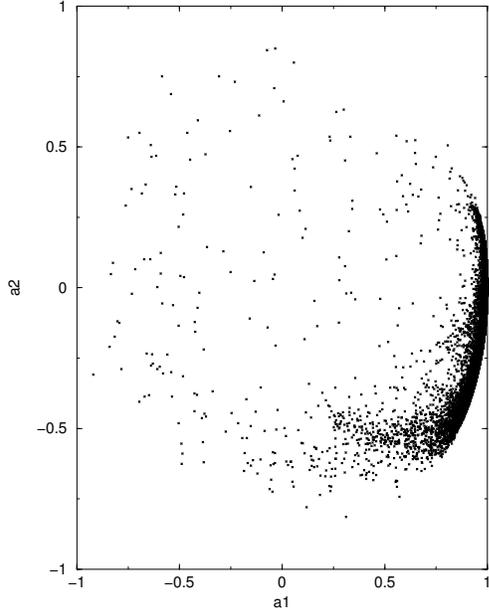
Figure 2.6 Eigencoefficients a_2 versus a_1 of our sample ($\sim$ 170,000 galaxy spectra). More than 90% of the whole galaxy population are located on this main locus. The trend is similar to that in previous works (Connolly et al. 1995), with red galaxies having larger, positive a_2 values and blue galaxies having smaller, or negative values.
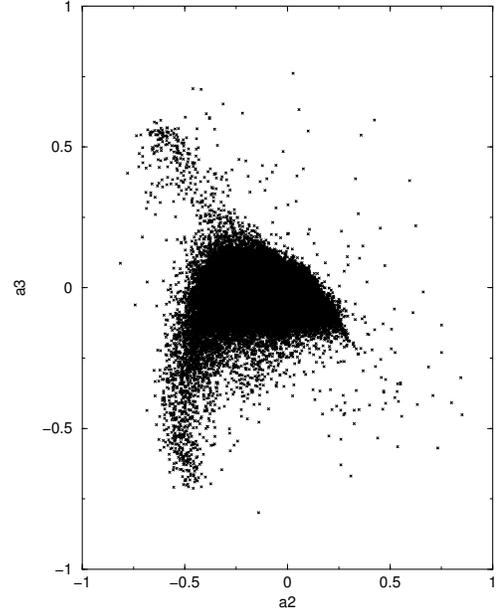
Figure 2.7 Eigencoefficients a_3 versus a_2 of our sample ($\sim$ 170,000 galaxy spectra). The introduction of the third eigencoefficient further discriminates galaxies with post-starburst activity (they are the group of galaxies with negative a_3 and a_2 values in this plot). Also, a group of outliers are apparent (a small group of objects with positive a_3 and negative a_2 values) which are explained in the text.

Table 2.2 The number of galaxies in the range $(\phi_{KL}^s, \phi_{KL}^e, \theta_{KL}^s, \theta_{KL}^e)$. These data are a subset of the full sample. The galaxy types listed are the possible morphological types, estimated by comparing the spectral features of the mean spectrum constructed in each range with spectra in Kennicutt's atlas and therefore they are for reference only.

$(\phi_{KL}^s, \phi_{KL}^e, \theta_{KL}^s, \theta_{KL}^e)$	Galaxy Type	Number [Number ratio relative to Sa]
(7.5, 20, 86, 92)	E/S0	6599 [0.34]
(5, 6, 80, 100)	Sa	19543 [1.00]
(0, 2, 80, 100)	Sb	13872 [0.71]
(-12, -8, 80, 100)	Sbc/Sc	11979 [0.61]
(-40, -30, 80, 100)	Sm/Im	140 [0.0072]
(-60, -40, 120, 135)	SBm	135 [0.0069]
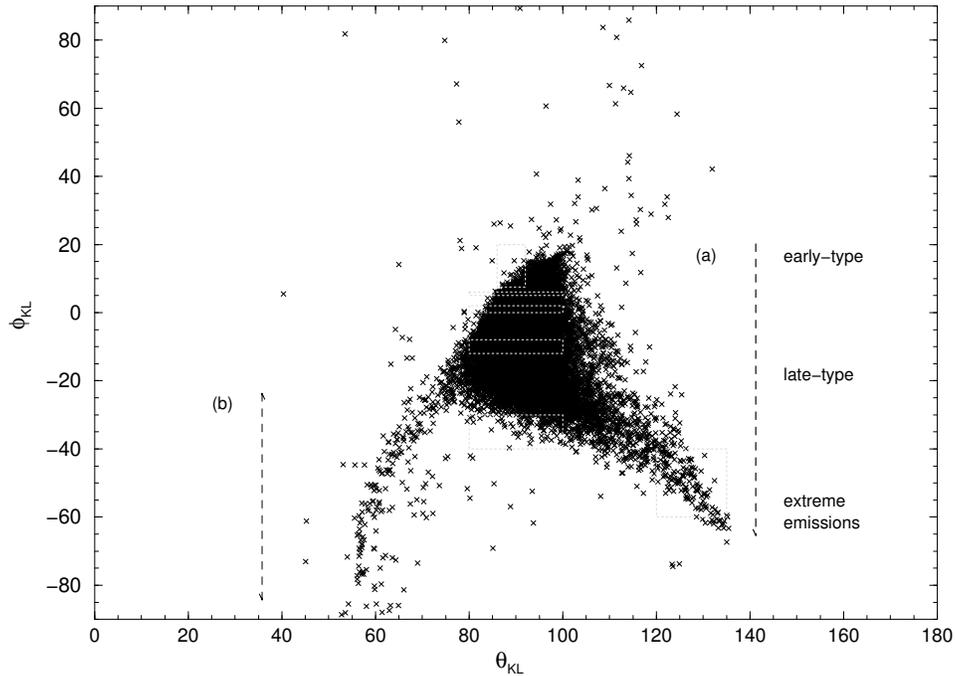
Figure 2.8 (ϕ_{KL}, θ_{KL})-classification of $\sim$ 170,000 SDSS galaxy spectra. (a) illustrates the sequence along which the galaxy spectral types are identified. (b) Outliers, mostly spectra without significant spectral features. The angles are in degrees. Most outliers have large errors in their redshift estimations, while 90% have low signal-to-noise ratios. The boxes are areas in which the mean of all of the observed spectra correspond to red, blue and emission line galaxies. See Fig. 2.9 for the mean spectra.
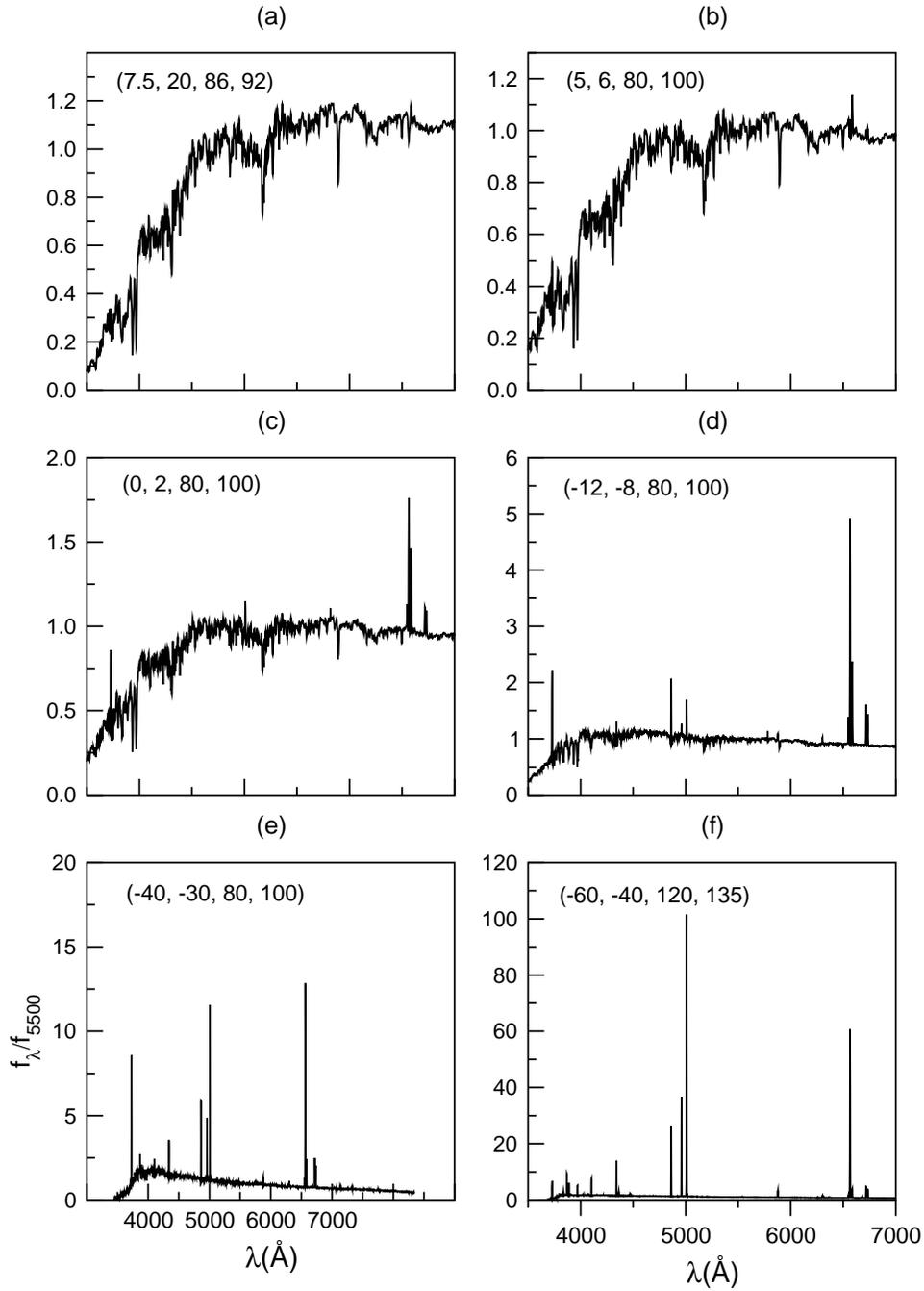
Figure 2.9 Mean of all the observed spectra in different ranges of $(\phi_{KL}^s, \phi_{KL}^e, \theta_{KL}^s, \theta_{KL}^e)$, with the classification angles being (a) (7.5, 20, 86, 92), (b) (5, 6, 80, 100), (c) (0, 2, 80, 100), (d) (-12, -8, 80, 100), (e) (-40, -30, 80, 100) and (f) (-60, -40, 120, 135) are shown.
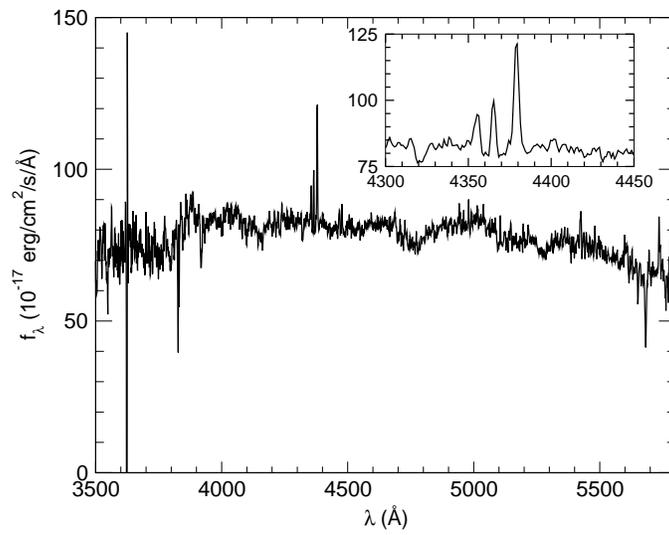
Figure 2.10 One of the outliers in the (ϕ_{KL}, θ_{KL})-plane. The redshift of this galaxy is incorrectly assigned by the spectroscopic pipeline.
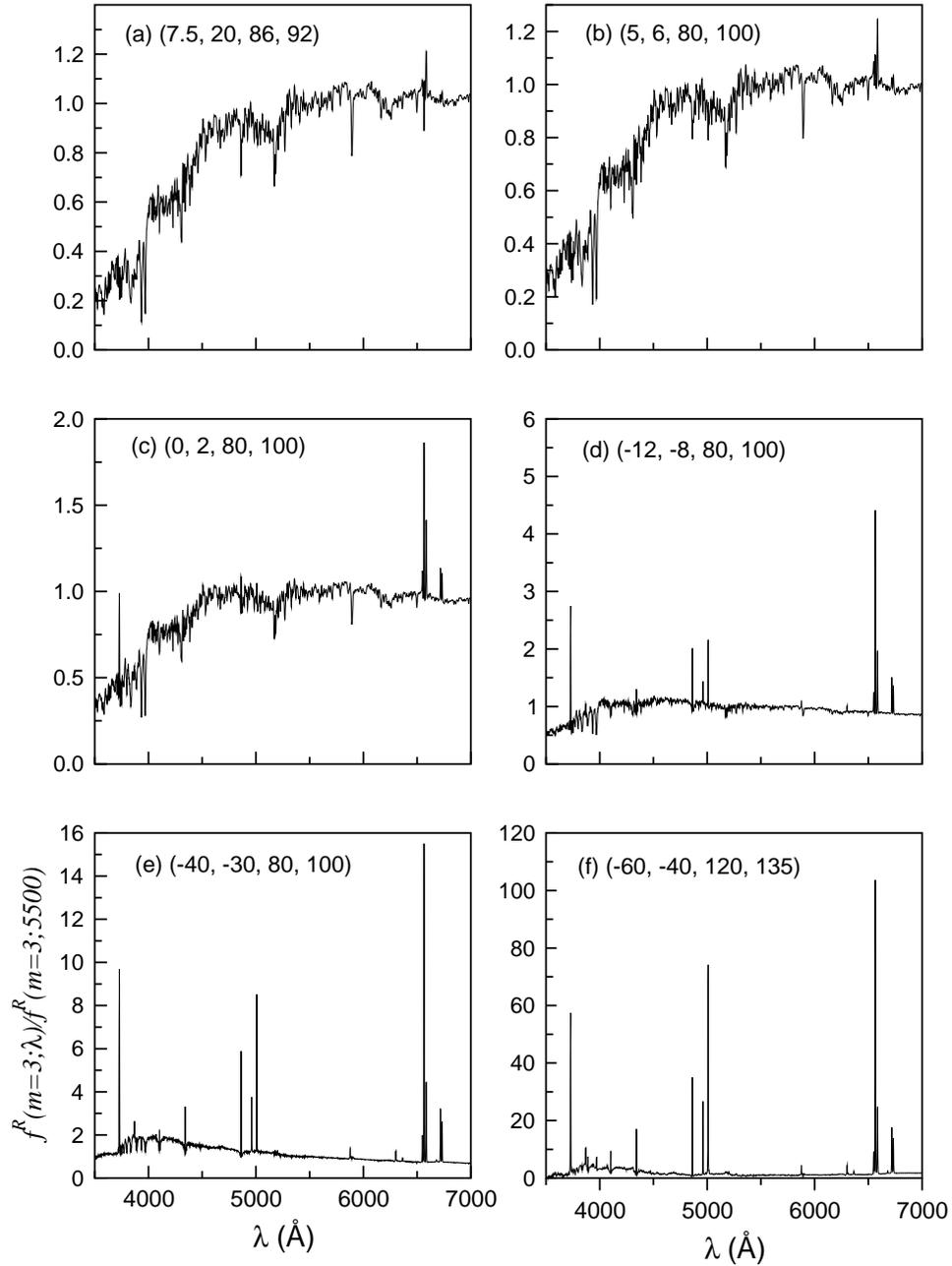
Figure 2.11 The means of all KL-reconstructed spectra (a-f) in different ranges of $(\phi_{KL}^s, \phi_{KL}^e, \theta_{KL}^s, \theta_{KL}^e)$ (the bounding boxes are the same as those in Fig. 2.9). The first three eigenspectra are used in the reconstruction. The continua and most of the line features are in excellent agreement with those of the mean spectra shown in Fig. 2.9.
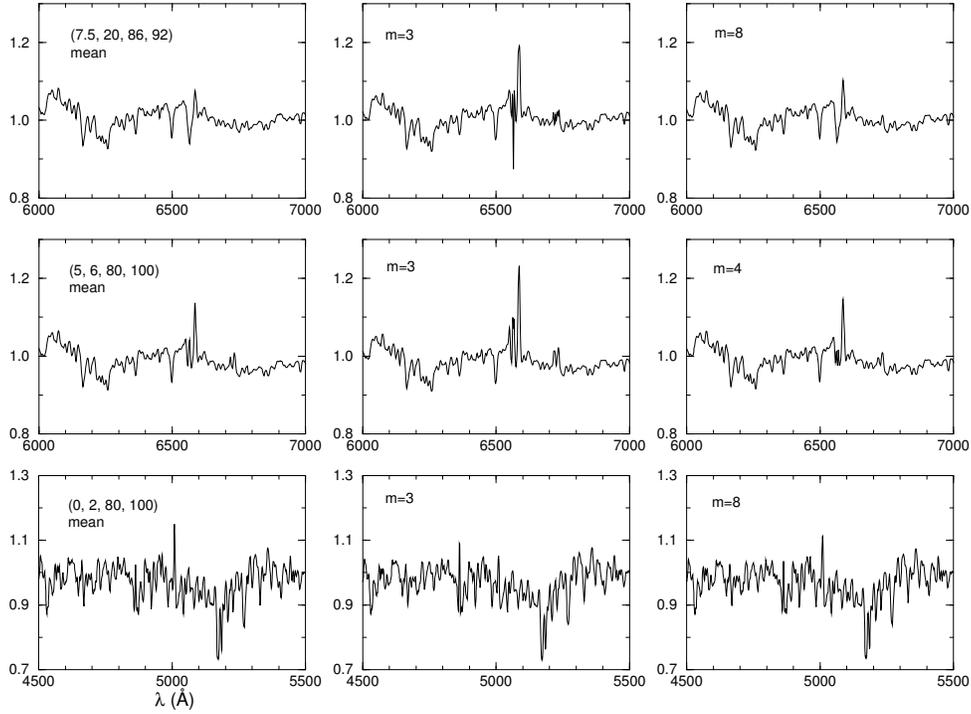
Figure 2.12 Number of modes needed to reconstruct some of the lines in early-type to early late-type galaxy, with the classification angles in the ranges (7.5, 20, 86, 92) (top panel), (5, 6, 80, 100) (middle panel) and (0, 2, 80, 100) (bottom panel). The figures on the leftmost panels show the mean spectra in each type, and the consecutive figures show the KL-reconstructed spectra with different numbers of modes. All spectra are normalized at 5500Å.
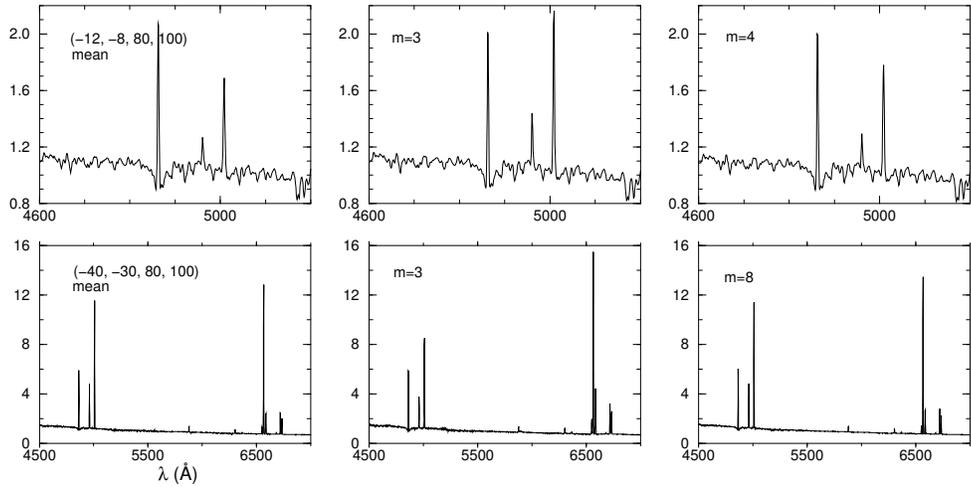
Figure 2.13 Same as Fig. 2.12, but for galaxies with the classification angles in the ranges (-12, -8, 80, 100) (top panel) and (-40, -30, 80, 100) (bottom panel).

36
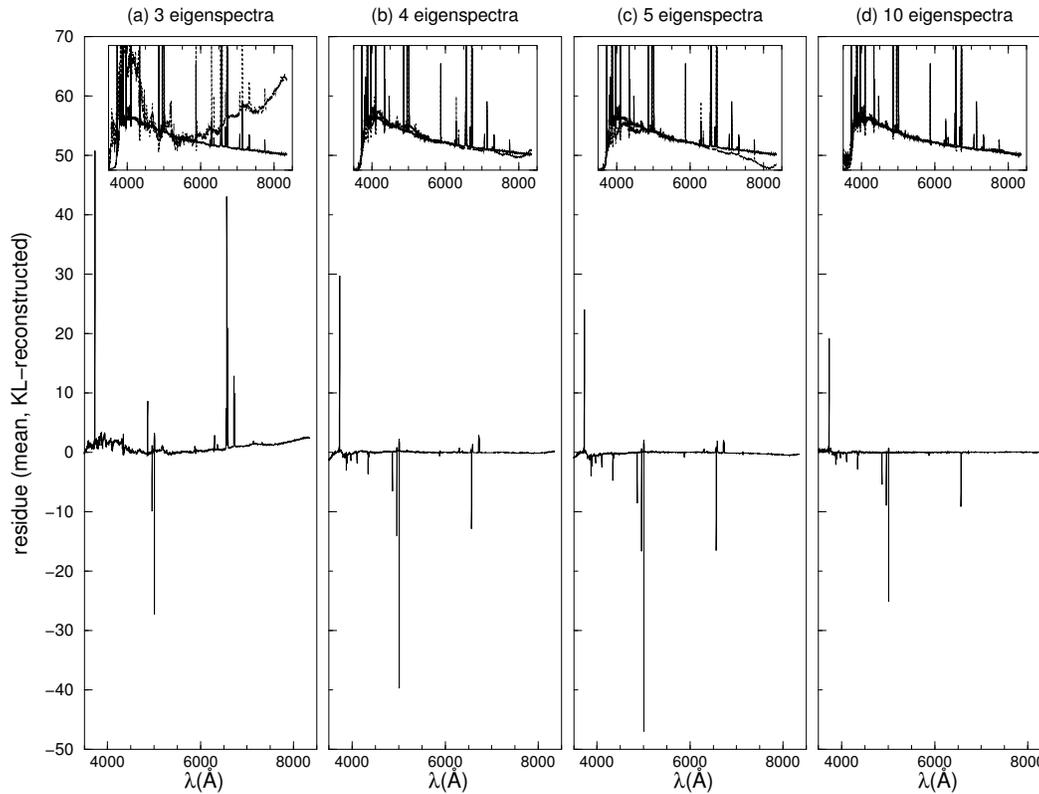
Figure 2.14 The residuals of the mean spectrum of extreme-emission galaxies (classification angles in the range (-60, -40, 120, 135)) with the KL-reconstructed spectrum using (a) 3, (b) 4, (c) 5, and (d) 10 eigenspectra. The inserts are the enlarged regions of the continuum levels, in each case the solid line is the mean spectrum and the dotted line is the mean of the KL-reconstructed spectra. The spectra are normalized at 5500Å.
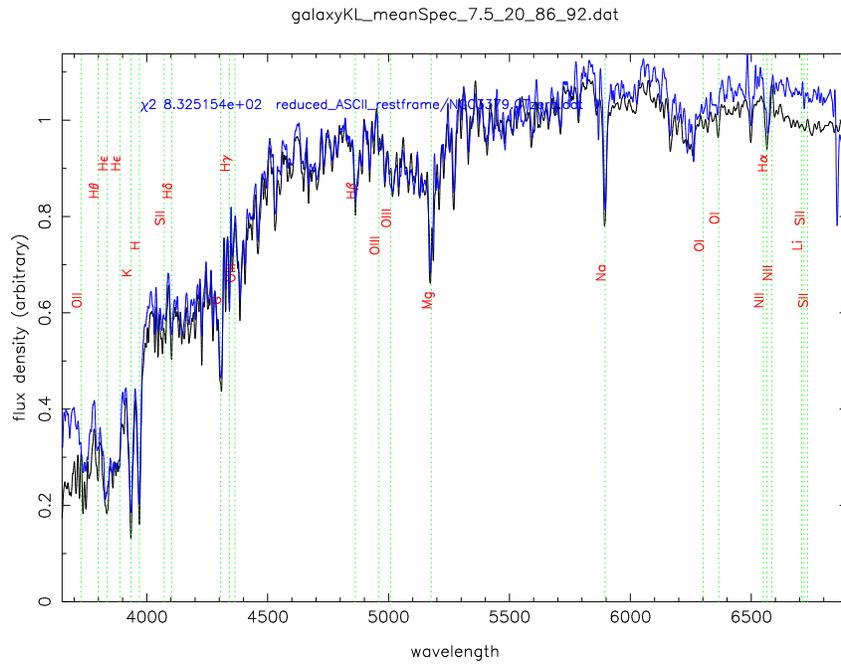
Figure 2.15 Black: mean galaxy spectrum for eClass type A. Blue: best fit spectrum in Kennicutt's atlas.
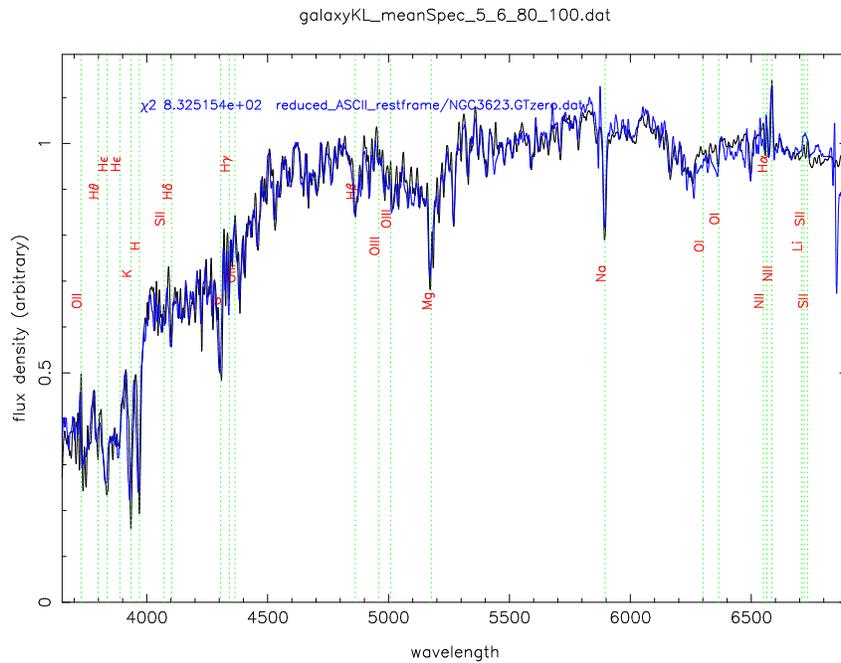
Figure 2.16 Black: mean galaxy spectrum for eClass type B. Blue: best fit spectrum in Kennicutt's atlas.
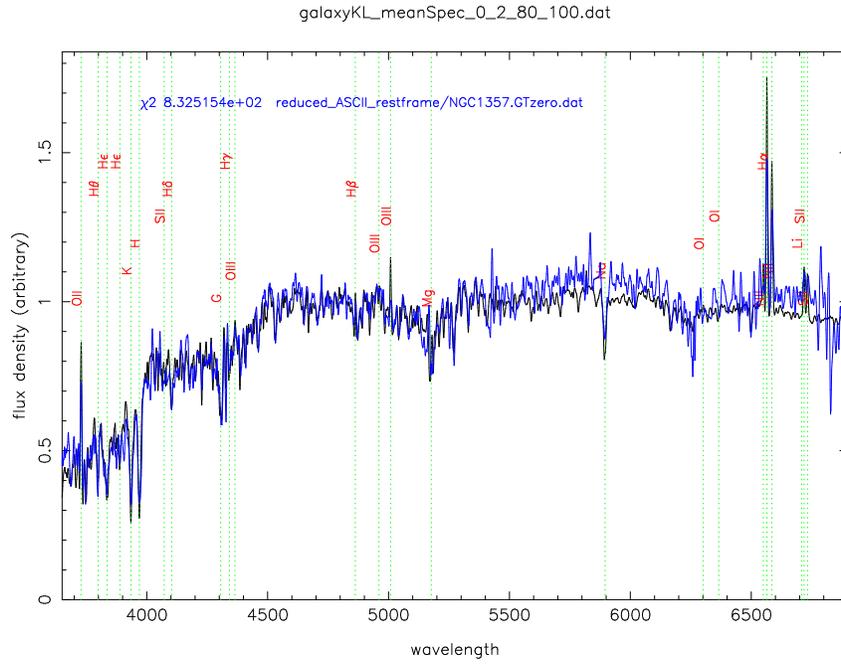
Figure 2.17 Black: mean galaxy spectrum for eClass type C. Blue: best fit spectrum in Kennicutt's atlas.
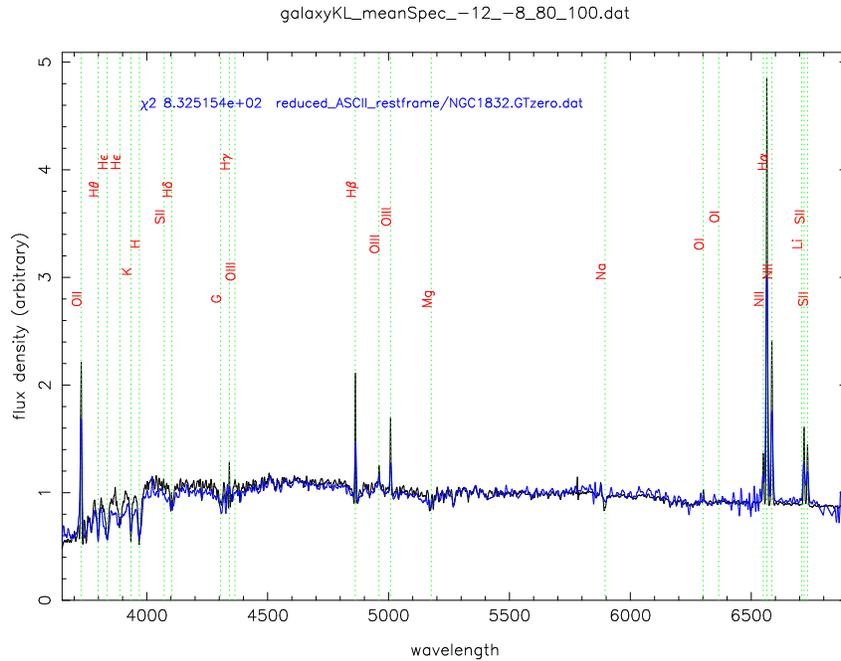
Figure 2.18 Black: mean galaxy spectrum for eClass type D. Blue: best fit spectrum in Kennicutt's atlas.
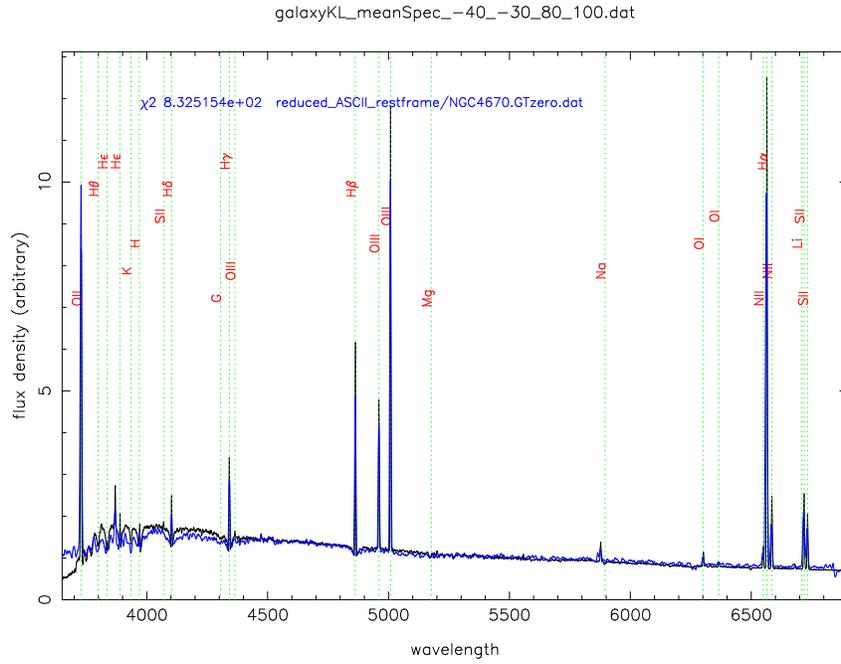
Figure 2.19 Black: mean galaxy spectrum for eClass type E. Blue: best fit spectrum in Kennicutt's atlas.
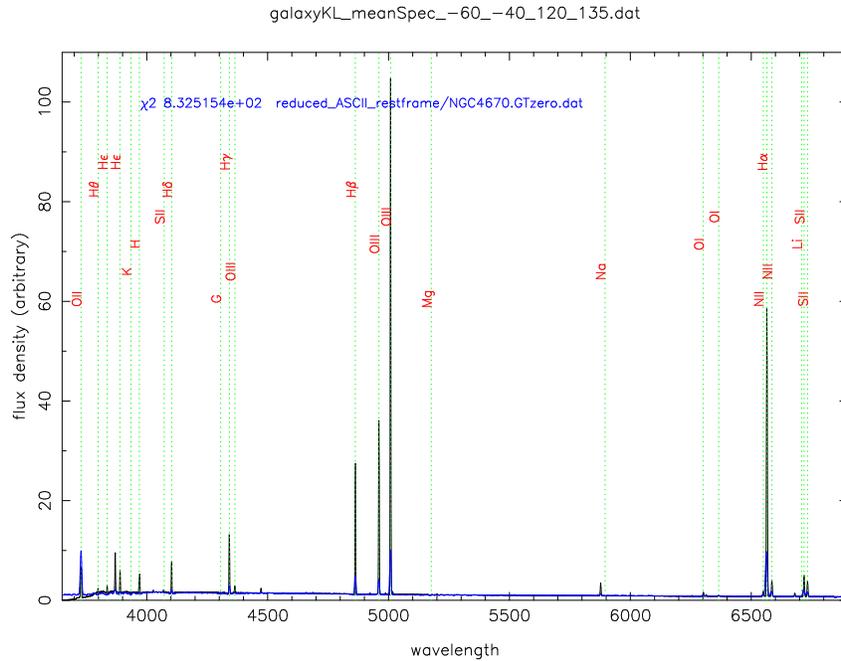
Figure 2.20 Black: mean galaxy spectrum for eClass type F. Blue: best fit spectrum in Kennicutt's atlas.

2.5 EFFECT OF FIXED APERTURES

The SDSS uses a fixed aperture of $3''$ diameter for its spectroscopic observations. This can, in principal, lead to biases in the current spectral classification scheme if, for example, a fiber samples only the central bulge of a nearby intermediate or late-type galaxy (resulting in the assignment of an early-type spectral class). As this effect depends on the apparent size of a galaxy when compared to the fiber diameter it has the potential to induce redshift and luminosity dependent biases in any analysis using the KL classifications (Kochanek et al. 2000). Studies of the effect of aperture bias on observed parameters (e.g., star-formation rate) can be found in, e.g., Baldry et al. (2002); Pérez-González et al. (2003); Brinchmann et al. (2004). The questions we address here are: (i) is there an aperture bias using the KL approach (ii) how can we quantify this bias and (iii) can we correct for the aperture effects to obtain bias-free galaxy types?

We estimate the effect of aperture bias by calculating, for a given galaxy, the difference in the classification (in this case the ϕ_{KL} angle) derived from the total galaxy flux compared to that derived from the central $3''$. The dependence of this classification error on the redshift and the physical size of the galaxy serves to quantify the bias in our sample. We assume that the apparent diameter of each galaxy can be approximated as twice the Petrosian half-light radius (petro50) in the r-band. The physical sizes of the galaxies are then calculated by assuming $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ and $H_0 = 71$. The aperture magnitudes of all galaxies are initially k-corrected to redshift $z = 0.1$ using the code by Baldwin (1977) version 1.16 prior to estimating the spectral types. Type assignment for the total flux and fiber flux is performed using the photometric redshift code of Connolly et al. (1999). The input spectral templates are constructed as linear-combinations of the first 3 eigenspectra from this work, with the resolution in both ϕ_{KL} and θ_{KL} set to 2°. In the following discussion we will express the distance dependence of the relation as function of z/z_{max}, where z_{max} is the highest redshift at which a galaxy of a given absolute magnitude would pass the sample selection criteria. This provides a pseudo volume independent analysis.

Fig. 2.22 shows the difference in the classifications of galaxies, $\phi_{KL}(\text{total}) - \phi_{KL}(3'')$ ($\equiv D\phi_{KL}$), as a function of z/z_{max} and galaxy type. The bin sizes of smoothing are 0.02 in
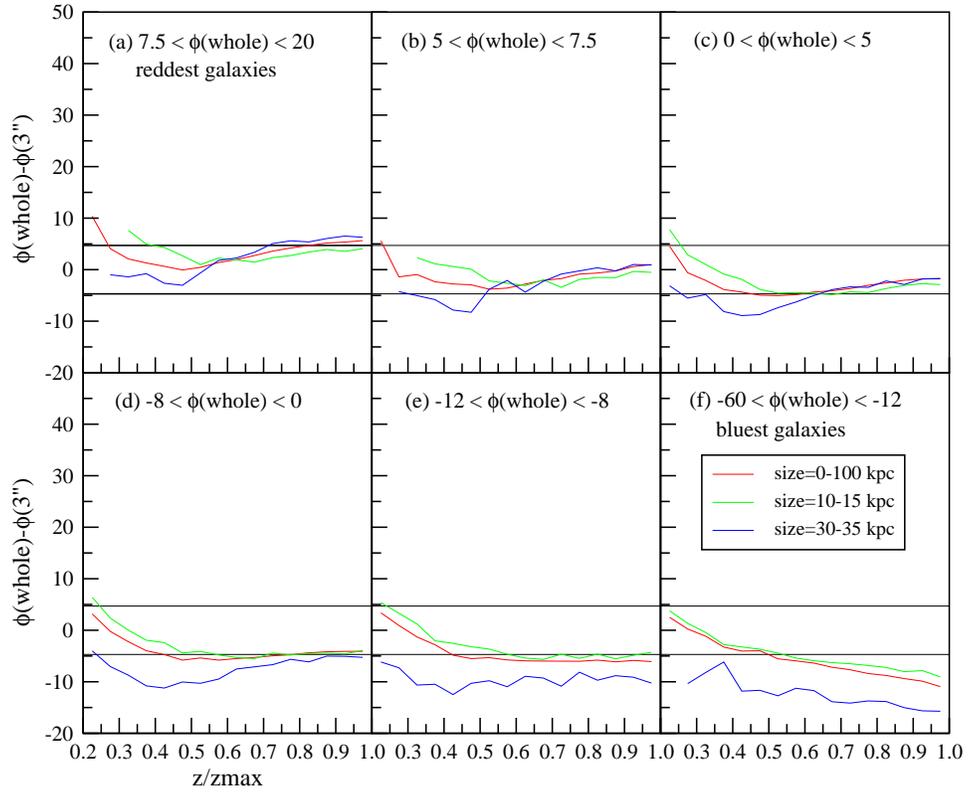
Figure 2.21 The difference in the classification for the whole galaxy and the inner $3''$ region as a function of z/z_{max} from the reddest (a) to the bluest (f) galaxies. In each sub-figure the galaxies are of sizes $0 - 100$ kpc (red), $10 - 15$ kpc (green) and $30 - 35$ kpc (blue).

42
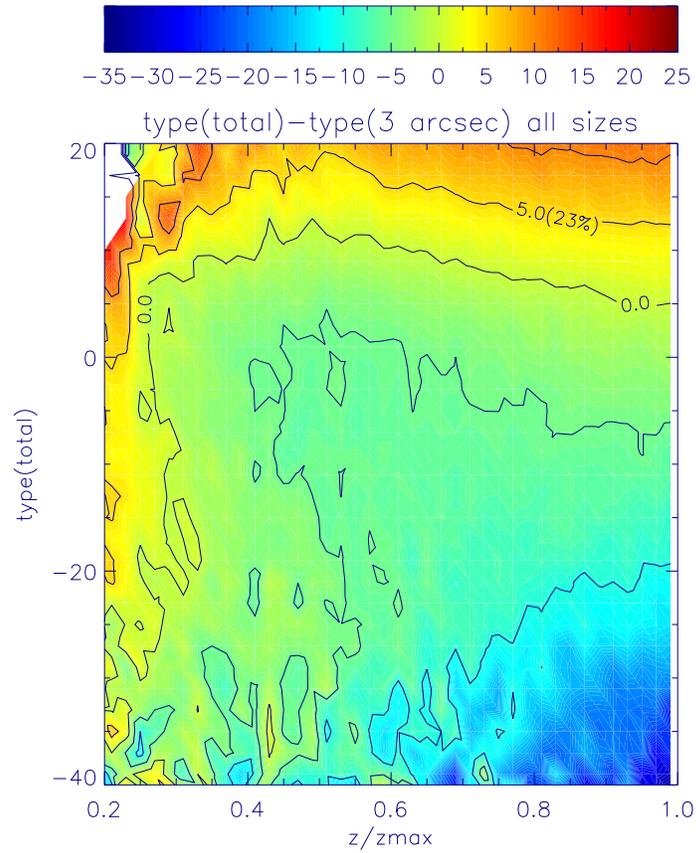
Figure 2.22 The contours of the difference in the classification ($D\phi_{KL}$) for the total flux and for the inner $3''$ region as a function of z/z_{max}. The ordinate is the type assigned from the total flux.
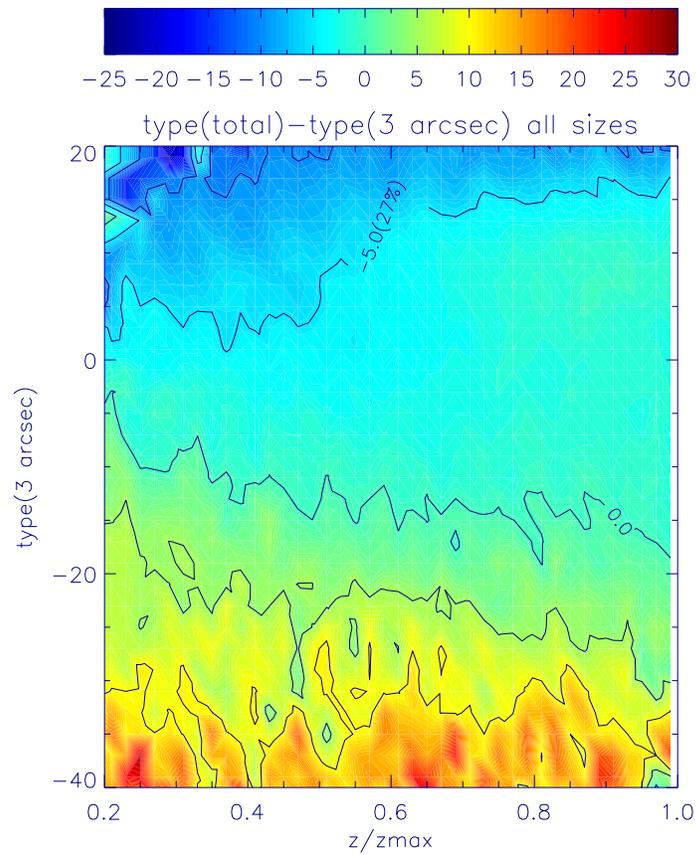
Figure 2.23 Same as Fig. 2.22 except that the ordinate is the type for the inner 3″ of each galaxy. This serves as the look-up table with which the aperture bias can be corrected.

Table 2.3 The number of galaxy spectra in our sample with the specified values of aperture bias $D\phi_{KL}$.

$D\phi_{KL}$-range (degree)	fraction of galaxies (0.1% accuracy)
$-20 < D\phi_{KL} <= -10$	18.5 %
$-10 < D\phi_{KL} <= -5$	16.4 %
$-5 < D\phi_{KL} <= -2.5$	9.2 %
$-2.5 < D\phi_{KL} <= 0$	17.7 %
$0 < D\phi_{KL} <= 2.5$	7.3 %
$2.5 < D\phi_{KL} <= 5$	6.1 %
$5 < D\phi_{KL} <= 10$	12.1 %
$10 < D\phi_{KL} <= 20$	7.2 %

z/z_{max} and 2° in $D\phi_{KL}$. Galaxies of sizes from $0 - 100$ kpc are included, whereas galaxies of ϕ_{KL}(total)$< -40^o$ are excluded for there are less then 1% of them. Redder-hue components in the contour image correspond to the fraction of galaxies that would be classified as an earlier type (i.e., redder) if the total flux was used rather than the $3''$ flux. Darker components correspond to galaxies that are of later type (bluer) when using the total flux. The percentages of galaxies residing within these contours are listed in Table 2.3. From our repeatability test, the mean signal-to-noise limited classification is $\langle|\delta(\phi_{KL})|\rangle = 2.35$. With the assumption that the typical signal-to-noise limit in the ϕ_{KL} angle estimation for the whole galaxy is the same as that for the inner $3''$, the derived signal-to-noise limit in $D\phi_{KL}$ is $2 \times \langle|\delta(\phi_{KL})|\rangle \sim 5$. There are about half of the galaxies ($\sim 40\%$) in our sample in which the type-differences are within the estimated signal-to-noise limit.

Aperture effects on the spectral classification clearly exist. For blue galaxies (i.e., $\phi_{KL} \sim 0^o - -40^o$), $D\phi_{KL}$ increases for nearby galaxies. This is to be expected as the flux from the inner $3''$ is more likely to be dominated by the presence of a bulge component. Similarly, for galaxies classified as red based on their total flux (i.e., $\phi_{KL} \sim 0^o - 20^o$), errors in the classification angles $D\phi_{KL}$ increase rapidly with decreasing distance (i.e., $z/z_{max} < 0.25$). This implies that the cores of red galaxies are redder than the color estimated from the total flux.

The dependence of the aperture bias on the physical size of a galaxy is illustrated in Fig. 2.21. We divide the above sample into 6 ranges of galaxy type, from the reddest in

Fig. (a) to the bluest in Fig. (f). The differences in the classification angles $D\phi_{KL}$ are plotted in each figure as a function of z/z_{max} for physical sizes ranging from $0 - 100$ kpc (black line), $10 - 15$ kpc (dotted line) and $30 - 35$ kpc (dashed line). The two horizontal lines mark the uncertainty on the classification due to the survey signal-to-noise limits. For the red galaxies in Fig. (a) to (c) the bias is constant or decreases with effective distance and is, essentially, negligible when compared to the uncertainties on the classification. For distances $z/z_{max} < 0.25$, the bias is above the signal-to-noise limit so that the type deduced from the total flux is redder than that from the central $3''$.

As we would expect Fig. (b) and (c) show a dependence on galaxy size for the classifications with larger galaxies exhibiting a redder classification when only considering the $3''$ flux. This size and redshift dependence extends to the blue galaxies (Fig. (d) through (f)). For these galaxies, however, a more pronounced dependency is shown on the classification bias with galaxy size. Overall there is a general aperture bias for all physical sizes of galaxy ($0 - 100$ kpc) that approaches the intrinsic error on the classification as the effective redshift z/z_{max} approaches unity. The exception to this arises when we consider galaxies with prominent emission lines (Fig. (f)) which, counter-intuitively, exhibit *larger* bias the more distant they are. One of the possible reasons for this is our use of three eigenspectra in constructing the spectral templates, whereas 10 modes and more are typically required to accurately reconstruct these observed spectra.

While we observe an aperture bias in the SDSS sample it is relatively small when compared to the intrinsic classification errors and is essentially negligible for most galaxies. Moreover its dependence on size and redshift is relatively mild and straightforward to correct. For those galaxies with non-negligible bias a simple correction can be made using the lookup table shown in Fig. 2.23, which is identical to Fig. 2.22 with the exception that the ordinate axes is the spectral type inside $3''$, $\phi_{KL}(3'')$. Given the $\phi_{KL}(3'')$ and z/z_{max} the correction to our spectral classification can be determined directly from Fig. 2.23 with the size-averaged bias.
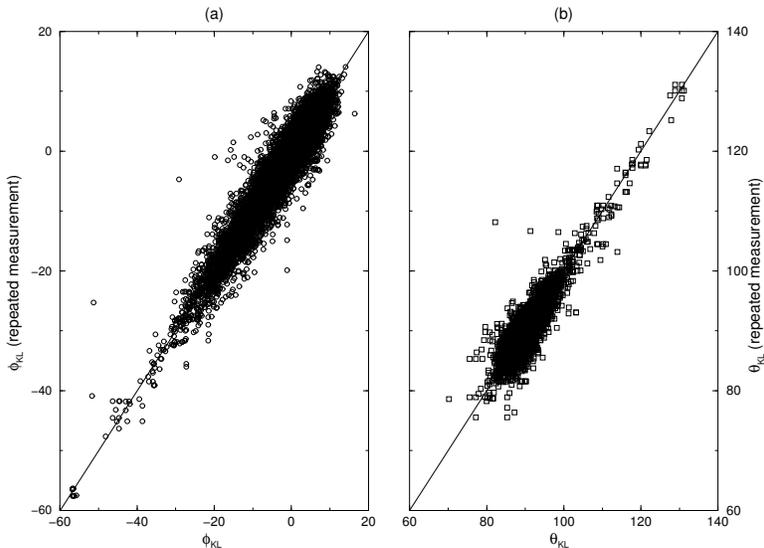
Figure 2.24 Repeatability of the KL classification of galaxies. The classification parameter of each object is plotted against that of the repeated measurement, for the cases (a) ϕ_{KL} and (a) θ_{KL}.

2.6   REPEATABILITY OF CLASSIFICATION

Any classification scheme has to be repeatable in order to be useful. If we measure the spectrum of a galaxy on different nights in different conditions, would the classification still be the same? To answer this question, we are fortunate in that many galaxy spectra in the SDSS data set are taken on multiple nights (Blanton et al. 2003). A total of 1,854 galaxies were found in our sample to be not unique (i.e., they have been observed and reduced independently). A further thirty thousand galaxies were found in the SDSS spectroscopic data to have been observed on multiple nights with different observing conditions (often these individual observations do not meet the signal-to-noise requirements of the SDSS spectral observations). The quantitative interpretation of the repeatability of classifications based on these plates may be difficult due to the variation in signal-to-noise ratios. Nevertheless all repeat observations are selected for this part of our work.

Of the thirty-thousand sources, only those with flags PRIMTARGET, OBJTYPE and CLASS equal GALAXY are selected (together with the requirement that all sources are
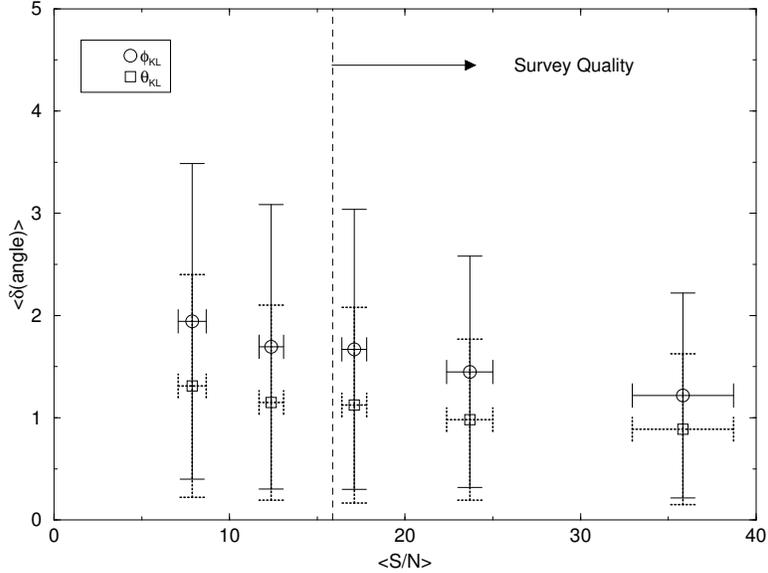
Figure 2.25 The mean discrepancy in the classification angles derived from the KL analysis (circles are $\langle|\delta\phi_{KL}|\rangle$ and squares, $\langle|\delta\theta_{KL}|\rangle$) versus the mean signal-to-noise ratio of all spectra. All spectra in this plot have been observed more than once.

present in the most up to date reductions). This selection results in thirteen-thousand objects in the final sample. Fig. 2.24a and Fig. 2.24b show a comparison between the ϕ_{KL} and θ_{KL} values assigned by our classification scheme to those galaxies with the highest signal-to-noise and the classification derived from the repeat observations. The solid line corresponds to the location of the one-to-one correspondence between the two measures. The dispersions in ϕ_{KL} and θ_{KL} are 2.35^o and 1.61^o respectively. This agreement is excellent, as these angle dispersions correspond to small changes in the resulting repaired spectra. The agreement also spans a large range in both classification angles. The implication of this finding is that a truly reliable and repeatable classification scheme is obtained which validates the repeatability of the spectrophotometry of the SDSS.

In order to determine a representative signal-to-noise ratio for each spectrum the median signal-to-noise ratio is adopted (the flag SN_MEDIAN in spZbest-*plate-mjd*.fits). The dependence of the rms error in the measured angles on the signal-to-noise of the observations (where the signal-to-noise is selected to be the lower one of any pair of observations) is shown in Fig. 2.25. From the current data set we observe a weak trend with larger dis-

crepancies in the classification for those observations with lower signal-to-noise ratios. The mean (absolute) discrepancies in the classification angles ($\langle|\delta(\phi_{KL})|\rangle$ and $\langle|\delta(\theta_{KL})|\rangle$) and mean signal-to-noise ratios are calculated in the ranges of signal-to-noise ratio $(0.0 - 10.0)$, $(10.0 - 15.0)$, $(15.0 - 20.0)$, $(20.0 - 30.0)$ and larger than 30.0. The dependence is very similar in both the ϕ_{KL} and θ_{KL} angles. The error bars are set by the root-mean-square fluctuations in both quantities. The vertical line marks the calculated mean signal-to-noise ratio of all the galaxies defined as meeting the survey quality (a signal-to-noise of 15.9).

For those sources meeting the survey quality signal-to-noise criteria, the maximum errors in the two mixing angles are, approximately, three degrees in ϕ_{KL} and two degrees in θ_{KL}. This result shows that the classifications based on the SDSS spectra are repeatable and robust to the variable signal-to-noise within the spectroscopic data. The fact that the signal-to-noise dependence is weak suggests that the noise within the spectra are essentially Poisson such that the projection of a noisy spectrum does not add substantial artifacts into the expansion coefficients. Despite this weak dependence in the distribution of expansion coefficients with signal-to-noise we do find instances where the spectral properties of the galaxies change between pairs of observations. For example, in one case we find that the strength of the O II lines change by about 20% between two separate observations. It is not clear whether this difference is due to a calibration error or due to variability in the source.

3.0 DECOMPOSITION OF QUASAR SPECTRA

It takes all sorts to make a world.

English Proverb

Quasar spectra pose different challenges in regard to their classifications compared with galaxy spectra mainly because of the diversity in the line kinematics such as the emission line profiles and line-widths. In this chapter we apply the Karhunen-Loève transform to decompose a sample of QSO spectra from the SDSS.

3.1 DATA

The sample we use is an early version of the First Data Release (DR1, Abazajian et al. 2003) quasar catalog (Schneider et al. 2003) from the SDSS, which contains 16,707 quasar spectra and was created on the 9th of July, 2003. The official DR1 quasar catalog includes slightly more objects (16,713) and was created on the 28th of August, 2003. All spectra in our sample are cataloged in the official DR1 quasar catalog except one: SDSS J150322.94+600311.3 (i.e., there are 7 DR1 QSOs not included in our sample). A description of the SDSS pipeline is given in §2.1. The details of the target selection, the spectroscopic reduction and the catalog format are discussed by Schneider et al. (2003) and references therein. About 64% of the quasar candidates in our sample are chosen based on their locations in the multi-dimensional SDSS color-space (Richards et al. 2002a), while $\sim 22\%$ are targeted solely by the Serendipity module. The remaining QSOs are primarily targeted as FIRST sources, ROSAT sources, stars or galaxies. All quasars in the DR1 catalog have absolute magnitudes (M_i) brighter

than -22.0, where M_i are calculated using cosmological parameters $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$; and that the UV-optical spectra can be approximated by a power law ($f_\nu \propto \nu^{\alpha_\nu}$) with the frequency index $\alpha_\nu = -0.5$ (Vanden Berk et al. 2001). The absolute magnitudes in five bands are corrected for Galactic extinction using the SFD dust maps (Schlegel et al. 1998). Quasar targets are assigned to the $3''$ diameter fibers for spectroscopic observations (the tiling process; Blanton et al. 2003). Spectroscopic observations are discussed in detail by York et al. (2000); Castander et al. (2001); Stoughton et al. (2002); Schneider et al. (2002). The SDSS Spectroscopic Pipeline, among other procedures, removes skylines and atmospheric absorption bands, and calibrates the wavelengths and the fluxes. The signal-to-noise ratios generally meet the requirement of $(S/N)^2$ of 15 per spectroscopic pixel (Stoughton et al. 2002). The resultant spectra cover $3800 - 9200$ Å in the observed frame with a spectral resolution of $1800 - 2100$. At least one prominent line in each spectrum in the DR1 quasar catalog is of FWHM ≥ 1000 km s^{-1}. Type 2 quasars and BL Lacs are not included in the DR1 quasar catalog.

All of the 16,707 quasars are included in our analysis, including quasars with broad absorption lines (BALQSOs). To perform the KL transforms, the spectra are shifted to their restframes, and linearly rebinned to a spectral resolution $\sim 1800/(1 + z_{min})$, with z_{min} being the lowest redshift of the whole sample (§3.2) or of the subsamples of different (M_i, z)-bins (defined in §3.3). Skylines and bad pixels due to artifacts are removed and fixed with the gap-correction procedure discussed in §2.2 and §3.7.

3.2 GLOBAL QSO EIGENSPECTRA

Models of accretion on black holes and scenarios for the formations of Fe II-blends often predict relationships between the UV and optical quasar spectral properties (for example, the strong anti-correlation between the "small bump" and the optical Fe II blends was suggested by Netzer and Wills 1983). Using our sample with 16,707 quasar spectra, we construct a set of eigenspectra covering 900 Å to 8000 Å in the restframe. For each quasar spectrum, the spectral regions without the SDSS spectroscopic data are approximated by the linear

Table 3.1 The partial sums of weights of the global QSO eigenspectra.

Number of first m-modes: m	Weight
1	0.560887
2	0.680197
3	0.755992
4	0.822280
5	0.849927
8	0.896213
10	0.919394
15	0.953041
20	0.968920
50	0.995680
75	0.998512
100	0.999199

combinations of the calculated eigenspectra by the gap-correction procedure described in §2.2 and §3.7. A quantitative assessment of this procedure on quasar spectra is discussed in detail in appendix 2.2. To determine the number of iterations needed for this gap-correcting procedure, we calculate the commonality between the two subspaces spanned by the eigenspectra in one iteration step and those in the next step. For the subspace spanned by the first two modes, the convergence rate is fast and it requires about three iterations at most to converge. Including higher-order components, in this case the first 100 modes, the subspace takes about 10 iteration steps to converge. In this work, all eigenspectra are corrected for the missing pixels with 10 iteration steps. The gaps in each spectrum are corrected for using the first 100 eigenspectra during the iteration.

The partial sums of weights (i.e., accumulative weights, where the weights are the eigenvalues of the correlation matrix) in different orders of the global eigenspectra are shown in Table 3.1. The first eigenspectrum accounts for about 0.56 of the total sample variance and the first 10 modes account for ~ 0.92. To account for ~ 0.99 of the total sample variance, about $50 - 60$ modes are required. The first four eigenspectra are shown in Fig. 3.1, and their physical attributes will be discussed below.
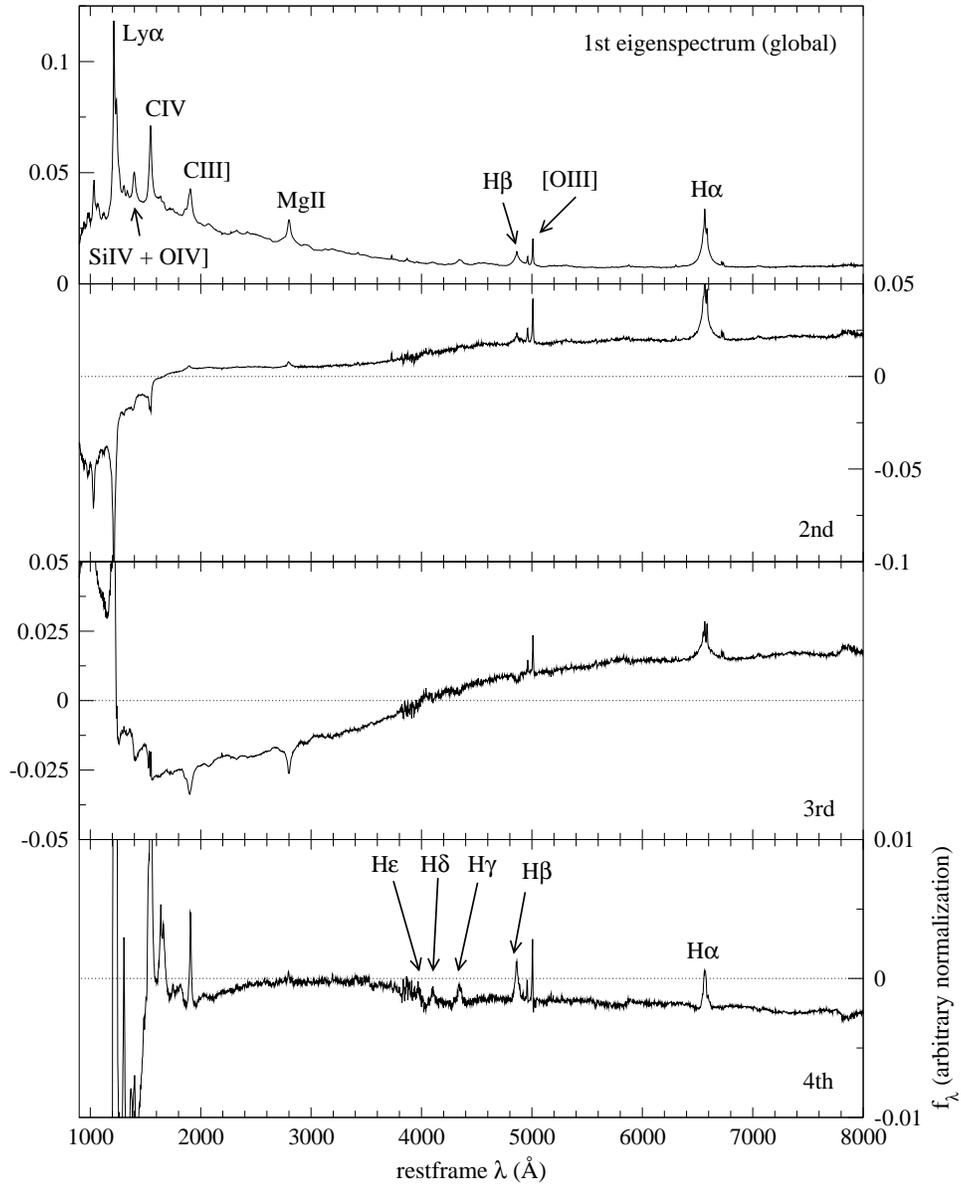
Figure 3.1 The first 4 eigenspectra of 16,707 SDSS quasars, in the rest wavelengths 900 − 8000 Å.
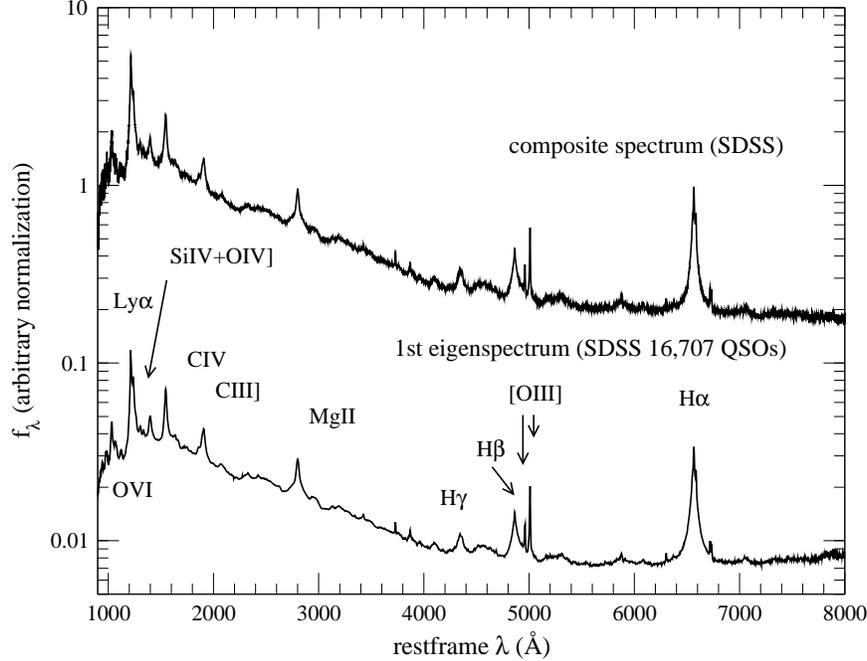
Figure 3.2 There is a good agreement between the first global QSO eigenspectrum and the SDSS QSO composite.

3.2.1 1st mode: composite spectrum

The first eigenspectrum (the average spectrum of the data set) reveals the dominant broad emission lines that exist in the range of $\lambda_{rest} = 900 - 8000$ Å. These, presumably Doppler-broadened lines, are common to most quasar spectra. As can be seen in Fig. 3.2, this eigenspectrum exhibits a high degree of similarity with the median composite spectrum (Vanden Berk et al. 2001) constructed using over 2200 SDSS quasars, but with lesser noise at the blue and red ends, probably due to the larger sample used in this analysis.

3.2.2 2nd mode: host-galaxy component

The 2nd eigenspectrum shows a striking similarity in the optical region ($\lambda_{rest} \geq 3500$ Å) with the 1st galaxy eigenspectrum (i.e., mean spectrum) from the SDSS galaxies (of $\sim 170,000$ galaxy spectra; Yip et al. 2004a). In Fig. 3.3 we show a comparison between the two. Besides the presence of the Ca K and Ca H lines and the Balmer absorption lines as

reported previously in the composite quasar spectrum, the Mg I triplet (which appears to be composed of two lines because of the limited resolution, i.e., Mg Iλ5169+λ5174, and Mg Iλ5185[1]) is also seen in this mode. The presence of the Balmer absorption lines (see the inset of Fig. 3.3) implies the presence of young to intermediate stellar populations near the nuclei (because of the SDSS 3″ spectroscopic fiber). The main differences between the quasar 2nd eigenspectrum and the galaxy mean spectrum lie in the Balmer lines Hα and Hβ, which are, as expected, Doppler-broadened for the QSO spectra. The quasar eigenspectrum also has a redder continuum, meaning that *if* this eigen-component represents all contributions from the host-galaxies, the galaxies would be of earlier spectral type than the average spectral type in the SDSS Main galaxy sample.

Our ability to detect significant host-galaxy features in this eigenspectrum triggers an important application, that is, the removal of the host-galaxy contributions from the quasar spectra. The properties the host-galaxies of quasars have recently attracted interest (e.g., Bahcall et al. 1997; McLure et al. 1999, 2000; Nolan et al. 2001; Hamann et al. 2003), mainly because of their obvious relationship with the quasars they harbor and the probable co-evolution that happens between them. Therefore, the evolution of massive galaxies, which are believed to be at one time active quasar hosts (see Hamann and Ferland 1999), can also be probed.

On the other hand, narrow emission lines in active galactic nuclei (AGNs) have been considered less useful than broad emission lines as diagnostic tools, because AGNs with prominent narrow lines have low luminosities (see, for example, the discussion in (chapter 10 of Krolik 1999), in which case contributions from the host galaxies may affect both the continuum and the lines, obscuring their true appearances. Hence, the removal of host-galaxy components can potentially fix the narrow emission lines and reveal their true physical nature. A recent study (Vanden Berk et al. 2005) shows that the method is applicable in removing the galaxy continuum in the lower-redshift quasars in the SDSS sample.

The second mode also shows slight anti-correlations between major broad emission lines which exist in λ_{rest} smaller and larger than $\sim$ 2000 Å (see Fig. 3.1).

---

[1]The restframe wavelength of the longest wavelength component of the Mg I triplet appears to be red-shifted by $\sim$ 290 km s^{-1} relative to the laboratory vacuum value of 5185 Å. This maybe due to the contamination by an unidentified absorption line redward of Mg Iλ5185.
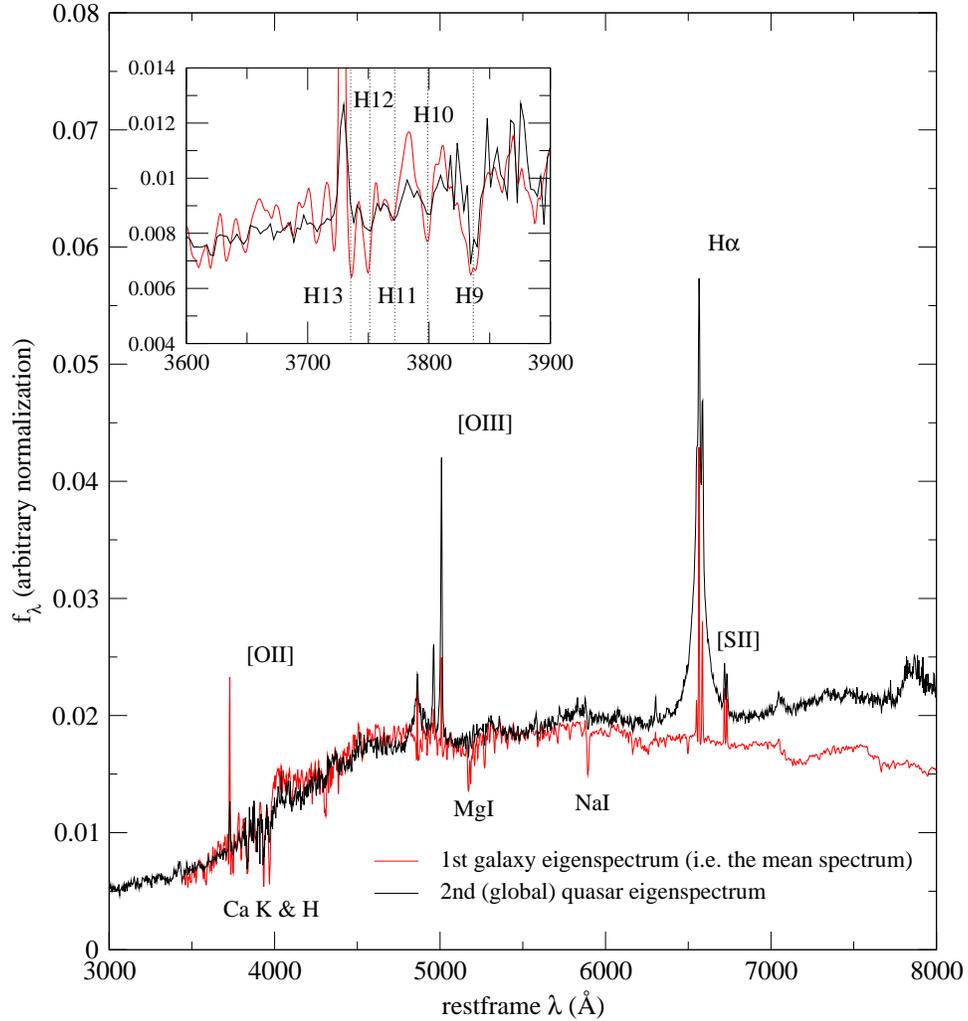
Figure 3.3 Comparison of the global quasar eigenspectrum of this work with the 1st eigen-spectrum (i.e., the mean spectrum) of the SDSS galaxies ($\sim 170,000$ galaxy spectra) in the rest-wavelengths $3000-8000$ Å. Not only the major emission lines and absorption lines noted in the graph are found in both cases, the "bumps and wiggles" in the continua also exhibit similarities. The inset shows the spectral region near the hydrogen absorption lines.

3.2.3 3rd mode: UV-optical continuum slope

The change of the continuum slope, with a zero-crossing (i.e., a node) at around 3990 Å, dominates this global eigenspectrum. The optical continuum appears to be galaxy-like, but not as much as the 2nd global eigenspectrum. For example, in this component the [O II]λ3728 is missing, and the nebular lines are generally weaker. The node at ~ 4000 Å is in partial agreement with the 2nd principal component of 18 low-redshift ($z < 0.4$; BALQSOs *excluded*) quasar spectra (Shang et al. 2003), which showed the UV-optical continuum variation (except the node is at ~ 2600 Å). This particular wavelength (4000 Å) marks the modulation of the slope between the UV and the optical regions. One related effect is the "ultra-violet excess", describing the abrupt rise of quasar flux densities from about 4000 Å to 3500 Å. This observed excess flux was suggested to be due to the Balmer continuum (Malkan and Sargent 1982), as there seem to be no other mechanisms which can explain this wavelength coincidence. In Malkan & Sargent's work, an exact wavelength for this onset was not clear. The node at ~ 4000 Å can serve the purpose of defining that wavelength. Other possible physical reasons for the modulations between the UV and optical continua are the intrinsic change in the quasar continuum (e.g., due to intrinsic dust-reddening) and the stellar light from the host galaxy. There is also a second node located in Lyα showing an anti-correlation between the continua blueward and redward of the Lyα. Since the number of quasars with spectroscopic measurements in the vicinity of Lyα is much smaller than those with measurements in the UV-optical regions that are redward of Lyα, the significance of this anti-correlation is less than that of the UV-optical continuum variation in this eigenspectrum.

3.2.4 4th mode: correlations of balmer emission lines

This mode shows the correlations of broad emission lines, namely, Lyα, C IV, Si IV+O IV], C III], Mg II, [O III]λ5008 and also the Balmer emission lines Hα, Hβ, Hγ, Hδ and Hϵ. These are in partial agreement with the 3rd eigenspectrum of Shang et al., in which emission lines C III], Mg II, Hα, Hβ are found to be involved. It seems natural that these Balmer lines are correlated, as presumably they are formed coherently by some photo-ionization processes. However, it is not known why they appear in this low-order mode. The fact that C III] and

Hβ vary similarly was seen previously (Wills et al. 2000), and it was suggested that Hβ and C III] may arise from the same optically-thick disk.

3.2.5   Higher orders

By construction, subsequent higher-order eigenspectra show more nodes, causing small modulations of the continuum slope. They also show broad absorption line features. Since quasars with BALs are not the dominating populations in our sample (there are 224 broad absorption line quasars in the 3814 quasars from the SDSS EDR quasar catalog, Reichard et al. 2003), their signatures preferentially show up at higher orders in this global set of eigenspectra. The BAL components are not confined to only one particular mode, but span a number of orders. To investigate the effects of BALQSOs on the global eigenspectra, our approach is to perform the KL transform on our original sample (including BALQSOs) and on the same sample but with the BALQSOs excluded, and make a comparison between them. There are 682 BALQSOs (with balnicity index > 0) found in our sample according to the BALQSO catalog for the SDSS spectra by Trump et al. (2003).

Fig. 3.4 compares the weights at different orders between the BALQSO-included and the BALQSO-excluded global eigenspectra. Since the BALQSO-included global eigenspectra contain information describing both the non-BALQSOs and the BALQSOs, the weight of each mode is larger than that of the BALQSO-excluded eigenspectra. That is, the BALQSO-excluded eigenspectra set is more compact. The magnitude of this offset, however, is small and is apparent only after the 5-th order, which is consistent with the fact that the BALQSOs form a minority population (about 4%). This difference is seen to extend to higher orders, implying that the features describing the BALQSOs span a number of higher-order eigenspectra and are not confined to only one particular mode.

A comparison of the 6th global eigenspectrum between the BALQSO included and excluded samples is shown in Fig. 3.5. Absorption features (in this case, in Si IV+O IV] and C IV) are found in the first set of eigenspectra but are missing in the latter. We have to note that the discrepancies in the spectral features of these two sets of eigenspectra attributed to the weight differences are not only confined to the existence or non-existence of BAL
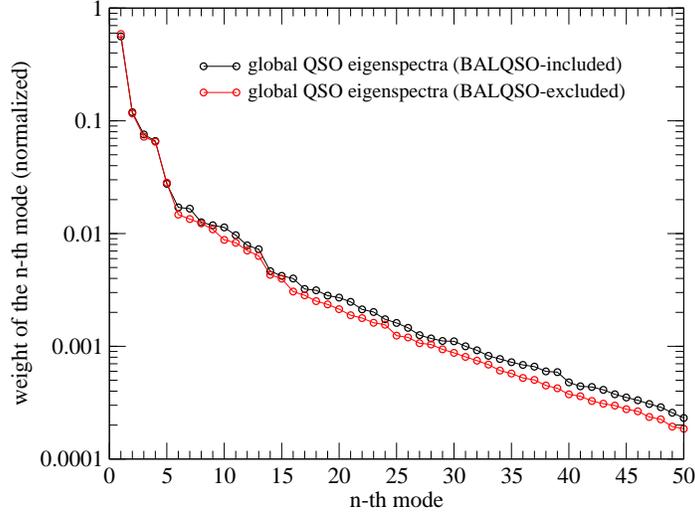
Figure 3.4 Comparison of the weights at different orders between the BALQSO-included global eigenspectra (i.e., the 16,707 QSOs; black) and the BALQSO-excluded global eigenspectra (red). Since the BALQSO-included global eigenspectra contain information for features of both typical quasars and broad absorption lines, the weight of each mode is larger than that of the BALQSO-excluded eigenspectra.

absorption troughs as shown here, as the difference in the normalizations between the two can in general also yield different eigenspectra sets. We will leave the discussion of the reconstruction of the BALQSO spectra using eigenspectra till §3.3.4 .

3.2.6 A non-unique set of eigenspectra by commonality analysis

To study the possible evolution and luminosity effects in the quasar spectra, our first step is to investigate whether the set of eigenspectra of a given order derived from quasar spectra in different redshift and luminosity ranges differ. The trace quantity mentioned in §2.2 is adopted for these quantitative comparisons.

As a null measure, two subsamples are chosen with approximately the same redshift and luminosity distributions, such that any differences in the two sets of eigenspectra would be due to noise and the intrinsic variability of the quasars. We fix the rest-wavelengths of this study to be $2000 - 4000$ Å, and require a full rest-wavelength coverage of the input quasars;
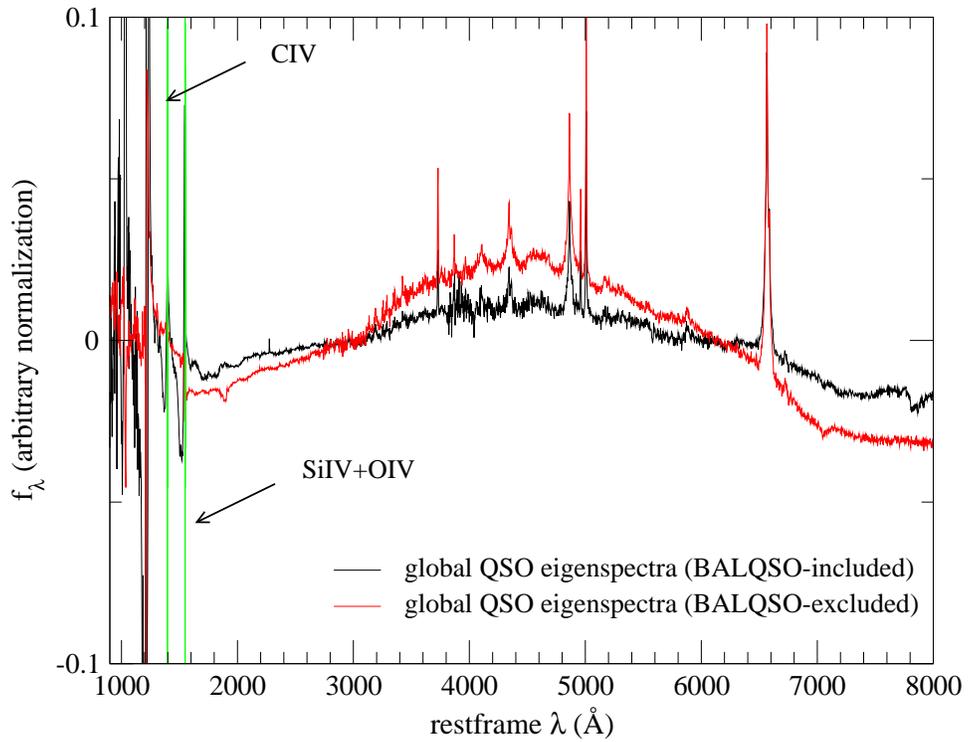
Figure 3.5 Comparison between the 6th global eigenspectra from the BALQSO-included sample and the BALQSO-excluded one. Absorption features in Si IV+O IV] and C IV exist in the first case and is missing in the latter, as expected.

redshifts are limited to 0.9 to 1.1. One subsample contains 472 objects (Subsample 1) and the other subsample, 236 objects (Subsample 2). Subsample 2 is, by construction, a subset of the original 472 objects. The reason behind this construction is to ensure a high commonality of the two sets of resultant eigenspectra. They both have luminosities from -24 to -25, and the actual distributions of redshifts and luminosities are similar. The line on the top in Fig. 3.6 shows the commonality of these two subsamples as we increase the number of eigenspectra forming the subspace. As higher orders of eigenspectra are included in the subspaces, the commonality drops, meaning that the two subspaces become more disjoint. As mentioned above, this disjoint behavior is mainly due to the noise and the intrinsic variability among quasars, both are unlikely to be completely eliminated. At about 20 modes and higher, the commonality levels off, which implies that the eigenspectra mainly contain noise.

With this null measure in place, the differences of our test subsamples are further relaxed to include luminosity effects alone (Subsamples 1 and 3, see Table 3.2), redshift effects alone (Subsamples 3 and 4), and lastly, both effects combined (Subsamples 1 and 4). The commonalities of these subsamples are overlaid in Fig. 3.6. The first modes constructed in all these subsamples, including the null measure, are always very similar to each other (more than 99% similar). This shows that a single mean spectrum can be constructed across the whole redshift coverage, which was presumed to be true in many previous constructions of quasar composite spectra. The validity of construction of the mean spectrum in a given sample may seem trivial, but it is not if we take into account the possibility that the quasar population may evolve at different cosmic epochs.

Similar to the null measure, as higher orders are included in the subspaces, the eigenspectra subspaces become more disjoint. In addition, the commonalities in these condition-relaxed cases actually drop *below* the null measure for orders of modes higher than ~ 10. Therefore, the eigenspectra of the same order but derived from quasars of different redshifts and luminosities describe different spectral features. In addition, our results show that both luminosity and evolution effects have detectable influences on the resultant sets of eigenspectra, very much to the same degree (in terms of commonality). In the case of the combined effects, the commonality drops to the lowest value among all cases, as expected.

The actual redshift and luminosity effects found in the quasar spectra will be presented
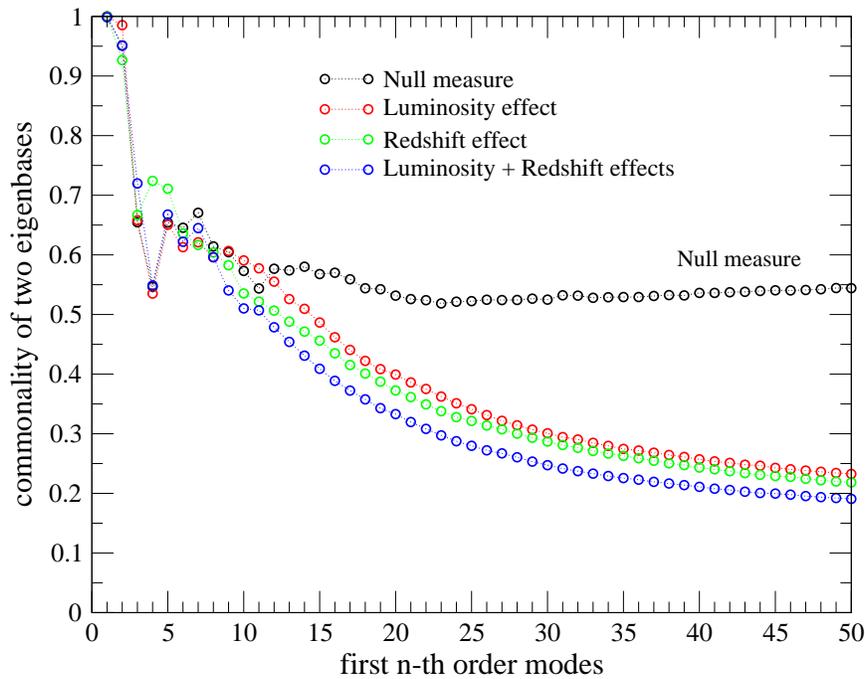
Figure 3.6 Commonality of two subsamples which are designed to be different from each other mainly due to noise and intrinsic variations (black), the additional luminosity effect (red), redshift effect (green), and both (blue); plotted against the common dimension of the eigenbases (which equals the number of modes constructing each eigenbasis). The commonality departs from unity and progressively drops below the null measure. This means that the two eigenbases under consideration are more disjoint from each other when higher the orders are included.

Table 3.2 The subsamples for performing the commonality analysis on the resultant sets of eigenspectra.

Name	Redshift range	Luminosity range (M_i)	Number of objects
Subsample 1	0.9 to 1.1	-24 to -25	472
Subsample 2	0.9 to 1.1	-24 to -25	236
Subsample 3	0.9 to 1.1	-25 to -26	442
Subsample 4	1.1 to 1.3	-25 to -26	469

in §3.4.2 and 3.4.3. We learn from this analysis that there does not exist a unique set of KL eigenspectra across the whole redshift range, with the number of modes equal or smaller than approximately 10. The implications are twofold. On one hand, the classification of quasar spectra, in the context of the eigenspectra approach, has to be redshift and luminosity dependent. In other words, the *weights* of different modes are in general different when quasars of different redshifts and luminosities are projected onto the same set of eigenspectra. So, eigenspectra derived from quasars of a particular redshift and luminosity range in general do not *predict* quasar spectra of other redshifts and luminosities. On the other hand, the existence of the redshift and luminosity effects in our sample can be probed quantitatively by analyzing the eigenspectra subspaces.

3.3 (M_i, z)-BINNED QSO EIGENSPECTRA

KL transforms are performed on subsamples with different redshift and luminosity ranges, that allow us to explicitly discriminate the possible luminosity effects on the spectra from any evolution effects, and vice versa[2]. The constructions of these bins are based on requiring that the maximum gap fraction among the quasars, that is, the wavelength region without the SDSS data, is smaller than 50% of the the total spectral region we use when applying the KL transforms. The total spectral region, by construction, is approximately equal to the largest common rest-wavelengths of all the quasars in that particular bin. We find that constraining the gap fraction to be a maximum of 50% improves the accuracy of the gap-correcting procedure for most quasars (see appendix 2.2 for further explanation). As a result, five divisions are made in the whole redshift range $0.08 < z < 5.13$ (where the quasars of redshifts larger than 5.13 are discarded to satisfy the constraint of 50% minimum wavelength-coverage in all related luminosity bins), and four in the whole luminosity range $M_i = (-30, -22)$. These correspond to **ZBIN 1** to **5** and the M_i bins **A** to **D** for the redshift and luminosity subsamples respectively. In the following, we denote each subsample

---

[2]Since the K-correction of our sample is calculated in the SDSS assuming a spectral index $\alpha_\nu = -0.5$, so in principle a color dependence is present in any redshift trend found.

Table 3.3 The number of QSOs in the (M_i, z)-bins.

	A $M_i = (-30, -28)$	B $M_i = (-28, -26)$	C $M_i = (-26, -24)$	D $M_i = (-24, -22)$
ZBIN 1: $0.08 < z < 0.53$ $2486 - 8000$ Å			109 $(95\%^a, 0\%^b)$	1597 $(81\%, 4\%)$
2: $0.53 < z < 1.16$ $1759 - 6018$ Å		178 $(94\%, 0\%)$	2752 $(79\%, 12\%)$	1351 $(30\%, 45\%)$
3: $1.16 < z < 2.06$ $1242 - 3800$ Å		3477 $(92\%, 4\%)$	4462 $(41\%, 46\%)$	
4: $2.06 < z < 3.33$ $900 - 3005$ Å	110 $(74\%, 0\%)$	1796 $(65\%, 11\%)$	477 $(0\%, 61\%)$	
5: $3.33 < z < 5.13$ $900 - 2123$ Å	85 $(94\%, 0\%)$	352 $(75\%, 1\%)$		

---

[a]The percentage of quasars (to the nearest unity) that are targeted by the quasar multi-dimensional color-space in the SDSS.

[b]The percentage of quasars targeted solely by the SDSS Serendipity module.

in a given luminosity and redshift range, for example, the bin **A4**. Such divisions are by no means unique and can be constructed according to one's own purposes, but we find that important issues such as the correlation between continua and emission lines remain unchanged as we construct bins with slightly different coverages in redshift, in luminosity and in the total rest-wavelength range. The actual rest-wavelength range and the number of spectra in each bin are shown in Table 3.3, which also lists the fractions of QSOs in each bin that are targeted either in the quasar color-space (Richards et al. 2002a) or solely by the Serendipity module. While the majority of the quasars from most of the bins are targeted by using the multi-dimensional color-space, in which the derived eigenspectra are expected to be dominated the intrinsic quasar properties, there is one bin (**C4**) in which most quasars are targeted by the Serendipity module. In principle, the eigenspectra in the latter case will represent the properties of the serendipitous objects and lack a well motivated color distribution.

In general for all (M_i, z)-bins, the first 10 modes or less are required to account for more than 92% of the variances of the corresponding spectra sets (Table 3.4). In the iterated calculation of the (M_i, z)-binned eigenspectra, the first 50 modes are used in the gap correction. The first 4 orders of eigenspectra of each (M_i, z)-bin are shown in Fig. $3.7 - 3.11$, arranged in

Table 3.4 The partial sums of weights of the (M_i, z)-binned QSO eigenspectra.

weight	1 mode	2 modes	3 modes	5 modes	10 modes	15 modes	20 modes	50 modes
ZBIN: 1	0.9284	0.9646	0.9729	0.9836	0.9915	0.9934	0.9942	0.9945
2	0.9317	0.9657	0.9752	0.9815	0.9874	0.9896	0.9905	0.9909
3	0.9232	0.9556	0.9651	0.9763	0.9841	0.9870	0.9880	0.9885
4	0.8737	0.9089	0.9298	0.9455	0.9608	0.9685	0.9719	0.9738
5	0.8122	0.8540	0.8783	0.8986	0.9247	0.9356	0.9398	0.9422
(M_i, z)-bin: **A4**	0.9134	0.9449	0.9551	0.9663	0.9781	0.9842	0.9866	0.9881
A5	0.8801	0.9076	0.9242	0.9384	0.9555	0.9637	0.9676	0.9699
B2	0.9789	0.9926	0.9958	0.9973	0.9984	0.9989	0.9991	0.9991
B3	0.9474	0.9701	0.9781	0.9859	0.9911	0.9931	0.9938	0.9941
B4	0.8794	0.9115	0.9292	0.9516	0.9673	0.9743	0.9771	0.9787
B5	0.8083	0.8487	0.8718	0.8941	0.9207	0.9314	0.9358	0.9383
C1	0.9893	0.9955	0.9969	0.9981	0.9991	0.9994	0.9995	0.9995
C2	0.9470	0.9709	0.9819	0.9870	0.9921	0.9938	0.9944	0.9947
C3	0.9226	0.9546	0.9631	0.9739	0.9805	0.9832	0.9842	0.9848
C4	0.8365	0.8756	0.8983	0.9191	0.9387	0.9494	0.9539	0.9565
D1	0.9289	0.9640	0.9715	0.9829	0.9908	0.9928	0.9937	0.9941
D2	0.9086	0.9469	0.9586	0.9698	0.9768	0.9797	0.9809	0.9816

5 different redshift ranges. In each figure, eigenspectra of different luminosities are plotted along with the ones which are constructed by combining all luminosities (shown in black curves). By visual inspection, the eigenspectra in different orders show diverse properties for each (M_i, z)-bin. In the following, properties associated with different orders are extracted by considering *all* (M_i, z)-bins generally. Eigenspectra which are distinct from the average population will be discussed separately.

3.3.1 1st mode: composite spectra

As in the global case, the lowest-order eigenspectra are simply the mean of the quasars in the given subsamples. For every redshift bin, the first eigenspectrum shows approximately a power-law shape (either a single or broken power law), with prominent broad emission lines. Different luminosity bins show differences in the overall spectral slopes to various degrees. In every redshift range, the spectra of higher-luminosity quasars are bluer than their lower luminosity counterparts. For example, **C1** (Fig. 3.7; $M_i = -26 - -24$) shows a harder spectral slope blueward of ~ 4000 Å than that of **D1** ($M_i = -24 - -22$). However, for the higher redshift ($z = 2.06 - 5.13$) quasars, e.g., in **ZBIN 4** (Fig. 3.10) and **5** (Fig. 3.11), the difference in spectral slope seems to be confined mainly to changes in the flux densities blueward of Lyα.
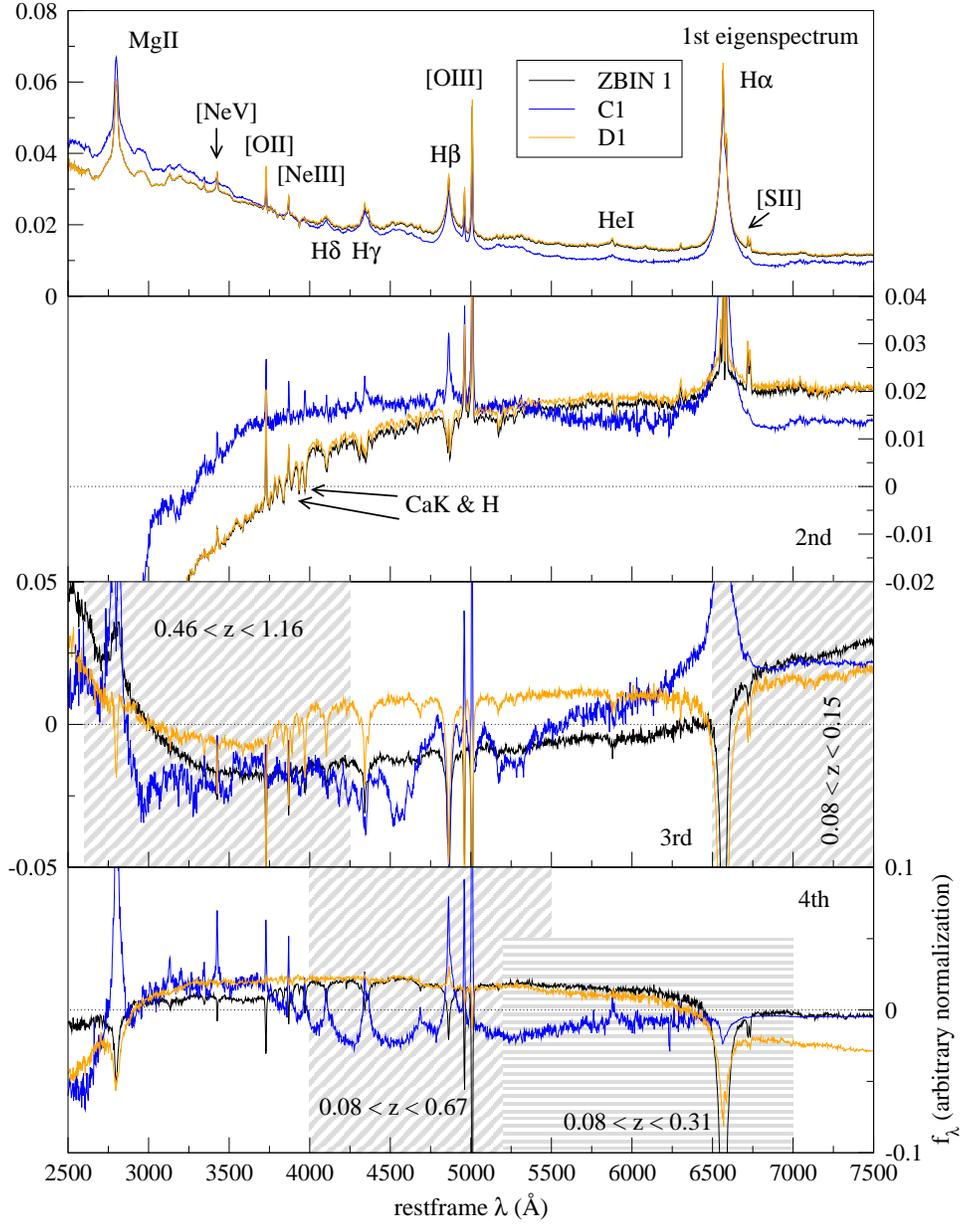
Figure 3.7 The first 4 orders of (M_i, z)-binned eigenspectra for the subsample **ZBIN 1**. The shaded areas correspond to the local spectral properties as found in the 2nd local eigenspectra sets using wavelength-selected quasar spectra.
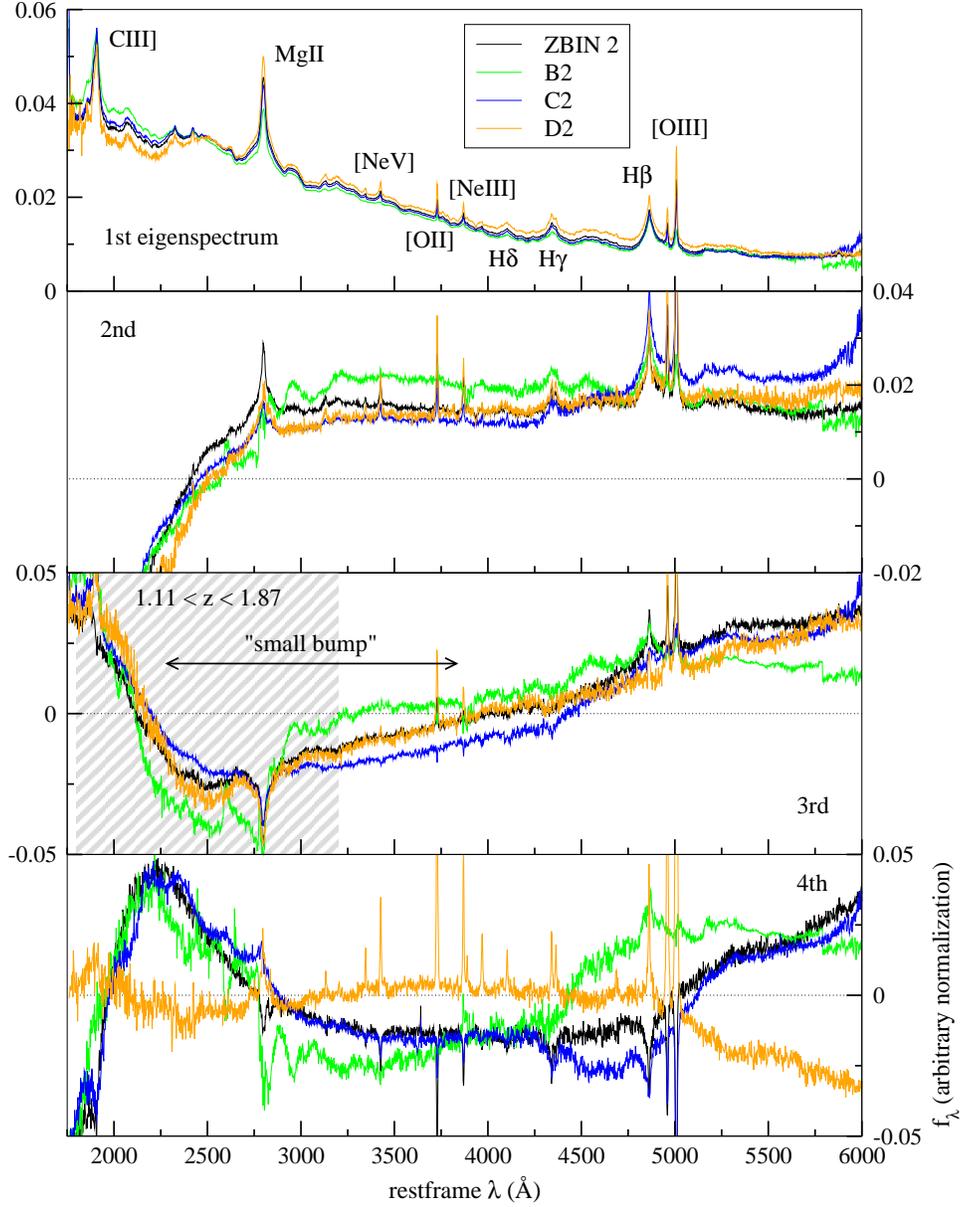
Figure 3.8 The first 4 orders of (M_i, z)-binned eigenspectra for the subsample **ZBIN 2**. The meaning of the shaded spectral region is explained in the caption of Fig. 3.7 and §3.6.4.
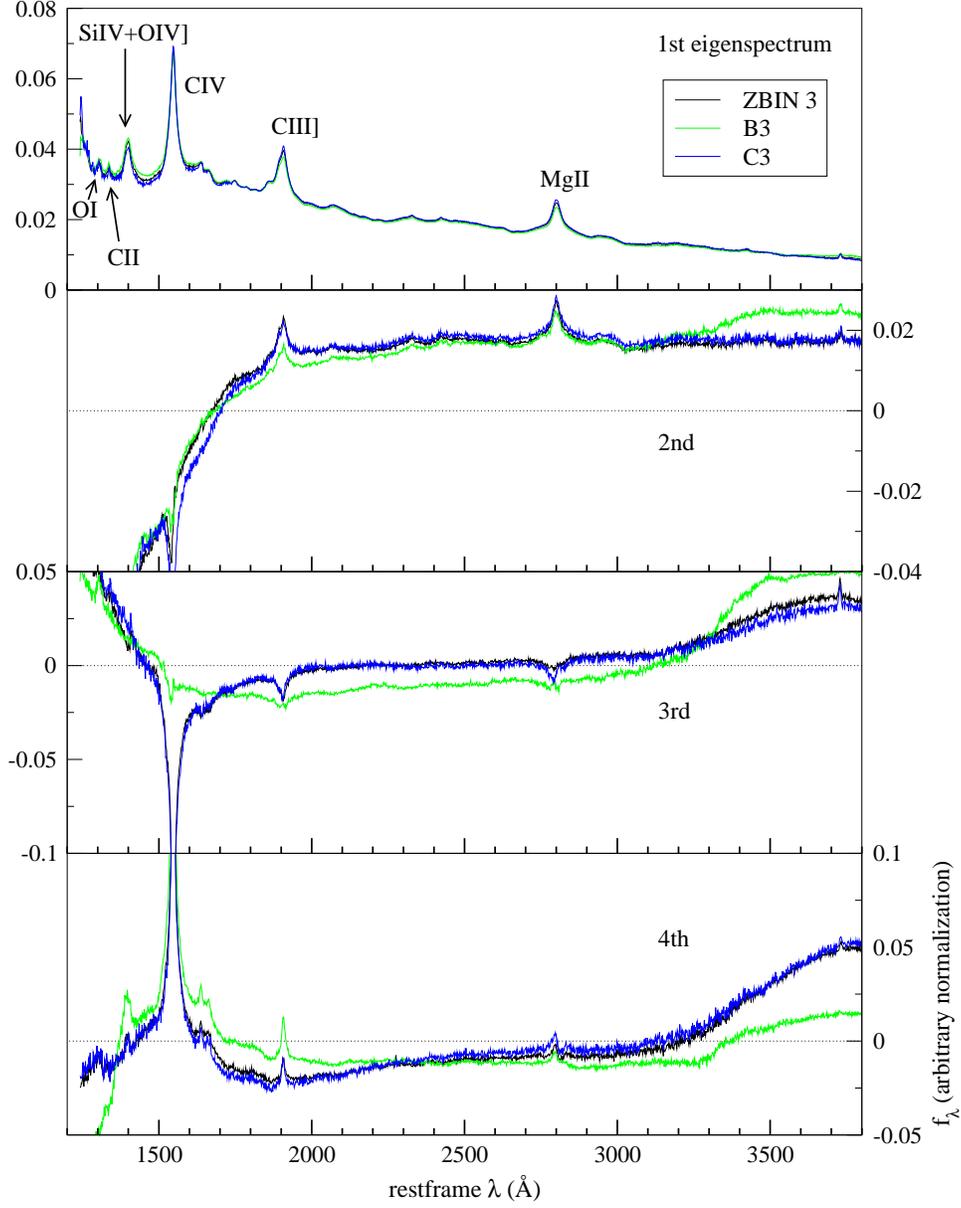
Figure 3.9 The first 4 orders of (M_i, z)-binned eigenspectra for the subsample **ZBIN 3**.
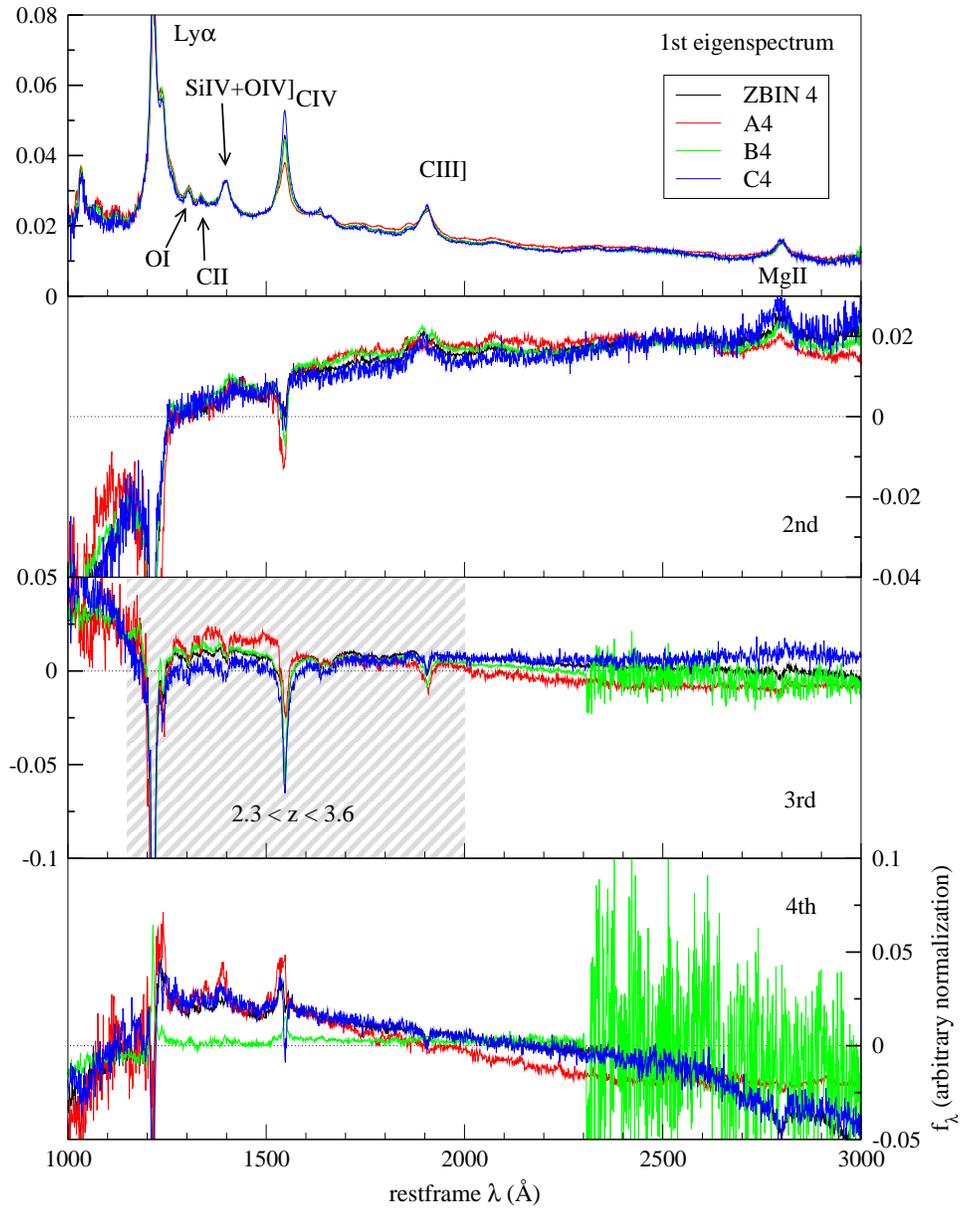
Figure 3.10 The first 4 orders of (M_i, z)-binned eigenspectra for the subsample **ZBIN 4**. The meaning of the shaded spectral region is explained in the caption of Fig. 3.7 and §3.6.4.
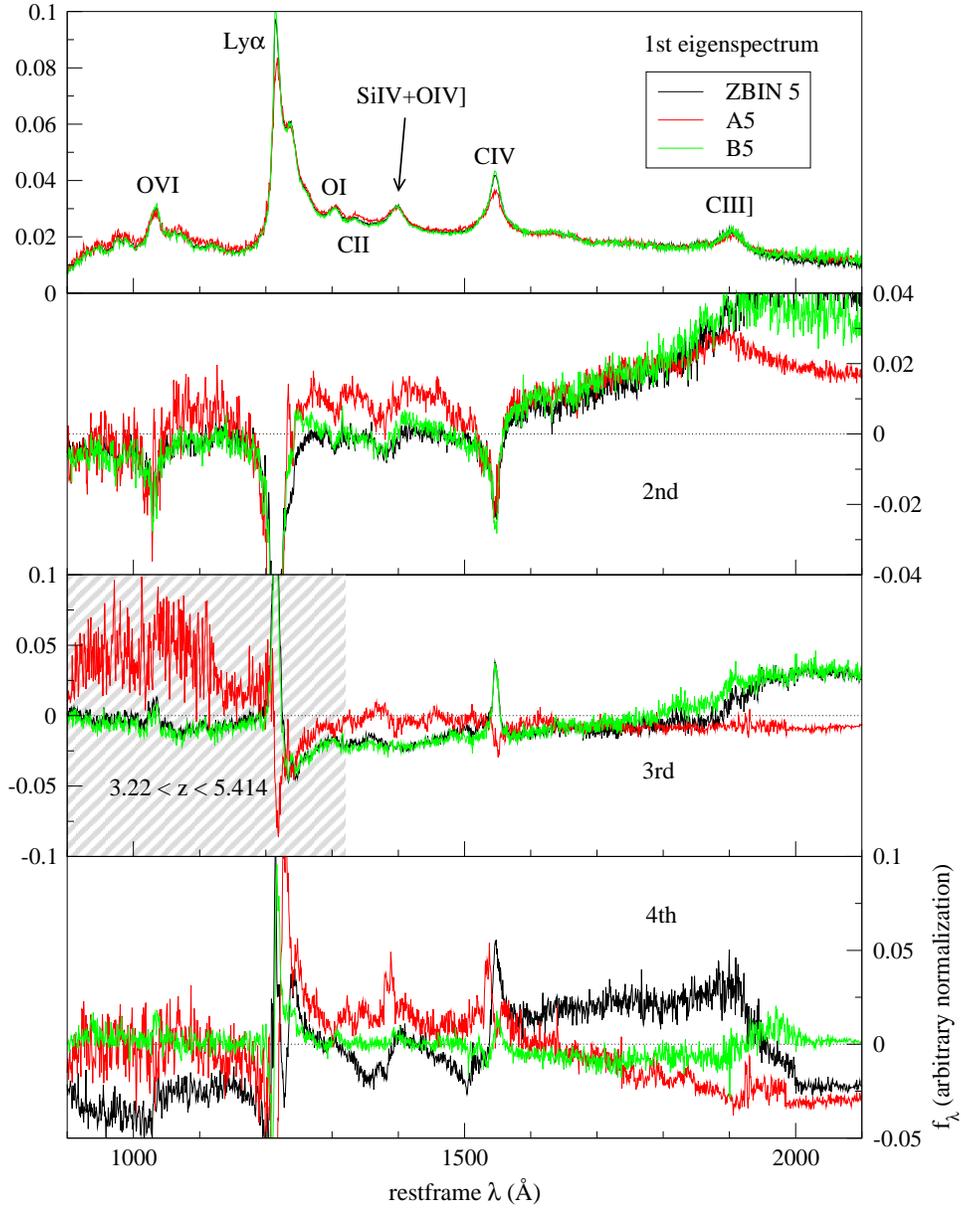
Figure 3.11 The first 4 orders of eigenspectra of the subsample **ZBIN 5**. The meaning of the shaded spectral region is explained in the caption of Fig. 3.7 and §3.6.4.

3.3.2 2nd mode: spectral slopes

The 2nd mode in every (M_i, z)-bin has one node at a particular wavelength. This implies that the linear-combination of the first 2 modes changes the spectral slope. This is similar to the galaxy spectral classification by the KL approach (Connolly et al. 1995), in which the first two eigenspectra give the spectral shape.

For the lowest redshift bin (**ZBIN 1**; Fig. 3.7), the node of the second eigenspectrum occurs at about 3850 Å for the lower luminosity QSOs (**D1**), but at $\sim$ 3300 Å for the higher luminosity ones (**C1**). Possible physical reasons underlying the modulation of the UV-optical slopes were discussed previously in §3.2.3. Interestingly, the luminosity averaged 2nd eigenspectrum (black curve) in this redshift range also shows galactic features (as found for the 2nd global eigenspectrum). The continuum redward of $\sim$ 4000 Å is very similar to that in galaxies of earlier-type. Absorption lines Ca K and Ca H, and the Balmer absorption lines H 9, H 10, H 11 and H 12 are seen in the lower-luminosity bin **D** (and are not present in the higher-luminosity bin **C**, hence a luminosity dependent effect is implied).

3.3.3 3rd mode: anti-correlation between Fe II (UV) and optical continuum around Hβ

In addition to the finer-modulation of the continuum slope provided by the 3rd eigenspectrum compared with the 2nd mode, in the redshift range $0.53 < z < 1.16$ (**ZBIN 2**; Fig. 3.8), averaging over all luminosities, this mode shows a strong anti-correlation between the quasi-continuum in the Fe II (UV) regions around Mg II (the "small bump", with its estimated location indicated in the 3rd eigenspectrum in Fig. 3.8) and the continuum in the vicinity of Hβ. Around the Hβ emission, the continuum is blended with the Fe II optical blends, the Hδ, Hγ and [O III] lines. The wavelength bounds are found to be $\sim 2120 - 4040$ Å for the Fe II ultraviolet blends and $\sim$ 4050 Å upward (to $\sim$ 6000 Å, which is the maximum wavelength of this redshift bin) for the optical continuum around Hβ. This appears to support the calculations that strong Fe II optical emissions require a high optical depth in the resonance transitions of the Fe II (UV) (Netzer and Wills 1983; Shang et al. 2003), hence a decrease in the strength of the latter. The actual wavelengths of the nodes bounding the

Fe II (UV) region are shown in Fig. 3.8. For brighter quasars (**B2**), the small bump is smaller ($\sim 2120 - 3280$ Å) than that found in fainter QSOs.

3.3.4   Reconstructing QSO with broad absorption lines

To examine the intrinsic broad absorption line features in the (M_i, z)-binned eigenspectra, we study the reconstructed spectra using different numbers of eigenspectra. Fig. 3.12 shows one of the EDR BAL quasars (Reichard et al. 2003) found in the bin **B3**, and its reconstructed-spectra using different numbers of eigenspectra. This HiBAL (defined as having high-ionization broad absorption troughs such as C IV) quasar is chosen for its relatively large absorption trough in C IV for visual clarity. The findings in the following are nonetheless general. The first few modes ($\lesssim 8$ for this spectrum) are found to fit mainly the continuum, excluding the BAL troughs. With the addition of higher-order modes the intrinsic absorption features (in this case, in the emission lines C IV and Si IV) are gradually recovered. Some intrinsic absorption features are found to require ~ 50 modes for accurate description, as was found in the global eigenspectra (§3.2.5). We should note that in the reconstructions using different numbers of modes; the *same* normalization constant is adopted (meaning the eigencoefficients are normalized to $\sum_{m=1}^{50} a_m^2 = 1$). Clearly, a different normalization constant in the case of reconstructions using fewer modes (e.g., Fig. 3.12a) will further improve the fitting in the least-squares sense.

While the fact that a large number of modes are required to reconstruct the absorption troughs probably suggests a non-compact set of KL eigenspectra (referring to those defined in this work) for classifying BAL quasars, the appropriate truncation of the expansion at some order of eigenspectra in the reconstruction process will likely lead to an *un-absorbed* continuum, invaluable to many applications. The proof of the validity of such a truncation will require detailed future analyses. One method is to construct a set of eigenspectra using only the known BAL quasars in the sample and to make comparisons between that and our current sets of eigenspectra. By comparing the different orders of both sets of eigenspectra we may be able to recover the BAL physics. We expect that this separate set of BALQSO-eigenspectra will likely reduce the number of modes in the reconstruction, which is desirable
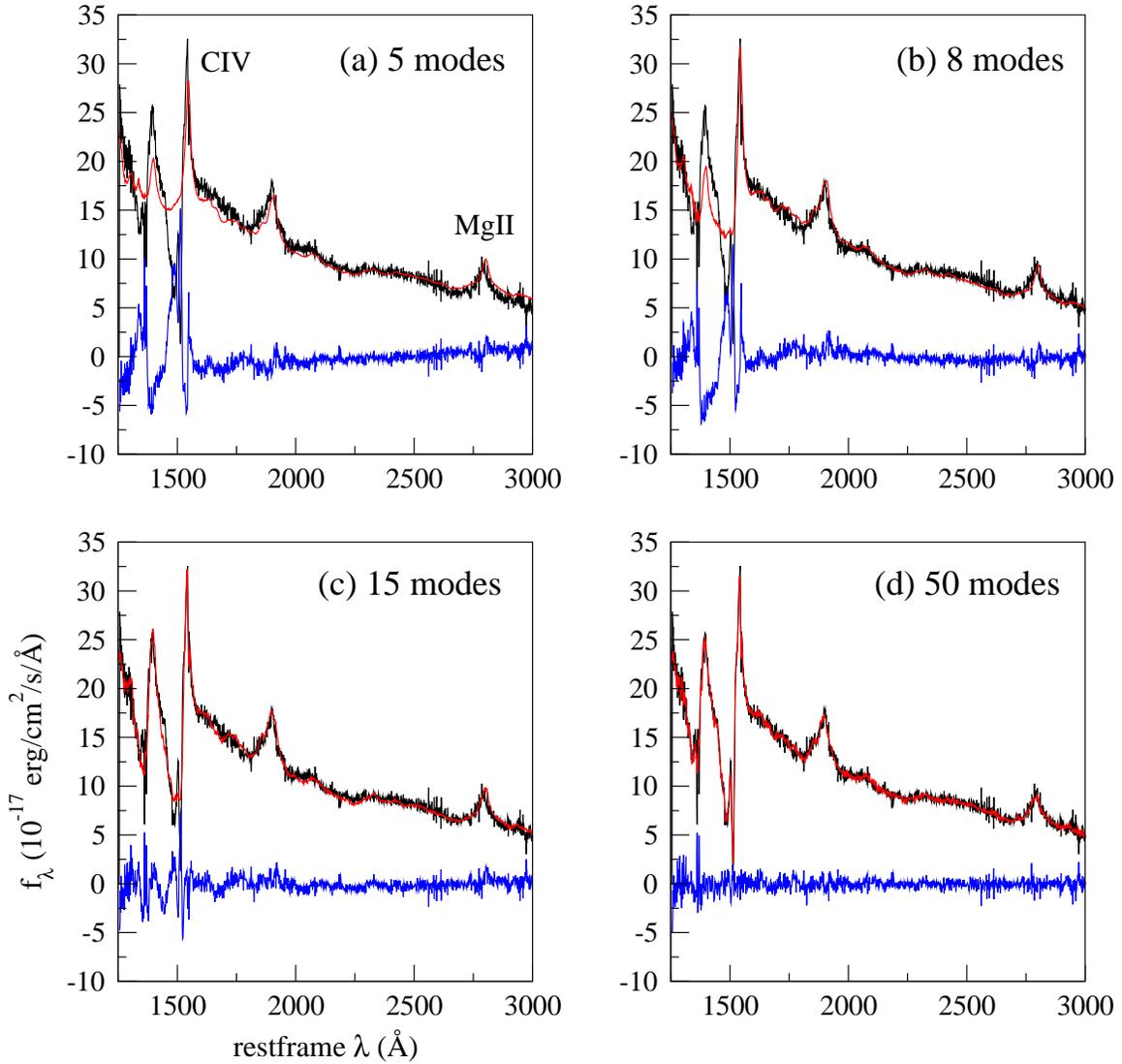
Figure 3.12 KL-reconstructed spectra (red) of a QSO spectrum (black) with broad absorption features in C IV and Si IV (SDSS J110041.19+003631.9, classified as HiBAL according to Reichard et al.) using the first (a) 5 modes (b) 8 modes (c) 15 modes (d) 50 modes, where the eigenspectra are constructed from the (M_i, z)-bin **B3**. The blue curves are the residuals. The broad absorption features span a number of higher-order modes. The observed spectrum is not smoothed.

from the point of view of classification.

3.3.5 KL-reconstructed spectra

Reconstructions of a typical non-BAL quasar spectrum are shown in Fig. 3.13, using from (a) 2 to (d) 20 orders of eigenspectra. This particular quasar is in the (M_i, z)-bin **C3**. The bottom curve in each sub-figure shows the residuals from the original spectrum. The first 10 modes are sufficient for a good reconstruction. The reconstructions of the same quasar spectrum but using the global set of eigenspectra are shown in Fig. 3.14, from (a) 2 modes to (f) 100 modes. To obtain the same kind of accuracy, more eigenspectra are needed in the global case; in this case about 50 modes. This is not surprising as the global eigenspectra must account for the intrinsic variations in the quasar spectra as well as any redshift or luminosity evolutions.

There are, therefore, two major factors we should consider when adopting a global set of quasar eigenspectra for KL-reconstruction and classification of quasar (instead of redshift and luminosity dependent sets). First, we need to understand and interpret about 10^2 global eigenspectra. This is significantly larger than found for galaxies (2 modes are needed to assign a type to a galaxy spectrum according to Connolly et al. 1995). This is a manifestation of the larger variations in the quasar spectra. Second, the "extrapolated" spectral region, $\lambda_{rest} < 1520$ Å, in Fig. 3.14 (which is the rest-wavelength region without spectral data) show an unphysical reconstruction even when 100 modes are used, although this number of modes can accurately reconstruct the spectral region with data. This agrees with the commonality analysis in §3.2.6, that there are evolutionary and luminosity effects in the QSOs in our sample. As such, eigenspectra derived in a particular redshift and luminosity range are in general not identical to those derived in another range.

The accuracy of the extrapolation in the no-data region using the KL-eigenspectra remains an open question for the (M_i, z)-bins. It will be an interesting follow-up project to confront the repaired spectral region with observational data, which ideally cover the rest-wavelength regions where the SDSS does not. For example, UV spectroscopic observations using the Hubble Space Telescope.
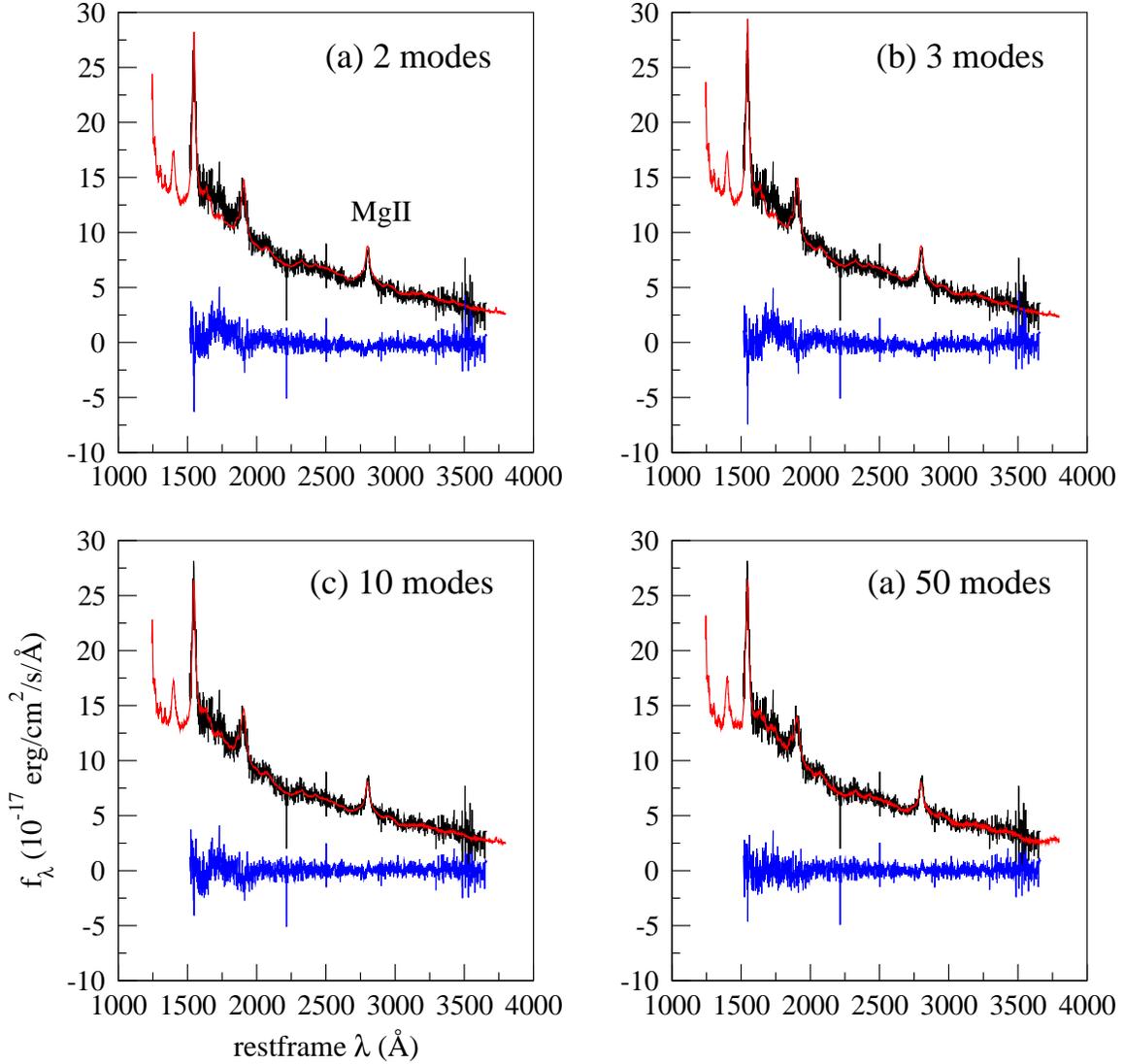
Figure 3.13 KL-reconstructed spectra (red) of an example QSO spectrum (black) (SDSS J015214.54+131532.0) using the first (a) 2 modes (b) 3 modes (c) 10 modes (d) 20 modes, where the eigenspectra are constructed from the (M_i, z)-bin **C3**. The blue curves are the residuals. The observed spectrum is not smoothed.
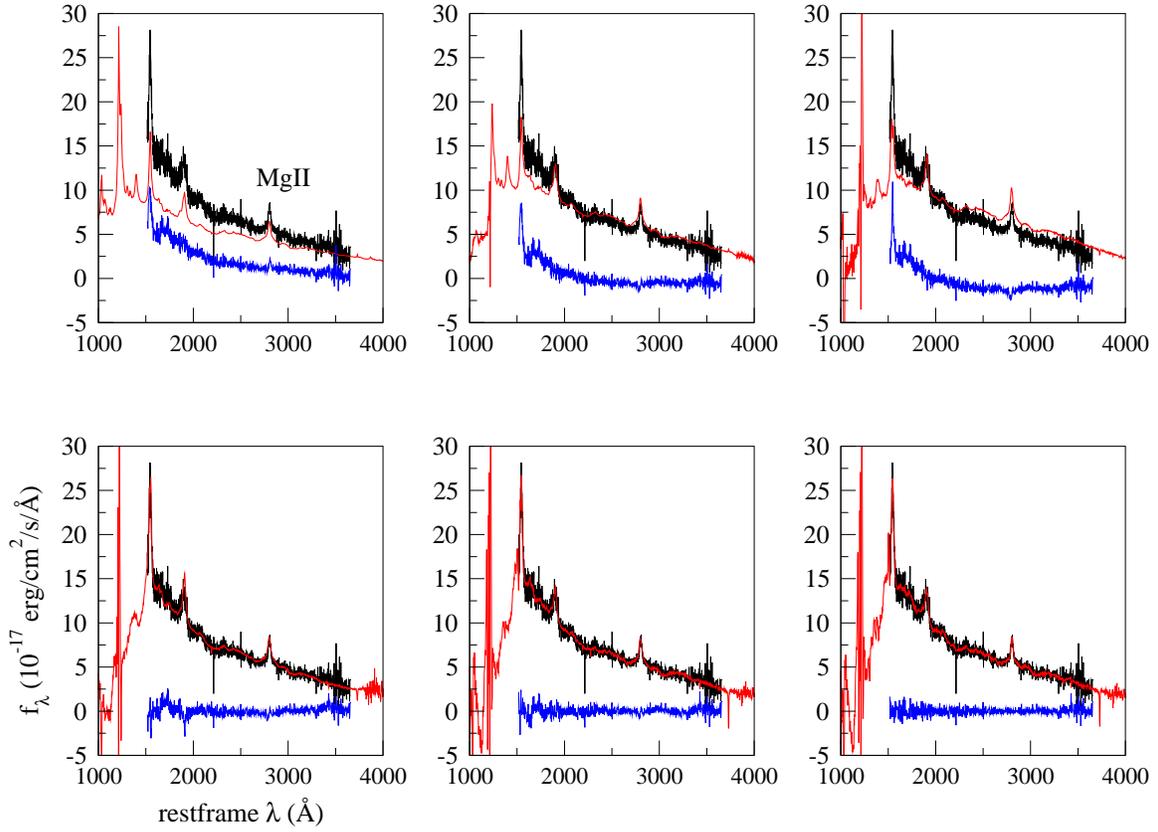
Figure 3.14 KL-reconstructed spectra (red) of the same quasar spectrum (black) as in Fig. 3.13 using the first (a) 2 modes (b) 3 modes (c) 10 modes (d) 20 modes (e) 50 modes and (f) 100 modes of the global eigenspectra. The blue curves are the residuals, and the observed spectrum is not smoothed.

3.4 EVOLUTIONARY AND LUMINOSITY EFFECTS

3.4.1 Cross-redshift and -luminosity bin projection

To study evolution in quasar spectra with the eigenspectra, we must ensure that the eigencoefficients reflect the same physics independent of redshift. We know however that the eigenspectra change as a function of redshift (see §3.2.6). To overcome this difficulty, and knowing that the overlap spectral region between the two sets of eigenspectra in any pair of adjacent redshift bins is larger than the common wavelength region ($2124 - 2486$ Å) for the full redshift interval, we study the differential evolution (in redshift) of the quasars by projecting the observed spectra at higher redshift onto the eigenspectra from the adjacent bin of lower redshift. In this way, the eigencoefficients can be compared directly from one redshift bin to the next.

Without the loss of generality, we project the observed quasar spectra in the higher redshift bin (or dimmer quasars for the cross-luminosity projection) onto the eigenspectra which are derived in the adjacent lower-redshift one (or brighter quasars for the cross-luminosity projection). For example, $spec(\mathbf{B3})$ (i.e., the spectra in the (M_i, z)-bin $\mathbf{B3}$) are projected onto $\{e(\mathbf{B2})\}$ (the set of eigenspectra from the (M_i, z)-bin $\mathbf{B2}$), and similarly for the different luminosity bins but the same redshift bin. From that, we can derive the relationship between the eigencoefficients and redshift (or luminosity).

3.4.2 Evolution of Iron

The most obvious evolutionary feature is the small bump present in the spectra at around $\lambda_{rest} \sim 2000$ Å to 4000 Å. This feature is mainly composed of blended Fe II emissions ($\sim 2000 - 3000$ Å, Wills et al. 1985) and the Balmer continuum ($\sim 2500 - 3800$ Å). When we project quasar spectra of redshifts $1.16 - 2.06$ (i.e., $spec(\mathbf{C3})$) onto eigenspectra constructed from quasars of redshifts $0.53 - 1.16$ (i.e., $\{e(\mathbf{C2})\}$), the coefficients from the second eigenspectrum show a clear trend with redshift, as shown in Fig. 3.15. In this figure, only those quasars with $M_i = -25.5 \pm 0.1$ are chosen (900 objects), as such the redshift trend does not primarily depend on the absolute luminosities of the quasars. To understand this

relation *observed* spectra are selected along the regression line in Fig. 3.15 (with the locations marked by the crosses) and are shown in Fig. 3.16. The two dotted lines mark the bandpass where the cross-redshift projection is performed. The small bump is found to be present and is prominent in the lower-redshift quasars, whereas it is small and may be absent in the higher-redshift ones. The spectra marked by the arrows in Fig. 3.16 lie relatively close to the regression line. An example of the range of evolution in the small bump as a function of redshift is shown by the remaining 3 spectra which deviate from the regression line. The observed evolution is present independent of which of the spectra we consider. The mean spectra (Fig. 3.17) as a function of redshift, constructed using a bin width in redshift (dz) of 0.2, show a similar behavior. Each mean spectrum is calculated by averaging the valid flux densities of all objects in each wavelength bin. The regression of the eigencoefficient-ratios with redshift (with outliers of $a_2/a_1 > 1$ removed from the calculation) is

$$(a_2/a_1)_{\{e(\mathbf{C2})\}} = -0.0820z + 0.0083 \; , \tag{3.1}$$

where the subscript $\{e(\mathbf{C2})\}$ denotes that the eigenspectra are from $\mathbf{C2}$. The correlation coefficient (r) is calculated to be 0.1206 with a two-tailed P-value[3] of 0.00027 (the probability that we would see such a correlation at random under the null hypothesis of $H_0 : r = 0$), as such the correlation is considered to be extremely significant by conventional statistical criteria.

This redshift dependency can be explained by either the evolution of chemical abundances in the quasar environment (Kuhn et al. 2001), or an intrinsic change in the continuum itself (which, of course, could also be due to the change in abundances through indirect photo-ionization processes). Green et al. (2001) found in the LBQS that the primary correlations of the strengths of Fe II emission lines are probably with redshift; an evolutionary effect is therefore implied. Kuhn et al. (2001) also supported the evolution of the small bump region $2200 - 3000$ Å from high-redshift ($\sim 3 - 4$) to lower-redshifts (< 0.3) by comparing two QSO subsamples with evolved luminosities.

As the second mode in the (M_i, z)-binned eigenspectra describes the change in the spectral slope of the sample, the above findings support the idea that the Balmer continuum, as

---

[3]The P-value for the t-test is calculated under the hypotheses $H_0 : r = 0$ and $H_1 : r \neq 0$.
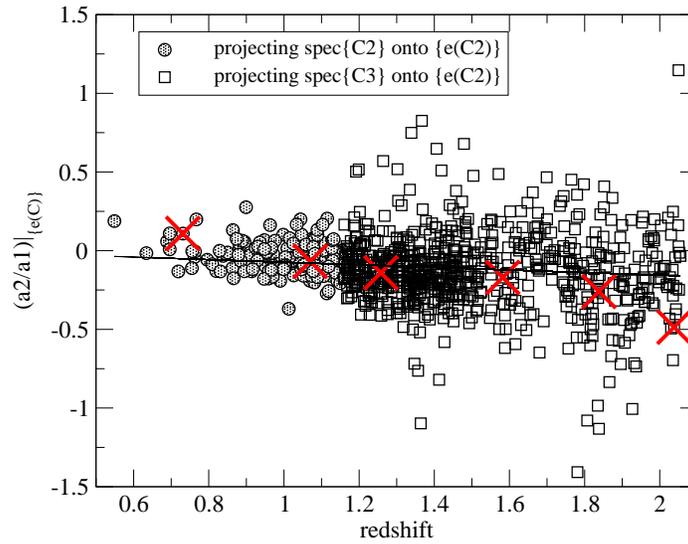
Figure 3.15 Dependence of a_2/a_1 on redshift, of quasars within the redshift range $0.53 - 2.06$, and at a fixed luminosity $M_i = -25.5 \pm 0.1$. The straight line is the regression of the data points: $(a_2/a_1)_{\{e(\mathbf{C2})\}} = -0.0820z + 0.0083$. The subscript $\{e(\mathbf{C2})\}$ denotes that the eigenspectra are constructed from the subsample $\mathbf{C2}$. The crosses mark the observed spectra illustrated in Fig. 3.16.
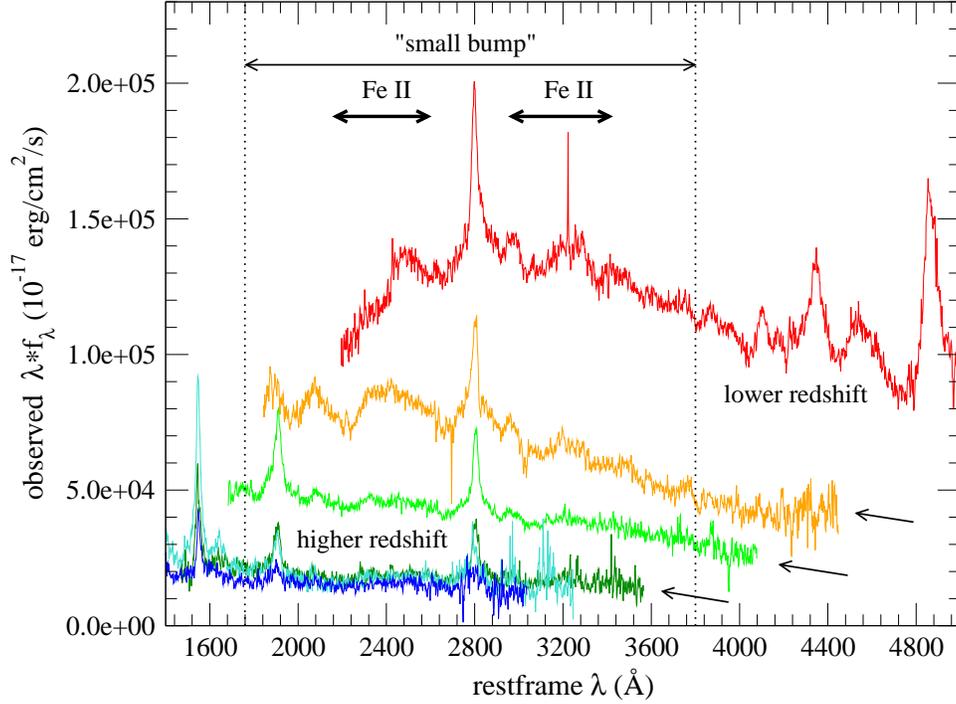
Figure 3.16 The observed quasar spectra located along the eigencoefficient-redshift relation in Fig. 3.15 in the original flux scales. The evolution of the small bump is evident, in that it is more prominent in the lower-redshift quasar spectra. The vertical dotted lines marked the common wavelength region on which the KL cross-projection is performed. The heavy arrows mark the wavelength regions in which the Fe II emissions are typically found. The spectra are smoothed with a FWHM = 3 Å Gaussian smoothing function for easier visualization. (The QSOs are, from the lowest to highest redshifts, SDSS J173052.71+602516.6, SDSS J015352.65-092010.7,   SDSS J090934.26+552944.1,   SDSS J033801.88+002718.8, SDSS J012858.45+152647.4, SDSS J021552.00-092310.3.)
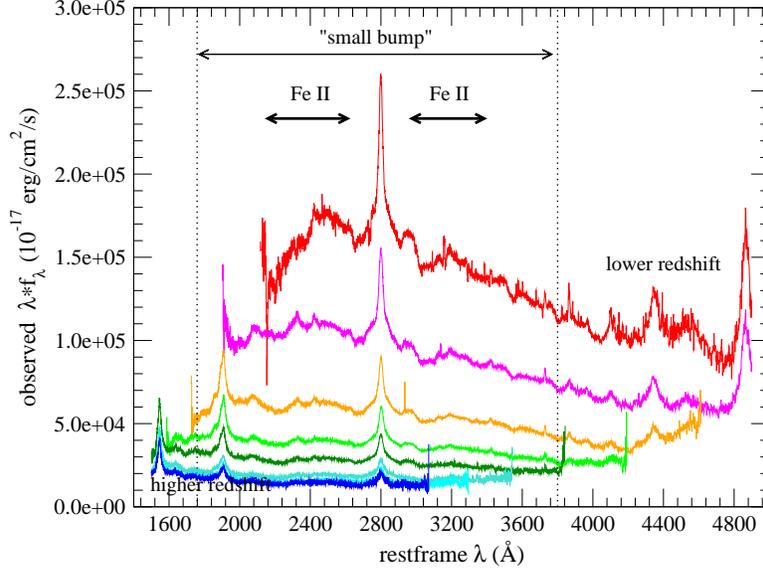
Figure 3.17 The mean spectra along the redshift trend in Fig. 3.15, from $z = 0.6$ to 2.2 with a bin width in redshift $dz = 0.2$. The observed spectra are not smoothed in the calculations of the mean spectra.

a part of the small bump, changes with redshift. To further understand this effect, the 3rd eigenspectrum in **C2** is taken into consideration, which presumably describes the iron lines (see §3.3.3). We find that the third eigencoefficient-ratio a_3/a_1 also shows a slight redshift dependency (not shown) with the regression relation (with outliers of $a_3/a_1 > 1$ removed from the calculation, resulting in 901 objects)

$$(a_3/a_1)_{\{e(\mathbf{C2})\}} = 0.0478z - 0.2063 \qquad (3.2)$$

and the correlation coefficient is calculated to be 0.0030 with a two-tailed P-value of 0.93, which is considered to be not statistically significant.

While the strength of this effect shown by the two ratios are of similar magnitude (0.0820 versus 0.0478), the difference in their correlation coefficients implies that the sample variation is much greater in the ratio a_3/a_1 than a_2/a_1. The non-trivial value of the regression slope in the case of a_3/a_1 agrees with the change in shape of the observed line profiles in the small bump regions seen in the local wavelength level (smaller in width than what is expected in the continuum change) with redshift. In conclusion, this implies that there exists the

81

possibility of an evolution in iron abundances but with a larger sample variation compared with that for the continuum change.

To our knowledge, our current analysis is the first one without invoking assumptions of the continuum level or a particular fitting procedure of the Fe II blends that finds an evolution of the small bump; directly from the KL eigencoefficients. Because of the large sample size, the conclusion of this work that the small bump evolves is drawn from spectrum-to-spectrum variation independent of the luminosity effect, in contrast to the previous composite spectrum approaches (Thompson et al. 1999), in which the authors found that the composite spectra in two subsamples with mean redshifts $< z >= 3.35$ and $< z >= 4.47$, and that from the Large Bright Quasar Survey of lower redshifts ($< z > \sim 0.8$) are similar in the vicinity of Mg II and hence did not suggest the existence of a redshift effect. The variation of the small bump with redshift is further confirmed with the study of composite quasar spectra of the DR1 data set (Vanden Berk et al. 2004b). At this point we make no attempt to quantitatively define and deblend the Fe II optical lines and the Balmer continuum, as that would be beyond the scope of this thesis. It is a well-known and unsolved problem to identify the true shape of total flux densities due to the Fe II emission lines. This difficulty arises because there are too many Fe II lines to model and they form a quasi continuum.

3.4.3   Luminosity dependence of broad emission lines

Luminosity effects on broad emission lines can also be probed in a similar way to the cross-redshift projection. One prominent luminosity effect is found by projecting $spec(\mathbf{D1})$ onto $\{e(\mathbf{C1})\}$. These samples have the same redshift range but different luminosities (for $\mathbf{D1}$, $M_i = (-24, -22)$ and for $\mathbf{C1}$, $M_i = (-26, -24)$). Fig. 3.18 shows the eigencoefficient $(a_2/a_1)_{\{e(\mathbf{C1})\}}$ as a function of absolute luminosity, with redshifts fixed at $z = 0.4 \pm 0.02$ (235 quasars). The ratio of the first 2 eigencoefficients decreases with increasing quasar luminosity. The regression line (with outliers of $a_2/a_1 > 1$ removed from the calculation) is

$$(a_2/a_1)_{\{e(\mathbf{C1})\}} = 0.0643 M_i + 1.5797 \ , \tag{3.3}$$

with a correlation coefficient of 0.2305 with an extremely significant two-tailed P-value of 0.0003.
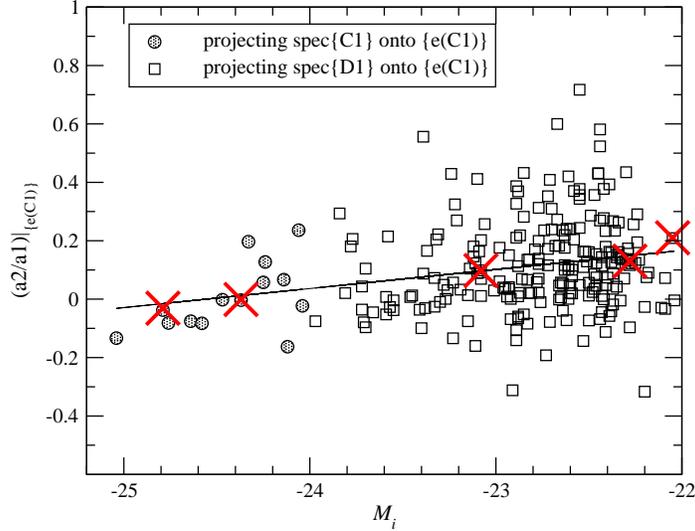
Figure 3.18 Dependence of a_2/a_1 on the i-band absolute luminosity, of quasars with fixed redshifts ($z = 0.4 \pm 0.02$). The regression line is: $(a_2/a_1)_{\{e(\mathbf{C1})\}} = 0.0643 M_i + 1.5797$. The crosses mark the observed spectra plotted in Fig. 3.19a.

Along this luminosity trend, the equivalent widths (EWs) of emission lines such as Hβ and [O III] lines are found to decrease typically, as a function of increasing absolute magnitude M_i (as shown in the spectra in Fig. 3.19a). This is the Baldwin (1977) effect. We note that the host-galaxy may come into play in this case (at low redshifts and low luminosities). The geometric composite spectra of different luminosities within the range from -22 to -25 are shown in Fig. 3.19b, in which a spectral index of $\alpha_\nu = -0.5$ for the continua is assumed. The Baldwin effect for the emission lines is also present.

In the highest redshift bins, the Baldwin effect can be found in the first and the second eigenspectra. Fig. 3.10 shows that the addition (with positive eigencoefficients) of the first two eigenspectra *enhances* the flux density around 1450 Å and reduces the equivalent width of C IV. Lyα and other major BELs are also shown to be anti-correlated with the continuum flux. Hence, the Baldwin effect is not limited to the C IV emission line, and is also observed in many broad emission lines (see, for example, a summary in Sulentic et al. 2000). The linear-combination of the first and third modes in this redshift range also shows a similar modulation between the flux density around 1450 Å and the line equivalent width. This
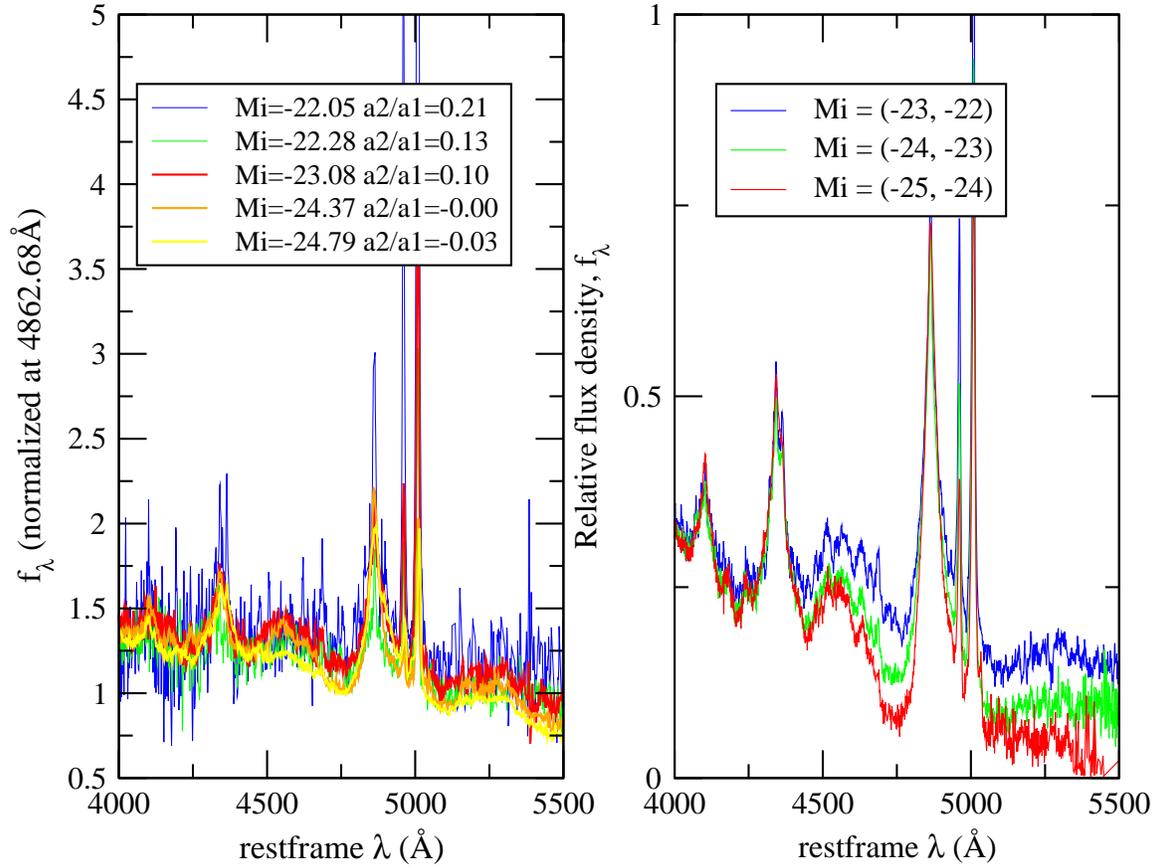
Figure 3.19 (a) The $H\beta$ region of spectra located along the regression line for $(a_2/a_1)_{\{e(\mathbf{C1})\}}$ versus M_i (marked with crosses in Fig. 3.18). The spectra are normalized by the continuum flux density at $\lambda = 4862.68$ Å and smoothed with a FWHM = 3 Å Gaussian smoothing function. For decreasing a_2/a_1 ratio, i.e., for brighter quasars, the emission line equivalent widths typically decrease. (The QSOs are, starting from the brightest, SDSS J013418.19+001536.6, SDSS J010342.73+002537.2, SDSS J011310.38-003133.1, SDSS J093409.17+023237.0, SDSS J092011.60+571718.2.) (b) The geometric composite spectra in different absolute luminosity bins with M_i ranges from -22 to -25 for the objects in Fig. 3.18.
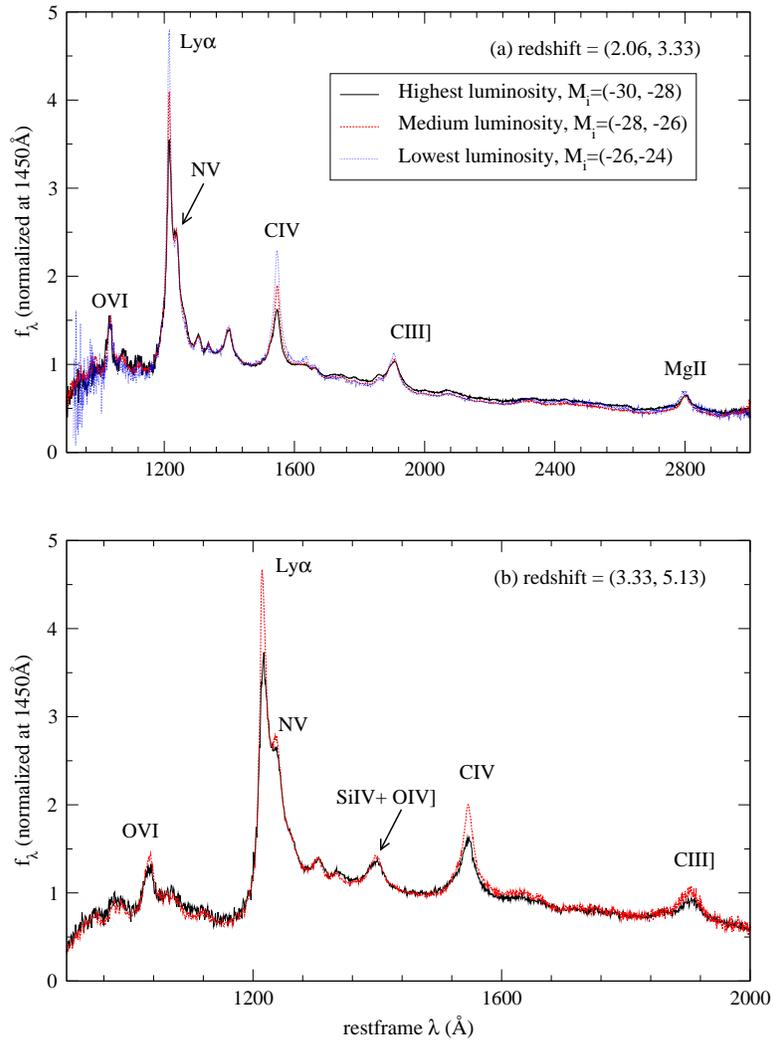
Figure 3.20 Baldwin effect of the broad emission lines.

effect is, however, not general for all luminosities, with the third eigenspectrum in **C4** showing only a small value in the 1450 Å flux density.

The Baldwin effect can also be seen by comparing the first eigenspectra constructed for different luminosity bins. Fig. 3.20 shows the first eigenspectra derived in different luminosities in the second highest redshift bin (i.e., the (M_i, z)-bins **A4**, **B4** and **C4**, with $2.06 < z < 3.33$) and the highest one (**A5** and **B5,** with $3.33 < z < 5.13$). The eigenspectra are normalized to unity at 1450 Å. The continua for wavelengths approximately greater than 1700 Å in Fig. 3.20a are not perfectly normalized (which is difficult to define in the first place), but a more careful normalization would only lead to an increase in the degree of the Baldwin effect in the emission lines C III] and Mg II. The Lyα and C IV lines demonstrate the most profound Baldwin effect. Other broad emission lines such as He IIλ1640, C III] and Mg II also exhibit this effect. For the controversial line N V, an "anti-Baldwin" correlation is found at redshifts $2.06 - 3.33$, such that flux densities are smaller for lower-luminosity quasars. At the highest redshifts in this study ($z = 3.33 - 5.13$, Fig. 3.20b), however, a normal Baldwin effect of N V is found. The redshift dependency in the Baldwin effect for N V may explain the contradictory results found in previous studies (a detection of Baldwin effect of N V in Tytler and Fan 1992; and non-detections in Steidel and Sargent 1991; Osmer et al. 1994; Laor et al. 1995). While most studies have shown little evidence of the Baldwin effect in the blended emission lines Si IV+O IV], our results support the existence of an effect (though at a much weaker level than that of Lyα and C IV). This is in agreement with two previous works: Laor et al. (1995) which used 14 HST QSOs, and Green et al. (2001) which used about 400 QSOs from the LBQS). In the optical region, at least He IIλ4687 was reported to show the Baldwin effect (Heckman 1980b; Boroson and Green 1992; Zheng and Malkan 1993).

To further verify that the luminosity dependency of the eigencoefficients implies a Baldwin effect, we also study the eigencoefficients corresponding to the Baldwin effect seen in Fig. 3.20. We find that when $spec(\mathbf{C4})$ are projected onto $\{e(\mathbf{B4})\}$ the luminosity dependency is also seen in the eigencoefficients, with $(a_2/a_1)_{\{e(\mathbf{B4})\}} = 0.0327M_i + 0.8794$ ($r = 0.1150$, and an insignificant two-tailed P-value of 0.14) and $(a_3/a_1)_{\{e(\mathbf{B4})\}} = -0.0616M_i - 1.6948$ ($r = 0.2177$, and a very significant two-tailed P-value of 0.0043), both for objects

with redshifts within 2.7 ± 0.1 (161 objects in the case of a_2/a_1 and 166 in that of a_3/a_1).

3.5 A SPECTRAL SEQUENCE ALONG EIGENCOEFFICIENTS $(a_1,\ a_2)$ IN $(M_i,\ z)$-BIN

Fig. 3.21 shows plots of the first five eigencoefficients of the (M_i, z)-bin **B3**, where the properties are typical for all (M_i, z)-bins. The eigencoefficients are normalized as: $\sum_{m=1}^{50} a_m^2 = 1$. The plot of a_2 versus a_1 shows a continuous progression in the ratio of these coefficients which is similar to that found in the KL spectral classification of galaxies (Connolly et al. 1995), in which the points fall onto a major "sequence" of increasing spectral slopes. As higher orders are considered, for example a_5 vs a_4 (Fig. 3.21d), no significant correlations are observed.

Observed quasar spectra are inspected along this trend of a_2 versus a_1 (Fig. 3.22). The top of each sub-figure shows the values of (a_1, a_2). Along the sequence with decreasing a_2 values, the quasar continua are progressively bluer. The relatively red continua in Fig. 3.22a to 3.22c may be due intrinsic dust obscuration (Hall et al. 2002). The quasar in Fig. 3.22c is probably a high-ionization BALQSO (HiBAL) according to the supplementary SDSS EDR BAL quasar catalog (Reichard et al. 2003). We do, however, emphasize that the appearance of this BALQSO (or any BALQSO in general) in this particular sequence of quasar in the a_2 versus a_1 plane does not imply two modes are enough to achieve an accurate classification for a general BALQSO (for the reasons described in §3.3.4). The steepness of the spectral slope of this particular BALQSO is the major reason which causes such values of a_1 and a_2 eigencoefficients.

On the variations of the emission lines along these major (M_i, z) sequences, we can appreciate some of the difficulties in obtaining a *simple* classification concerning *all* emission lines by inspecting the examples listed in Table 3.5. The addition of the 2nd eigenspectrum to the 1st, weighted with (signed) medians of the eigencoefficients for all objects in a given sample, broadens some emission lines while making others narrower; a similar effect is seen for the addition of the 3rd eigenspectrum to the 1st, but in two *different* sets of lines. This
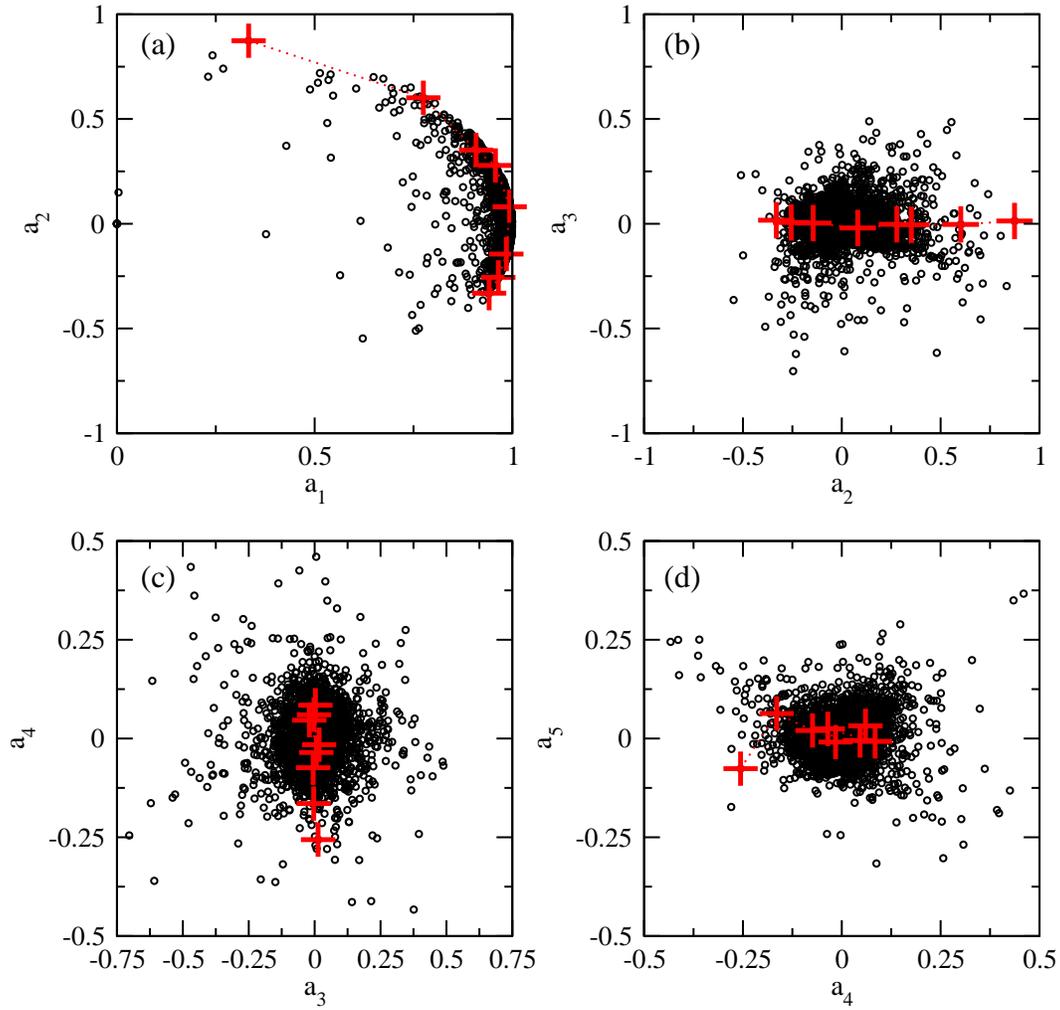
Figure 3.21 The distributions of the first 5 eigencoefficients of the (M_i, z)-bin **B3**, in (a) a_2 versus a_1 (b) a_3 versus a_2 (c) a_4 versus a_3 and (d) a_5 versus a_4. The crosses mark the observed QSO spectra illustrated in Fig. 3.22.
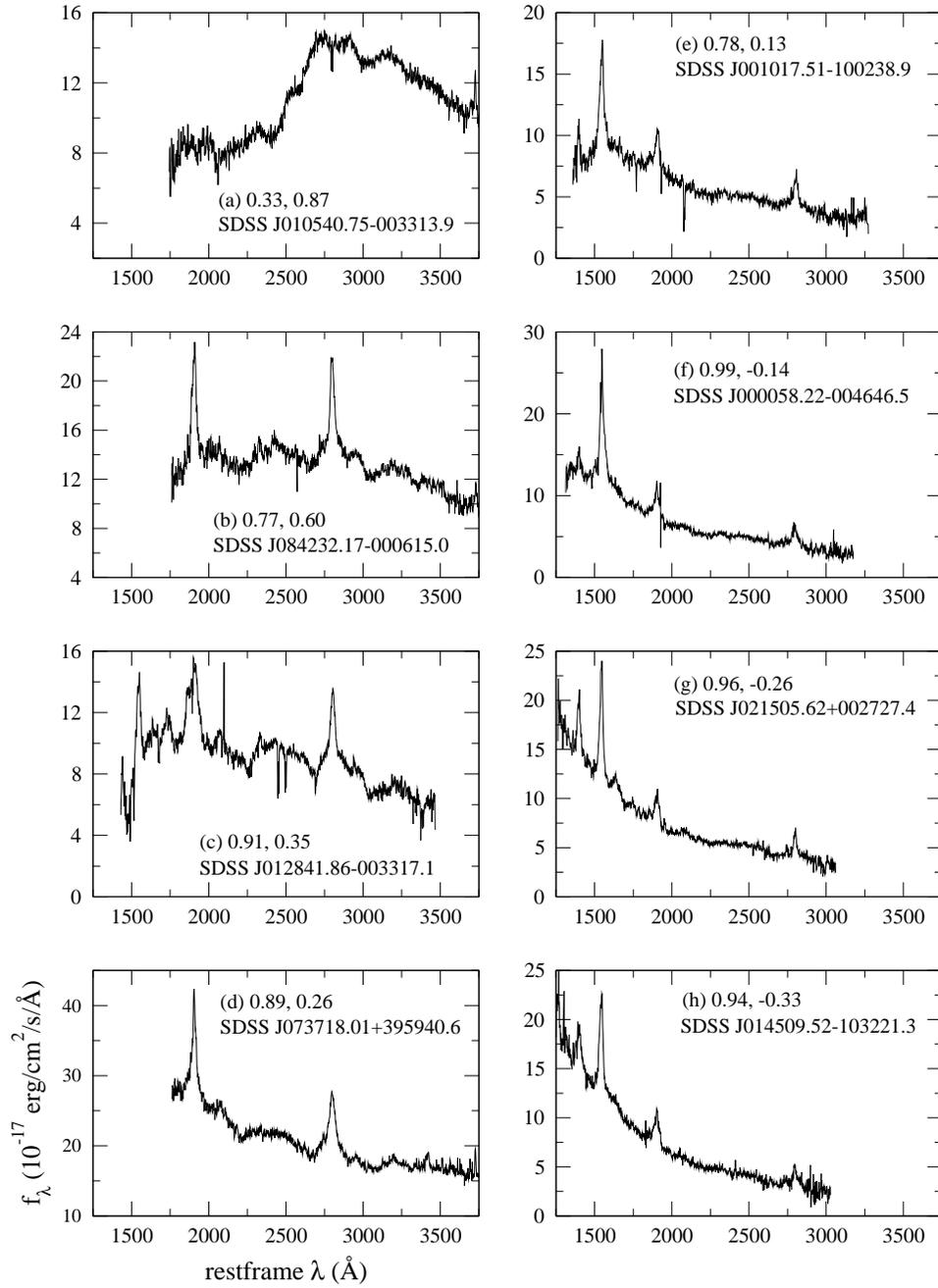
Figure 3.22 The observed quasar spectra picked along a sequence formed by the first two eigencoefficients a_1 and a_2 valued at the two numbers in each figure (see Fig. 3.21 for actual locations of a_1 to a_5). Proceeding along the sequence, the spectral slopes of quasars progressively vary from redder to bluer. The spectra are smoothed by a FWHM = 3 Å Gaussian smoothing function for easier visualization. The (M_i, z)-bin under consideration is **B3**.

Table 3.5 The average FWHMs of major broad emission lines of the **ZBIN** QSO eigenspectra.

Percentage of Lines[b]	1st−3rd	1st−2nd	1st	FWHM[a](km s^{-1}) (**ZBIN**) 1st+2nd[c]	1st+3rd	
Narrower than 1st mode	77%	40%	$\cdots$	61%	39%	
Wider than 1st mode	23%	60%	$\cdots$	39%	61%	
Example:						Redshift bin
Lyα+N V $(1160, 1290)$Å[d]	8999	3343	4405	7918[e]	3116	**ZBIN 5**
C IV $(1494, 1620)$Å	3579	4110	4140	4179	4913	**ZBIN 4**
C III] $(1830, 1976)$Å	5291	5597	5905	6296	6588	**ZBIN 4**
Mg II $(2686, 2913)$Å	4274	4318	4204	4096	4126	**ZBIN 3**
Hβ $(4050, 4152)$Å	1661	2035	1795	1443	1998	**ZBIN 1**
[O III]λ5008 $(4982, 5035)$Å	493	494	495	495	497	**ZBIN 1**

---

[a] The values are to the nearest unity.

[b] Only the emission lines with line-widths > 1000 km s^{-1} are counted, in both cases of reconstructions using the first mode and the concerned linear-combination.

[c] The expansion coefficients in the linear-combination are taken to be the (signed) medians of the eigencoefficients of all objects in the concerned sample.

[d] The restframe wavelength window across which the continuum underneath the line is approximated by linear-interpolation.

[e] Though not completely deblended from N V, the 2nd eigenspectrum mainly contains the Lyα component.

shows the large intrinsic variations in the emission line-widths of the QSOs.

3.5.1 host-galaxy removal and spectral sequence

In §3.3.2 we describe that 2nd eigenspectrum in the (M_i, z)-bin show host-galaxy features. To apply the eigenspectrum in host-galaxy removal, we construct the "cleaned" QSO spectrum as follows

$$f_\lambda(\text{QSO; cleaned}) = f_\lambda(\text{QSO; observed}) - a_2 e_2(\text{host galaxy}) . \qquad (3.4)$$

Since the host galaxy is prominently represented by the 2nd eigenspectrum, Eqn. 3.4 is a good assumption to the first order.

For this part of the work we use the SDSS DR3 catalog (Schneider et al. 2005) to maximize the number of objects. The sample contains $\sim 5,000$ QSOs in the redshifts of $0.2 - 0.6$, so that all the QSOs have full wavelength coverage in the restframe wavelengths $4,000 - 5,500$ Å. The wavelength range is considered for the most prominent host-galaxy continuum.

Firstly, the cleaned QSO spectrum reveals a power-law continuum (Fig. 3.24). There are some minor unphysical residuals of narrow emission lines remain, which may hinder the application of this method to some line diagnostics. The most prominent sample variation along a_1 is an increase in the QSO luminosity.

In Fig. 3.24 we show the spectral sequence along a_3. There is an anti-correlation seen between the Fe II emissions and [O III]λ5008. The four "extreme" spectral types picked respectively at the corners of the parameter space a_3 vs. a_1 are shown in Fig. 3.25. The top/bottom panels show clear change in the Fe II emissions; while the left/right panels show change in the QSO luminosity.
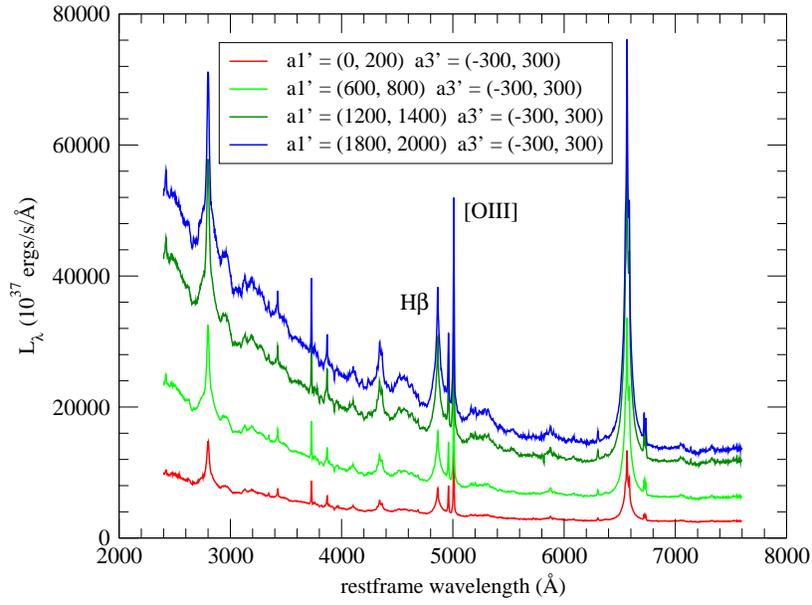
Figure 3.23 Spectral sequence along a_1, in the vicinity of Hβ.
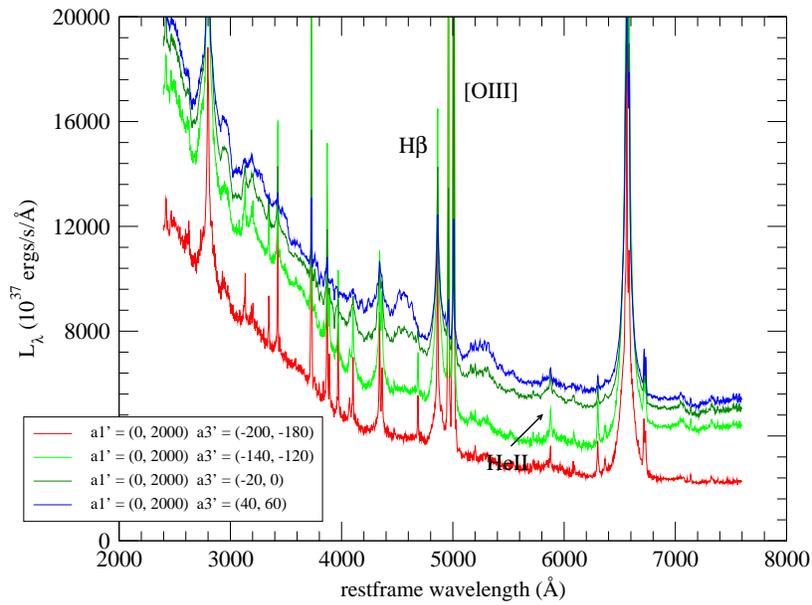
Figure 3.24 Spectral sequence along a_3, in the vicinity of Hβ.
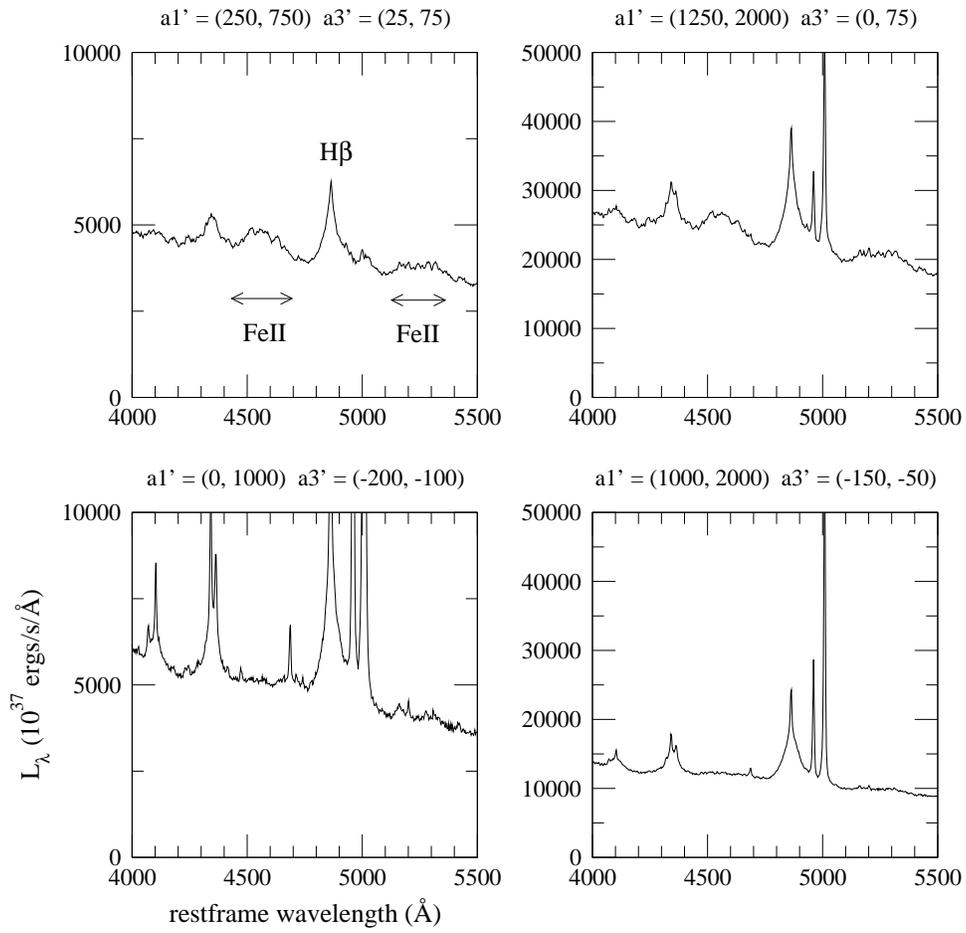
Figure 3.25 Four extreme spectral types picked on the plane a_3 vs. a_1, in the vicinity of Hβ and after the host-galaxy removal. The y-scale in the left and right panels are different.

3.6 LOCAL QSO EIGENSPECTRA

One of the utilities of the KL transform is to study the linear correlations among the input parameters, in this case, the pixelized flux densities in a spectrum. Due to possible uncertainties in any continuum fitting procedure in quasar spectra and the fact that no quasar spectrum in our sample completely covers the rest wavelength range $900 - 8000$ Å, correlations among the broad emission lines are first determined locally around the lines of interest by studying the first two eigenspectra in a smaller restricted wavelength range using the wavelength-selected QSO spectra. This process is then repeated from 900 Å to 8000 Å. Each local wavelength region is chosen to be $\sim 500 - 1800$ Å wide in the restframe. Empirically, we find that at these spectral widths the correlations among broad emission lines can be isolated in the first two eigenspectra without interference by the continuum information (except in the vicinity of Mg II doublet, for which the adjacent strong emission lines are located well beyond the Fe II (UV) region, which can be as broad as $\lambda_{rest} \sim 2000 - 4000$ Å), *in contrast* to the property of the (M_i, z)-bins in which the 2nd eigenspectra generally describe the variations in the spectral slopes.

The actual procedures to determine the correlations among the strengths of the major emission lines are as follows: (i) in each bin, the eigencoefficients of all objects are computed, and the distribution of the first two eigencoefficients, a_2 versus a_1, are divided into several (~ 10) sections within $\pm 1\sigma$ of the a_2 distribution. In each section the mean eigencoefficients, $< a_1 >$ and $< a_2 >$, are calculated (discarding outliers $a_1 < 0$). (ii) Along this trend of mean eigencoefficients, synthetic spectra are constructed by the linear-combination of the first two eigenspectra using the weights defined by the mean eigencoefficients. (iii) The equivalent widths of emission lines in the synthetic spectra are calculated along the trend of mean eigencoefficients, so that the correlations among the strengths of the broad emission lines can be deduced. Linear regression and linear correlation coefficients are calculated from the EW-sequence of a particular emission line relative to that of another line, which is fixed to be the emission line with the shortest wavelength of each local bin. The equivalent widths are calculated by direct summation over the continuum-normalized flux densities within appropriate wavelength windows. From such procedures, the correlations found are

ensemble-averaged properties of redshifts and luminosities over the corresponding range, and are physical. Table 3.6 shows the rest-wavelength bounds, the redshift range, the number of quasar spectra in each bin, and regression and correlation coefficients for each major emission line. The range of the possible restframe equivalent widths (EW_{rest}) along (a_1, a_2) is listed in decreasing a_2 values. Since the redshifts are chosen such that each quasar spectrum has a full coverage in the corresponding wavelength region, the gap-correcting procedure is implemented to correct only for skylines and bad pixels.

Table 3.6.   Correlations among emission lines in local QSO eigenspectra.

Line	λ_{lab} (Å)	$(\lambda_{low}, \lambda_{upp})$[a] (Å)	λ_{rest}[b](Å)	z (num. of obj.)[b]	EW_{rest}[c](Å)	Reg. Coeff.[f]	Corr. Coeff.[f]	P-value[g]
Lyβ+OVI	1025.72+1033.83	(1012, 1055)	900 − 1320	3.22 − 5.414 (496)	9.08 − 14.30	1.0000	1.0000	0.000000
Lyα+NV	1215.67+1240.14	(1160, 1290)			74.17 − 140.28	12.6411	0.9985	0.000000
NV	1240.14	(1230, 1252)			2.31 − 1.05	−0.2405	0.9332	0.000080
OI+SiII	1304.35+1306.82	(1290, 1318)			1.10 − 2.59	0.2833	0.9931	0.000000
Lyα+NV	1215.67+1240.14	(1160, 1290)	1150 − 2000	2.3 − 3.6 (1277)	122.96 − 65.82	1.0000	1.0000	0.000000
NV	1240.14	(1230, 1252)			1.87 − 2.06	−0.0032	0.9760	0.000001
OI+SiII	1304.35+1306.82	(1290, 1318)			2.40 − 1.22	0.0206	0.9999	0.000000
CII	1335.30	(1325, 1348)			0.72 − 0.67	0.0009	0.9862	0.000000
SiIV+OIV]	1396.76+1402.06	(1360, 1446)			9.12 − 9.24	−0.0021	0.9812	0.000001
CIV	1549.06	(1494, 1620)			35.45 − 17.56	0.3129	0.9997	0.000000
HeII	1640.42	(1622, 1648)			1.06 − 0.35	0.0125	0.9921	0.000000
OIII]+AlIII+FeII(UV40)	1664.74[d]	(1648, 1682)			1.24 − 0.13	0.0195	0.9534	0.000020
NIII]	1750.26	(1735, 1765)			0.74 − 0.30	0.0077	0.9966	0.000000
FeII(UV191)	1788.73[d]	(1771, 1802)			0.59 − 0.25	0.0059	0.9978	0.000000
AlIII	1857.40	(1840, 1875)			0.49 − 0.66	−0.0030	0.9601	0.000011
CIII]+Iron lines	1905.97[d]	(1830, 1976)			24.93 − 21.82	0.0541	0.9891	0.000000
AlIII	1857.40	(1840, 1875)	1800 − 3200	1.1 − 1.87 (6998)	0.44 − 0.53	1.0000	1.0000	0.000000
CIII]+Iron lines	1905.97[d]	(1830, 1976)			25.47 − 23.14	−26.7053	0.9962	0.000000
FeIII(UV48)	2076.62	(2036, 2124)			2.67 − 3.06	4.4128	0.9999	0.000000
CII	2326.44	(2312, 2338)			0.42 − 0.35	−0.8192	0.9933	0.000000
[NeIV]+FeIII(UV47)	2423.46[d]	(2402, 2448)			0.83 − 0.85	0.1746	0.9987	0.000000
MgII	2798.75	(2686, 2913)			35.81 − 33.98	−20.8265	0.9973	0.000000
OIII+FeII(Opt82)	3127.70[d]	(3100, 3153)			0.66 − 0.48	−1.9915	0.9883	0.000000
MgII	2798.75	(2686, 2913)	2600 − 4250	0.46 − 1.16 (4647)	34.97 − 42.33	1.0000	1.0000	0.000000
OIII+FeII(Opt82)	3127.70[d]	(3100, 3153)			1.34 − 1.05	−0.0403	0.9903	0.000000
[NeV]	3346.82	(3329, 3356)			0.60 − 0.22	−0.0522	0.9404	0.000051
[NeV]	3426.84	(3394, 3446)			2.04 − 0.99	−0.1434	0.9619	0.000009
[OII]	3728.48	(3714, 3740)			3.36 − 0.71	−0.3591	0.9041	0.000329
[NeIII]	3869.85	(3850, 3884)			2.34 − 0.87	−0.1986	0.9441	0.000040
[NeIII]+Hε	3968.58+3971.20	(3950, 3978)			0.73 − 0.33	−0.0533	0.9593	0.000011
Hδ	4102.89	(4050, 4152)			4.83 − 6.32	0.2001	0.9996	0.000000
Hδ	4102.89	(4050, 4152)	4000 − 5500	0.08 − 0.67 (2534)	6.96 − 5.30	1.0000	1.0000	0.000000
Hγ	4341.68	(4285, 4412)			14.72 − 11.08	2.1956	1.0000	0.000000
[OIII]	4364.44	(4352, 4372)			1.09 − 0.24	0.5133	0.9520	0.000022
FeII(optical, blended lines)[e]	4600.00	(4469, 4762)			14.69 − 21.28	−3.9796	0.9795	0.000001
HeII	4687.02	(4668, 4696)			0.99 − 0.03	0.5797	0.8904	0.000551
Hβ	4862.68	(4760, 4980)			59.38 − 43.26	9.7307	0.9999	0.000000
[OIII]	4960.30	(4945, 4972)			10.82 − 0.08	6.4826	0.8847	0.000671
[OIII]	5008.24	(4982, 5035)			40.46 − 2.64	22.8252	0.9063	0.000301
FeII(optical, blended lines)[e]	5250.00	(5100, 5477)			13.80 − 26.49	−7.6594	0.9600	0.000011
[FeVII]+FeII(Opt49)	5277.92[d]	(5273, 5287)			0.16 − 0.13	0.0204	0.9999	0.000000
[FeVII]+FeII(Opt49)	5277.92[d]	(5273, 5287)	5200 − 7000	0.08 − 0.31 (456)	0.15 − 0.18	1.0000	1.0000	0.000000
HeI	5877.29	(5805, 5956)			8.60 − 3.09	−258.9123	0.9423	0.000045
Hα+[NII]	6564.61+6585.28	(6400, 6765)			340.52 − 138.68	−9493.8027	0.9525	0.000021
[NII]	6585.28	(6577, 6593)			1.21 − 2.63	65.5191	0.9796	0.000001
[SII]	6718.29	(6708, 6726)			1.28 − 1.13	−7.2297	0.9967	0.000000
[SII]	6732.67	(6726, 6742)			0.91 − 0.65	−12.5154	0.9888	0.000000
[NII]	6585.28	(6577, 6593)	6500 − 8000	0.08 − 0.15 (29)	2.96 − 2.39	1.0000	1.0000	0.000000
[SII]	6718.29	(6708, 6726)			1.30 − 1.10	0.3411	0.9998	0.000000

Table 3.6 (cont'd)

Line	λ_{lab} (Å)	$(\lambda_{low}, \lambda_{upp})^{\mathrm{a}}$ (Å)	$\lambda_{rest}{}^{\mathrm{b}}$(Å)	z (num. of obj.)$^{\mathrm{b}}$	$EW_{rest}{}^{\mathrm{c}}$(Å)	Reg. Coeff.$^{\mathrm{f}}$	Corr. Coeff.$^{\mathrm{f}}$	P-value$^{\mathrm{g}}$
[SII]	6732.67	(6726, 6742)			$0.62 - 0.94$	-0.5424	0.9800	0.000001
[ArIII]	7137.80	(7131, 7148)			$0.63 - 0.11$	0.8965	0.9424	0.000045

$^{\mathrm{a}}(\lambda_{low}, \lambda_{upp})$ is the rest-wavelength window within which the local continuum is estimated by linear-interpolation. These wavelength windows are determined by Vanden Berk et al. (2001).

$^{\mathrm{b}}$The restricted wavelengths, the corresponding redshifts and the numbers of the quasars chosen for the KL transforms. In a given redshift range, all appropriate quasars in our sample are included regardless of their absolute magnitudes, hence the correlations listed here are the ensemble-averaged properties.

$^{\mathrm{c}}$The range of restframe equivalent widths of the emission line along the spectral sequence (a_1, a_2) (in decreasing a_2 values).

$^{\mathrm{d}}$The observed wavelengths (λ_{obs}; in vacuum) as listed in Vanden Berk et al. (2001).

$^{\mathrm{e}}$Uncertain λ_{lab} and ($\lambda_{low}, \lambda_{upp}$) due to the unknown number of lines and their blended nature.

$^{\mathrm{f}}$The linear regression and the linear correlation coefficients are given in 4 significant figures.

$^{\mathrm{g}}$The two-tailed P-value for the correlation coefficient.

The EW$_{rest}$ of the emission lines vary at different magnitudes along the (a_1, a_2) sequence; some change by nearly a factor of two (e.g., Lyα, C IV), while some show smaller changes (e.g., Si IV+O IV], C IIIλ1906). Within a single local bin, the rest equivalent widths of some emission lines increase while others decrease along the trend (a_1, a_2) with decreasing a_2 values. These results are the testimonies to the fact that quasar emission lines are diverse in their properties.

We also note that some pairs of emission lines change their correlations as a function of redshift (i.e., different local bins). For example, Mg II is correlated with O III+Fe II(Opt82) in the local bin of $z = 1.1 - 1.87$ but anti-correlated in that of $z = 0.46 - 1.16$. Another example is the [S II]λ6718 and [S II]λ6733 pair. Hence if correlations are interpreted between the emission lines from one local bin with those from an adjacent bin, caution has to be exercised. The uncertainty in the continuum estimation (e.g., the iron contamination in the continuum in the vicinity of Mg II) prevents us from drawing an exact physical interpretation of this phenomenon.

3.6.1 Francis's PCs and Boroson & Green's "Eigenvector-1"

Two examples of the locally-constructed eigenspectra are shown in Fig. 3.26 and 3.27. In Fig. 3.26, the eigenspectra are constructed using wavelength-selected QSO spectra in the rest-wavelengths $1150 - 2000$ Å (with $2.3 < z < 3.6$), so that both Lyα and C IV are covered. Excellent agreement is shown between our eigenspectra and those selected from the Large Bright Quasar Survey in the $1.8 < z < 2.7$ range (Francis et al. 1992). The second eigenspectrum (corresponding to the first principal component in Francis et al.) shows the line-core components of emission lines. In contrast, the 3rd mode (corresponding to their 2nd principal component) shows the continuum slope, with the node located at around 1450 Å. Besides, the addition (with positive eigencoefficient) of the 3rd eigenspectrum to the 1st one enhances the fluxes at shorter wavelengths while *increases* the C IV blueshift. This supports the finding of a previous study (Richards et al. 2002b) that C IV blueshift is greater in bluer SDSS QSOs.

At longer wavelengths, the SDSS quasars with redshifts $0.08 - 0.67$ show the anti-
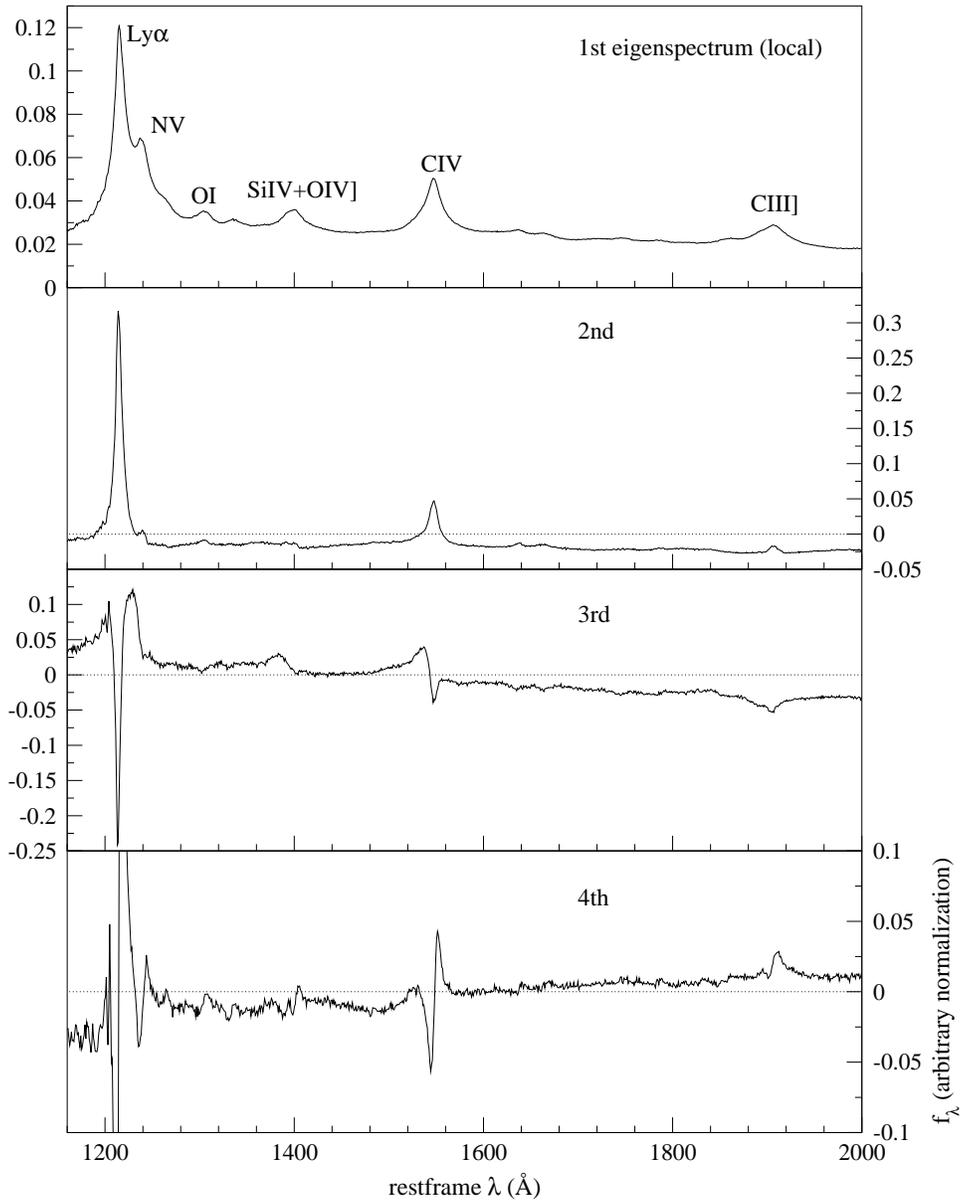
Figure 3.26 The first 4 orders of locally constructed eigenspectra in the restricted wavelength region $1150 - 2000$ Å, in which the quasar spectra are chosen to have full wavelength coverage. In this rather narrow rest-wavelength coverage, the emission lines in the 2nd eigenspectrum show the low-velocity core components (i.e., relatively narrower than the first eigenspectrum). The correlations among the relevant board emission lines are probed in the 2nd eigenspectrum.
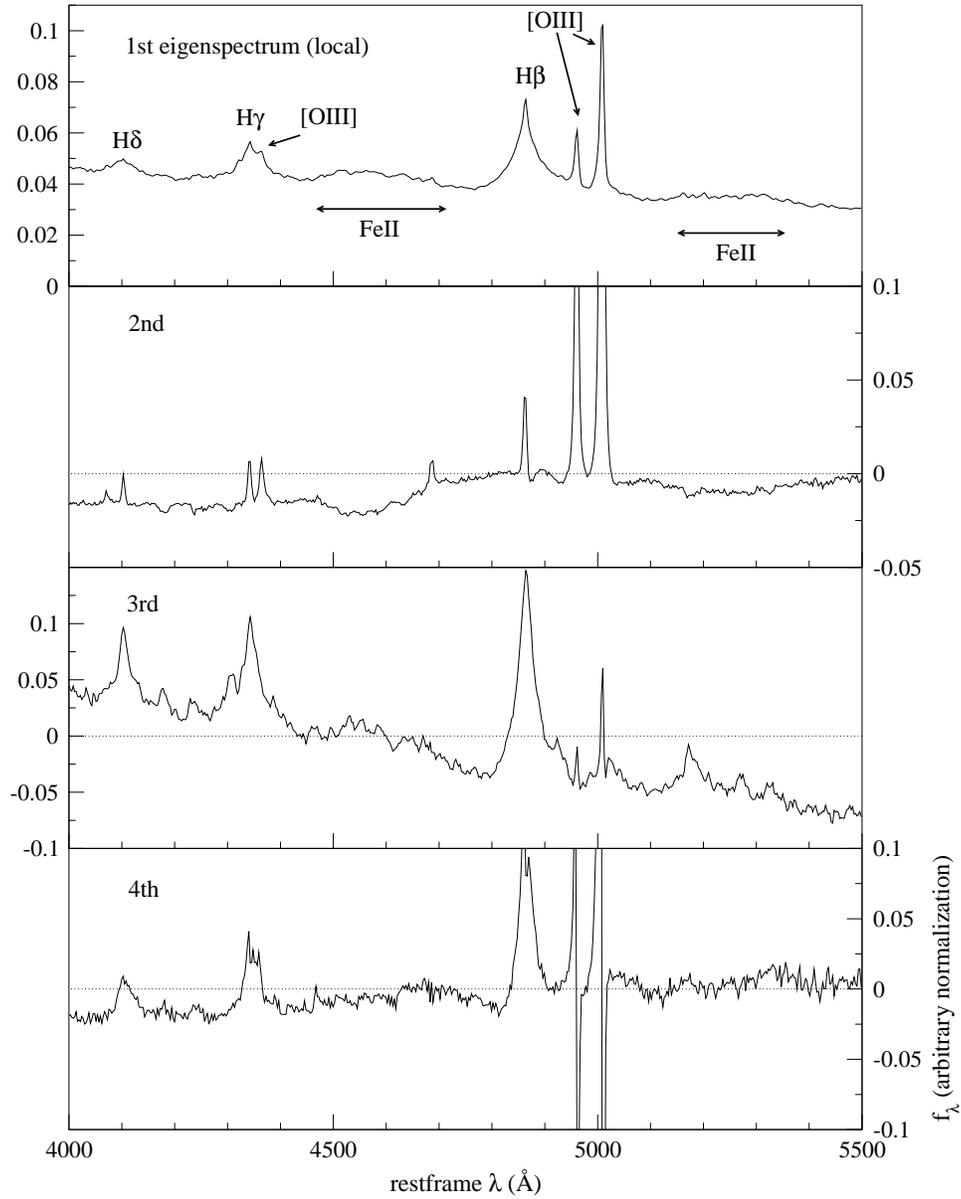
Figure 3.27 The first 4 orders of locally constructed eigenspectra in the restricted wavelength region $(4000, 5500)$ Å, in the vicinity of $H\beta$.

correlation between Fe II (optical) and [O III] (Fig. 3.27), in agreement with the Eigenvector-1 (Boroson & Green 1992). The first two eigenspectra in Fig. 3.27 demonstrate that both the Hβ and the nearby [O III] forbidden lines are anti-correlated with the Fe II (optical) emission lines, which are the blended lines blueward of Hβ and redward of [O III]. In the 3rd local eigenspectrum, the Balmer emission lines are prominent, which was noted previously in the PCA work by Shang et al. (2003). In addition, we find a correlation between the continuum and the Balmer lines in this local 3rd eigenspectrum, so that their strengths are stronger in bluer quasars.

To date, it is generally believed that the anti-correlation between Fe II (optical) and [O III] is not driven by the observed orientation of the quasar. One of the arguments by Boroson & Green was that the [O III]λ5008 luminosity is an isotropic property. Subsequent studies of radio-loud AGNs have put doubt on the isotropy of the [O III] emissions. Recent work by Kuraszkiewicz et al. (2000), however, showed a significant correlation between Eigenvector-1 and the evidently orientation-independent [O II] emission in a radio-quiet subset of the optically selected Palomar BQS sample, which implies that external orientation probably does not drive the Eigenvector-1. An interesting future project to address this problem is to relate the quasar eigenspectra in the SDSS to their radio properties.

3.6.2   Weight of line-core

Enlargements of the first two locally constructed eigenspectra focusing on major broad emission lines are illustrated in Fig. 3.28. Except for the almost perfectly symmetric and zero velocity of the line centers of the 1st and 2nd eigenspectra exhibited by [O III]λ5008, most broad emission lines do show asymmetric and/or blueshifted profiles. These demonstrate the variation of broad line profiles of quasars and the generally blueshifted broad emission lines relative to the forbidden narrow emission lines. The forbidden lines in the narrow line regions of a QSO are always adopted in calculating the systemic host-galaxy redshift, so the clouds associated with blueshifted BELs probably have additional velocities relative to the host. This line-shift behavior was found in many other studies (see references in Vanden Berk et al. 2001). The behavior of the C IV shift led Richards et al. (2002b) to suggest that

Table 3.7 The average FWHMs of major broad emission lines of the local QSO eigenspectra.

Percentage of Lines	1st−3rd	1st−2nd	1st	FWHM (km s^{-1}) (local) 1st+2nd[a]	1st+3rd	
Narrower than the 1st mode	72%	31%	$\cdots$	76%	52%	
Wider than the 1st mode	28%	69%	$\cdots$	24%	48%	
Example:						redshifts
Lyα+N V $(1160, 1290)$Å	2789	9139	3820	3104[b]	5549	$2.3 - 3.6$
C IV $(1494, 1620)$Å	3000	4594	3763	3339	4778	$2.3 - 3.6$
C III] $(1830, 1976)$Å	5904	6009	5802	5552	5721	$1.1 - 1.87$
Mg II $(2686, 2913)$Å	4194	4171	4149	4113	4117	$1.1 - 1.87$
Hβ $(4050, 4152)$Å	1978	2238	1997	1584	1999	$0.08 - 0.67$
[O III]λ5008 $(4982, 5035)$Å	551	571	535	521	522	$0.08 - 0.67$

[a]The 2nd eigenspectrum shows low-velocity (line-core) components of the broad emission lines.

[b]The Lyα and N V lines are not deblended in the 1st order (i.e., the mean spectrum), and are in the 2nd order, meaning that it is partly due to the deblending which causes the narrower width in the 2nd mode.

orientation (whether external or internal) may be the cause of the effect.

It is also obvious from Fig. 3.28 that the 2nd eigenspectra are generally narrower (except for Mg II, in which the conclusion is complicated by the presence of the surrounding Fe II lines) than their 1st eigenspectra counterparts. The line-widths of the sample-averaged KL-reconstructed spectra using only the first eigenspectrum or the first two eigenspectra are listed in Table 3.7. The addition of the first two modes, weighted by the medians of the eigencoefficients, causes the widths of 76% of the emission lines (with FWHM > 1000 km s^{-1}) to be narrower than those reconstructed from the first mode only. Hence, most broad emission lines can be mathematically decomposed into broad, high-velocity components and narrow, low-velocity components. Appearing in the second local eigenspectra, the line-widths are thus the most important variations of the quasar broad emission lines. The line-core components were reported by Francis et al. (1992) for C IV and Lyα; and Shang et al. (2003) for some major broad emission lines. One nice illustration of the line-core component of the 2nd mode is the splitting of Hγ and its adjacent [O III] in Fig. 3.27, for they are blended in the 1st mode.

Similar properties may be expected in the 2nd (M_i, z)-binned eigenspectra. Table 3.5

Table 3.8 The average FWHMs of major broad emission lines of the global QSO eigenspectra.

| Percentage of Lines | FWHM (km s^{-1}) (global) | | | | |
	1st$-$3rd	1st$-$2nd	1st	1st$+$2nd	1st$+$3rd
Narrower than 1st mode	26%	50%	$\cdots$	48%	73%
Wider than 1st mode	74%	50%	$\cdots$	52%	27%
Example:					
Lyα+N V $(1160, 1290)$Å	6837	5045	5103	9105[a]	3678
C IV $(1494, 1620)$Å	4787	4536	4518	4485	4341
C III] $(1830, 1976)$Å	6405	5867	6004	6195	5626
Mg II $(2686, 2913)$Å	4035	3997	3980	3959	3918
Hβ $(4050, 4152)$Å	2605	3090	2506	2244	2351
[O III]λ5008 $(4982, 5035)$Å	! 566	525	546	551	538

[a]The Lyα and N V are deblended in the 2nd eigenspectrum and are blended in the 1st one.

lists the average FWHM of different linear combinations using the first 3 eigenspectra in constructing some major broad emission lines. Comparatively, for most emission lines the second (M_i, z)-binned eigenspectra do not show as narrow line components as the second eigenspectra, in which the widths of 61% of the emission lines with FWHM > 1000 km s^{-1} become narrower by adding the 2nd eigenspectrum to the 1st one. This effect is mainly due to the difference in the numbers of quasars, and more importantly, the inclusion of a wider spectral region causes the ordering of the weights of different physical properties to re-arrange. In this case, the spectral slope variations are more important than those of the line-cores. While the 3rd (M_i, z)-binned eigenspectra (weighted by medians of the eigencoefficients of the sample) also do not represent prominent changes in the emission line-cores, except for Lyα and C IV (the FWHM of C IV appears to be larger because the line-core 3rd mode is pointing downward in **ZBIN 4**), on average the quasar populations with *negative* 3rd eigencoefficients do show narrower widths for 77% of the emission lines. Similarly, the 2nd global eigenspectrum does not carry dominant emission line-core components, which are found to be represented more prominently by the 3rd mode (Table 3.8).
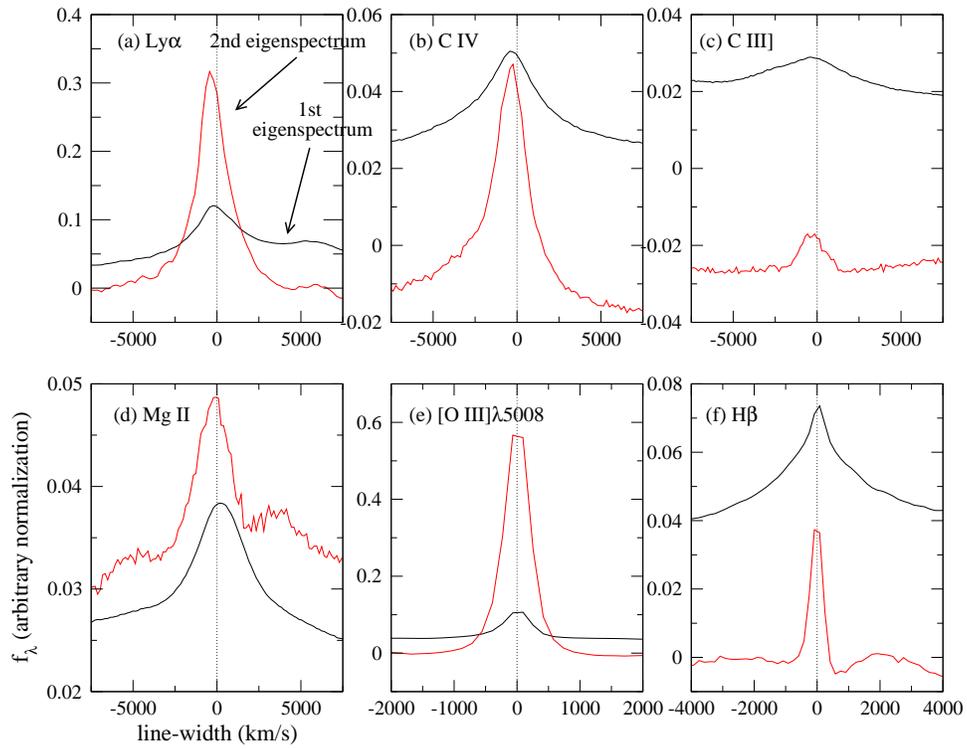
Figure 3.28 The first (black) and the second (red) locally constructed eigenspectra in the regions around different emission lines.

3.6.3   FWHM-EW anti-correlation in broad emission line: classification?

The narrower emission features in the 2nd local eigenspectrum compared with the 1st one, and the fact that almost every broad emission line is pointing towards positive flux values in both of these two modes, imply that there is an anti-correlation between FWHMs and the equivalent widths of broad emission lines. In fact, as suggested by Francis et al. (1992), this may form a basis for the classification of quasar spectra in $\lambda_{rest} = 1150 - 2000$ Å, by arranging them accordingly into a sequence varying from narrow, large-equivalent-width to broad, low-equivalent-width emission lines. From the locally constructed eigenspectra, such an anti-correlation is not generally true for every broad emission line as we find that there exists at least one exception: a positive correlation between the FWHM and the EW of Mg II in the local bin of the redshift range $0.46-1.16$. An assumption in these measurements is that the continuum underneath can be approximated by a linear-interpolation across the window $2686 - 2913$ Å. One complication, however, is the contamination due to the many Fe II emission lines in the vicinity of Mg II, so the true continuum may be obscured. The positive FWHM-EW correlations appear to exist in some other weaker emission lines as well, but the weak strengths of those lines do not permit us to draw definitive conclusions under the current spectral resolution. In conclusion, the FWHM-EW relation can help us to classify most broad emission lines individually, but *this relation cannot be used in a general sense, nor does it represent the most important sample variation*, if the surrounding continua are included to the extent of the rest-wavelength ranges of the (M_i, z)-binned spectra. Nonetheless, most broad emission lines can be viewed mathematically as the combinations of broad and narrower components. A future study will focus on finding the best physical parameters for classifying the spectra in the wide spectral region. One possible approach is to study the distributions of the eigencoefficients and their relations with other spectral properties (e.g., Francis et al. 1992; Boroson and Green 1992).

3.6.4   Local spectral properties in the (M_i, z)-binned eigenspectra

The shapes of the continua and the correlations among the broad emission lines of the second locally constructed eigenspectra are all identified in either the 3rd or the 4th (M_i, z)-binned

eigenspectra. We do expect, and it is indeed found to be true, that the local properties of the spectra can be found in the latter, though the ordering may be different. The identifications are marked in Fig. $3.7 - 3.11$ by the redshift ranges of the local eigenspectra, with reference to the luminosity averaged **ZBIN** eigenspectra. The correlations of broad emission lines are generally found in higher-order (M_i, z)-binned eigenspectra compared with the orders representing the spectral slopes.

3.7   GAP CORRECTION IN QSO SPECTRA

The construction of the (M_i, z)-bins in this work (§3.3) is performed by constraining the gap fraction to be smaller than 50% for each spectrum to improve the accuracy of spectral reconstructions using eigenspectra. Here we discuss in detail how this value is arrived at. We artificially mask out (i.e. assign a zero weight) to given spectral intervals and study how well we can reconstruct these "gappy" regions from the eigenspectra (Connolly and Szalay 1999). The comparison of the KL-reconstructed spectrum with the original unmasked spectrum gives a direct assessment to the accuracy of the gap-correction procedure. We perform this test for the (M_i, z)-binned quasar spectra from this work. To simulate the effects of un-observed spectral regions due to different rest-wavelength coverage for quasars at different redshifts (the principal reason for gaps in the quasar spectra in our sample), each spectrum in all (M_i, z)-bins is artificially masked at the short- and the long-wavelength ends. The masked spectra are then projected onto the appropriate eigenspectra and the reconstructed spectra are calculated using the first 50 modes. The fractional change in the flux density per wavelength bin (weighted by w_λ), $\sqrt{\sum_\lambda w_\lambda (f_\lambda - f_\lambda^{Recon})^2} / \sum_\lambda w_\lambda f_\lambda$, between the observed spectrum f_λ and the reconstructed spectrum f_λ^{Recon}, averaged over all quasar spectra in each bin, are shown in Fig. 3.29a as a function of the spectral gap fraction. The gap fraction is calculated relative to the full restframe wavelength range, a variable for each quasar spectrum. The reconstruction from 50 modes has an intrinsic error of approximately 6.2% (due to the noise present in each spectrum, and the existence of 3.4% bad pixels on average for each spectrum), which is estimated by reconstructing the spectra with no artificial gaps.

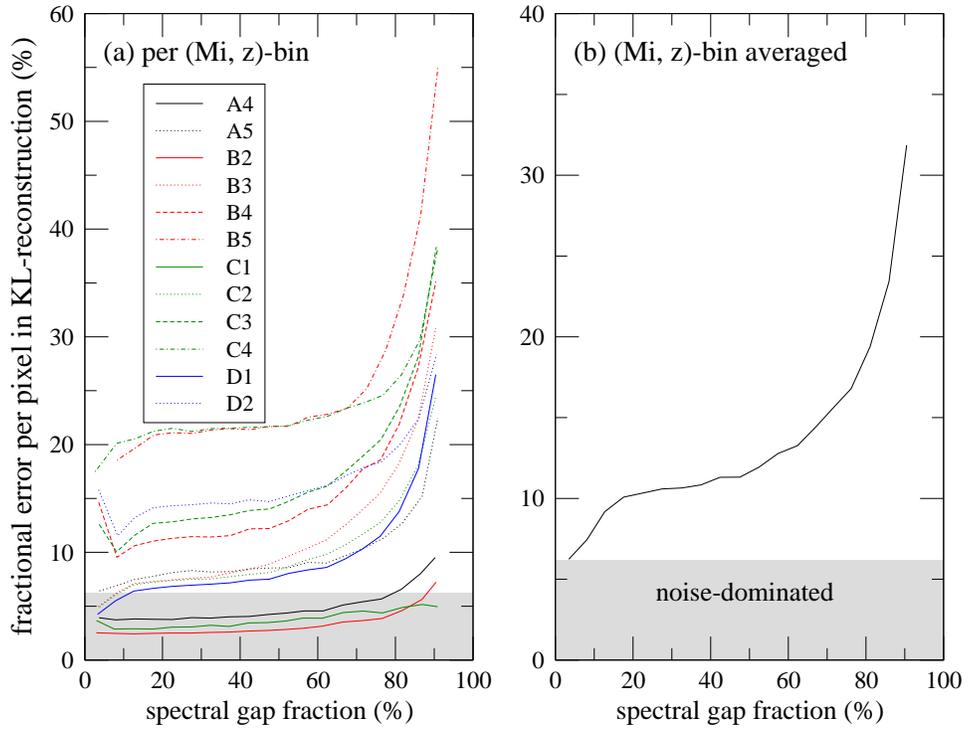Figure 3.29 The fractional error of KL-reconstruction in flux density per wavelength bin as a function of the spectral gap fraction in the observed quasar spectra, where in (a) each curve is averaged over all quasars in the corresponding (M_i, z)-bin, and in (b) the curve is averaged over all quasars in all (M_i, z)-bins. The gray region in each is the excluded region due to the average intrinsic noise per wavelength bin.

As expected, the difference between the unmasked observed spectrum and the reconstructed spectrum increases gradually with gap fraction.

Averaging over all (M_i, z)-bins (Fig. 3.29b), at a spectral gap fraction of 52.5% the mean error in the 50-mode reconstruction is $\sim 11.9\%$, which is 5.7% above the noise-dominated average reconstruction error in the flux. While a smaller gap fraction is in principle more desirable, 50% is chosen to be the upper bound to compromise the fewer (M_i, z)-bins.

In the construction of the global eigenspectra set covering the rest-wavelength range $900 - 8000$ Å, there are 89% of the QSOs (Table 3.9) having spectral gap fractions larger than 50%. From Fig. 3.29b, we find that a gap fraction larger than 76% gives substantial reconstruction errors ($> 16.8\%$), implying $\sim 17\%$ of the QSOs used in defining the global eigenspectra may be poorly constrained when correcting for the missing data. We stress that in defining the global eigenspectra from the SDSS this is strictly the best estimation that can be made at present, as no SDSS spectroscopic observations are available in the gap regions at the red and the blue ends of the spectrum. The impact of this gap correction is, as expected, wavelength dependent. Wavelengths shortward of 5000 Å are very well constrained even with the global eigenspectra with less than 1% of QSOs having gap corrections in excess of 76% (Table 3.9). Determining the impact of the gaps and the use of additional spectroscopic observations to complement the SDSS data should be addressed in a future.

We also find that quasar broad emission lines can be reconstructed locally using the (M_i, z)-binned eigenspectra with errors that are typically small relative to the noise level. For example, if C III] is masked (over the region of influence $1830-1976$ Å), averaging over all QSOs in the bins **B3** and **C3**, the 50-mode reconstruction error described above is 10.4%; and for Mg II (over the region of influence $2686 - 2913$ Å), 11.3%. For the case in which at least one broad emission line is masked and with a substantial total gap fraction (in our case, C III]; and a mean spectral gap fraction of 60.0%), the average reconstruction error per pixel is found to be 12.5% when averaging over the bins **B3** and **C3**. Fig. 3.30 shows the observed and the reconstructed spectra of an object with a reconstruction error approximately equal to the average value. While the reconstructed continuum has a small difference from the observed continuum, the emission line C III] is reconstructed well, extremely well if considering the fact that the whole region of influence is within the masked region. The actual quality of
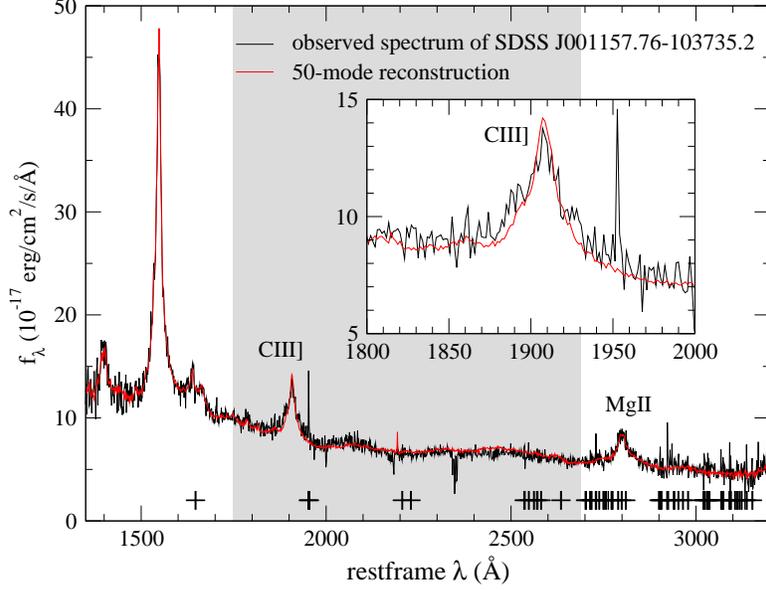
Figure 3.30 An example of a 50-mode reconstruction (red) using the (M_i, z)-binned eigenspectra. The spectrum has a total gap fraction of 56.5%. The gray area is the artificially masked spectral region, and the crosses mark the bad pixels in the original spectrum. The fractional reconstruction error per pixel is 12.1%. The inset shows the emission line locally.

Table 3.9.   Spectral gap fraction of the QSO sample.

Spectral gap fraction larger than	Number (fraction) of QSOs		
λ_{rest}-range of the eigenspectra:	$900 - 8000$ Å	$900 - 7000$ Å	$900 - 5000$ Å
0.4	16,420 (0.98)	15,313 (0.92)	10,423 (0.62)
0.5	15,050 (0.90)	13,561 (0.81)	6,421 (0.38)
0.6	12,696 (0.76)	10,275 (0.62)	1,682 (0.10)
0.7	7,424 (0.44)	3,519 (0.21)	423 (0.025)
0.75	2,920 (0.17)	1,131 (0.068)	100 (0.0060)
0.8	873 (0.052)	416 (0.025)	0 (0.00)
0.9	0 (0.00)	0 (0.00)	0 (0.00)

the reconstruction depends on the individual spectrum and position and size of the gaps.

4.0   SPECTRAL VARIABILITY OF GALAXIES

Nothing endures except change.

<div align="right">*Heraclitus*</div>

Active galactic nuclei (AGNs) are found to vary in many frequencies. The underlying physical mechanism driving the variability seen in AGN spectra is unknown (Peterson 2001; Bono et al. 2003). Several models have been proposed including accretion disk instability (Kawaguchi et al. 1998), starburst model (Aretxaga et al. 1997) and gravitational lensing (Hawkins 1993). At present, models involving events or instabilities happening on AGN accretion disk are most favored.

Most variability studies, however, have been focused on Type 1 (broad line) AGN. Emission lines of Type 2 (narrow line) AGNs, while were found to vary in the X-rays on time scales from hours to months (Guainazzi et al. 1998; Mueller et al. 2003), their optical continua and emission lines are generally considered to be non-variable, nor was there any strong observational evidence to either support or dispute the idea. If the continuum is generated by an accretion disk as in broad line AGNs, in which variability is seen, narrow line AGNs can in principle exhibit variability in their continua.

From a 4 year monitoring of the broad-band optical variability of 35 Type 1 and 2 Seyfert galaxies, Winkler et al. (1992) demonstrated that most galaxies in the sample are variable and that the nuclear flux in the optical should be of thermal origin. Another promising work to probe Type 2 AGN variability in the optical is the intranight variability of Seyfert 2 galaxies observed by Jang (2001). Unfortunately, the sample size is small (2 out of 3 objects were found to vary in ~ 0.25 mag).

In contrast, the variability of broad line AGNs is better established with an amplitude of

$\sim 10\%$. Wavelength dependence of QSO variability in the Sloan Digital Sky Survey (SDSS, York et al. 2000) have been studied by Wilhite et al. (2005b), and the C IV dependence by Wilhite et al. (2005a). In these studies the authors have developed a method to remove systematics which may exist in the spectra from a spectroscopic plate at one epoch of observation when compared to the same plate at another epoch. Correcting for this they were able to select variable QSOs and demonstrated a power-law liked QSO difference spectrum. QSO photometric variability in the SDSS has been studied by Vanden Berk et al. (2004a) for its dependency on physical parameters such as redshift, luminosity and radio properties, and by Ivezić et al. (2004a) in which the authors use the SDSS repeated imagings in combination with the POSS data to probe long term variability (up to ~ 50 years).

The goal of this chapter is to probe spectral variability of galaxies and narrow line AGN using the SDSS multi-epoch observations. The SDSS data have the advantage of large sample sizes with the same spectrophotometric reduction, which is critical in our work because we expect any variability present in galaxies to be small (compared with QSO variability, for example). Spectroscopic variability studies are complemented by analyses of the repeated imagings of the SDSS Southern Stripes which provide more epochs over a longer time baseline. Techniques similar to Wilhite et al. (2005b) are used to refine the spectroscopic calibration.

4.1   DATA

The preliminary sample is chosen from the SDSS Main Galaxy sample (Strauss et al. 2002) (spectro1d reduction 23), by selecting spectra which are both spectroscopically targeted and classified as galaxies, and rejecting spectra in which the redshift measurements failed or have not been made. From this sample pairs of spectral observations by matching sources in both RA and DEC to within $1''$, and in redshift to within 0.001. This results in $\sim 49,000$ mutually exclusive pairs. In order to increase the quality of the re-calibration we drop any plate from our analysis in which there are less than 20 pairs of galaxy spectra (referred as the initial sample). The initial sample comprises galaxies of different eClass types (Connolly et al.
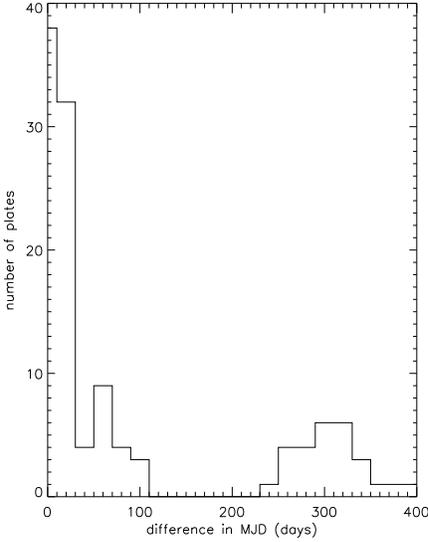
Figure 4.1 Number of plates vs. difference between the MJDs of the two epochs in the final sample (bin size is 20 days).

1995) and narrow-line AGNs. The subsequent sample selections (mainly for the purpose of increasing the signal-to-noise of the spectra) will be described in §4.2 and §4.3. The distribution of time baseline in the final sample of the SDSS spectroscopy is shown in Fig. 4.1. The sample covers $\sim 0-400$ days in the observed frame. The spectra are de-reddened against Galactic extinction using the library written by Simon Krughoff (the library is available from the author upon requests), which adopts the SFD dust maps (Schlegel et al. 1998) and the extinction curve by O'Donnell (1994).

4.2 REFINING THE SPECTROSCOPIC CALIBRATION

4.2.1 Removing systematics

We adopt the method developed by Wilhite et al. (2005b) to remove systematics which may present in each pair of galaxy spectra between two MJDs of a given plate. Calibration correlations as a function of observed wavelength are calculated by a origin-passing linear
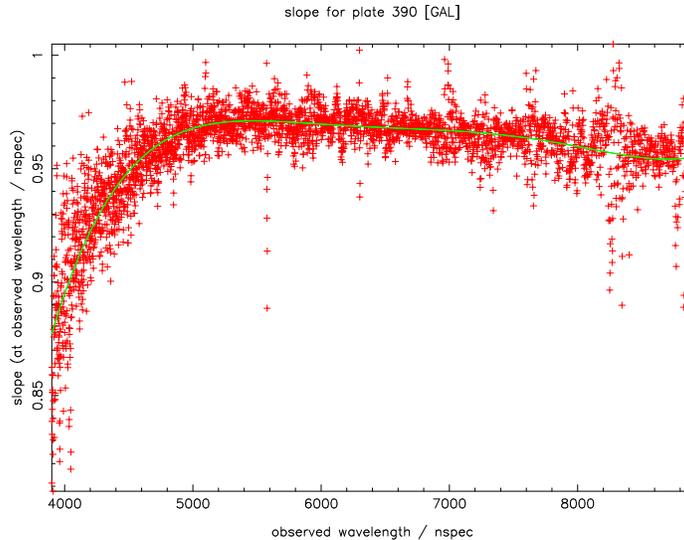
Figure 4.2 Correction slope as a function of observed wavelength and its 5th-order polynomial fit, of plate 390.

fit to the flux density (f_λ) of all the objects at the two epochs. This results in a calibration spectrum of one plate. We adopt a slightly different methodology from Wilhite et al. (2005b) in which stellar spectra in each plate were used. Our calibration spectrum per plate is calculated using all galaxies present in that plate, as we want to maximize the number of objects in generating the calibration spectrum, and avoid potential systematic effects when applying point sources calibration to extended sources. An example calibration spectrum is plotted in Fig. 4.2. In general the calibration spectra calculated using the SDSS plates are smooth. To focus on lower-order correction to the galaxy spectra, and to remove noise in the calibration spectrum, we follow Wilhite et al. (2005b) and set the final calibration spectrum per plate as its 5th-order polynomial least-squares fit.

4.2.2 Restframe band scaling vs. absolute flux calibration

After removing systematics, each spectral pair can in principle be subjected to a wavelength-independent normalization in their flux densities. In order to obtain an accurate variable measurement of each spectral pair, the total observed continuum fluxes of the two spectra at a chosen common wavelength region are set equal. Specifically, for each pair the galaxy

spectrum at the epoch of lower S/N is shifted in a wavelength-independent fashion according to

$$f_\lambda^{(\text{low S/N})} = C_{\text{band}} \, f_\lambda^{(\text{high S/N})} \, , \tag{4.1}$$

where C_{band} is a constant for each spectral pair, calculated by

$$C_{\text{band}} = F_{\lambda \in band}^{(\text{low S/N})} / F_{\lambda \in band}^{(\text{high S/N})} \, , \tag{4.2}$$

where $F_{\lambda \in band}$ is the total observed flux in the wavelength range (which is referred to a "rest-band" for convenience). In this work the rest-band is chosen to be 6450 ± 260 km s^{-1}, as such no prominent emission lines are present in the galaxy spectra and the spectral region represents mainly continuum. Only the galaxy spectra in the bad epoch are re-calibrated, where a bad epoch is determined by the one with lower S/N in the spectrograph over three different magnitude limits of 19.90, 20.20, 20.25. The combined procedure of systematics removal and restframe band scaling is referred to "SYS+BAND" in the following.

Base on the initial sample (§4.1), we impose S/N criteria in the concerned rest-band (or the [O III]λ5008 line, later described in the absolute flux calibration), that results in the parent sample. We reject any spectral pair from the analysis if there is one or more pixels flagged in either or both spectra in the wavelength region of interest, and to require flux larger than 2σ from the sky. The flags under considerations are BADTRACE, BADFLAT, BADARC, NEARBADPIX, LOWFLAT, FULLREJECT, SCATLIGHT, BRIGHTSKY, NODATA, and COMBINEREJ. The parent sample contains $33,270$ galaxies for the band-scaling method.
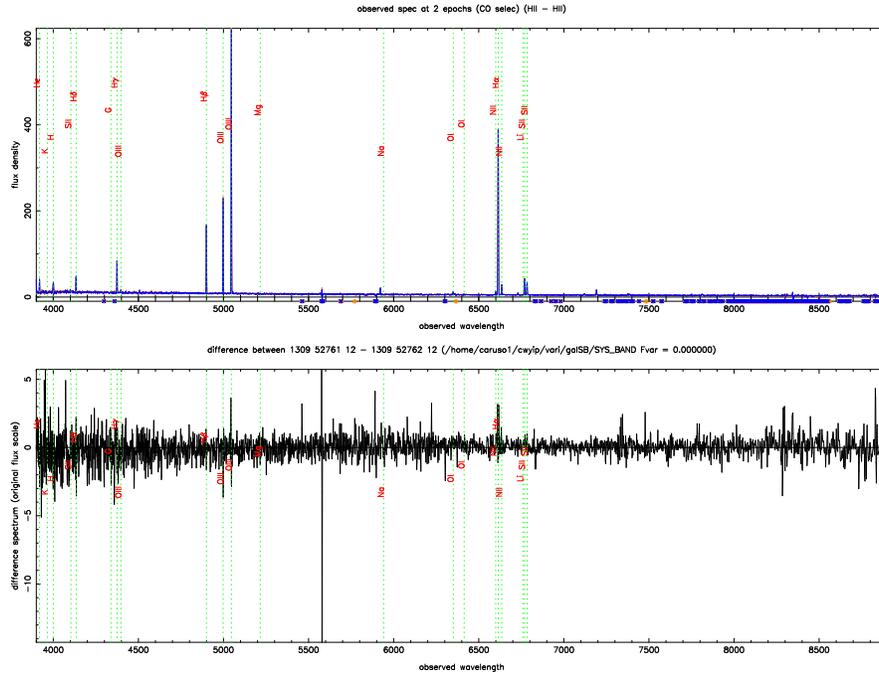
Figure 4.3 An example of HII galaxies with zero variability after the SYS+BAND recalibration. Top: spectra at two different epochs. Bottom: difference spectrum. The spectra are in observed frame.
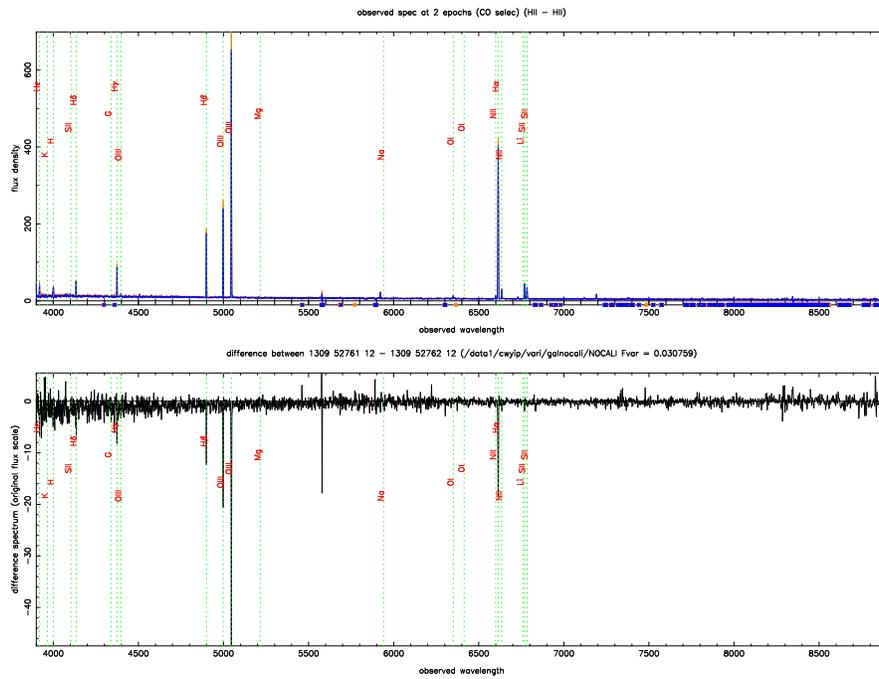

Figure 4.4 The same HII galaxy from the pipeline (the scale of the y-axis is different). Emission line residuals are evident.

In Fig. 4.3 and 4.4 we show the observed galaxy spectra in two epochs after the calibration refinement and those from the SDSS pipeline respectively. Firstly, we note that the pipeline performs very well (the average accuracy of the spectrophotometry is 3%), which is sufficient for most spectral analyses. However, emission line residuals in Fig. 4.4 are evident, which would affect the variability analysis. Similarly, in Fig. 4.5 and 4.6 we show the spectra at two epochs for a narrow line AGN, before and after the re-calibration, and the difference spectrum from the pipeline shows residuals in the continuum. In both cases, after the calibration refinement we can be able to obtain zero variabilities for the objects.

For comparison with the rest-band scaling previously described, we also adopt the absolute flux calibration described in Peterson et al. (1998). In this procedure, the spectrum in the bad epoch (as defined previously using S/N criteria) in the duo is scaled wavelength-independently to have the same [O III]λ5008 total observed flux as the spectrum in the good epoch. The physical justification is that the [O III]λ5008 line generally appears as a narrow line which is believed to be non-variable in time (Peterson et al. 1995). The combined procedure of systematics removal and absolute [O III] calibration is called "SYS+OIII" in the following.

When judging which method performs better in analysing spectral variability, we consider the better method to be the one which minimizes the median, mean and scatter (quantified by one standard deviation, σ) of the variability at a fixed spectral S/N. In Fig. 4.7 and 4.8 we compare the above mentioned statistics between SYS+BAND and SYS+OIII in the parent sample. In each figure, the left panel shows F_{var} vs. S/N and the right panel shows the histogram of F_{var} at S/N = 10 $\pm$1 with the peak normalized to unity and a bin size of 0.005. At S/N = 10 the method SYS+BAND is smaller than SYS+OIII by $\sim 25\%$ in the scatter, $\sim 39\%$ in the mean, and $\sim 41\%$ in the median of variability. The result suggests that [O III]λ5008 in some spectra may be subjected to the effect of seeing so that the total observed line flux is different for the two epochs. Another possible reason for this discrepancy is that some spectra may exhibit variability in the emission line.

To further test the method SYS+BAND, we divide the parent sample into red and blue galaxies by using the eClass classification (described in §4.5.1.1). Red spectra are defined as objects with classification parameter $\phi_{KL} > 0^o$, and blue galaxies, $\phi_{KL} < 0^o$. For each plate,
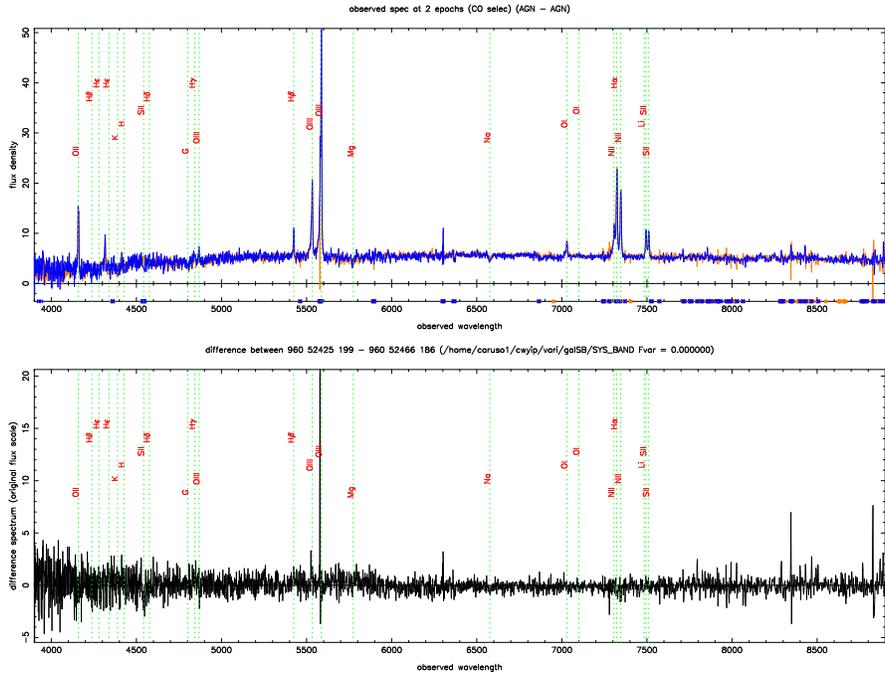
Figure 4.5 An example of narrow line AGNs with zero variability after the SYS+BAND re-calibration. Dots on the x-axis show the bad pixels.
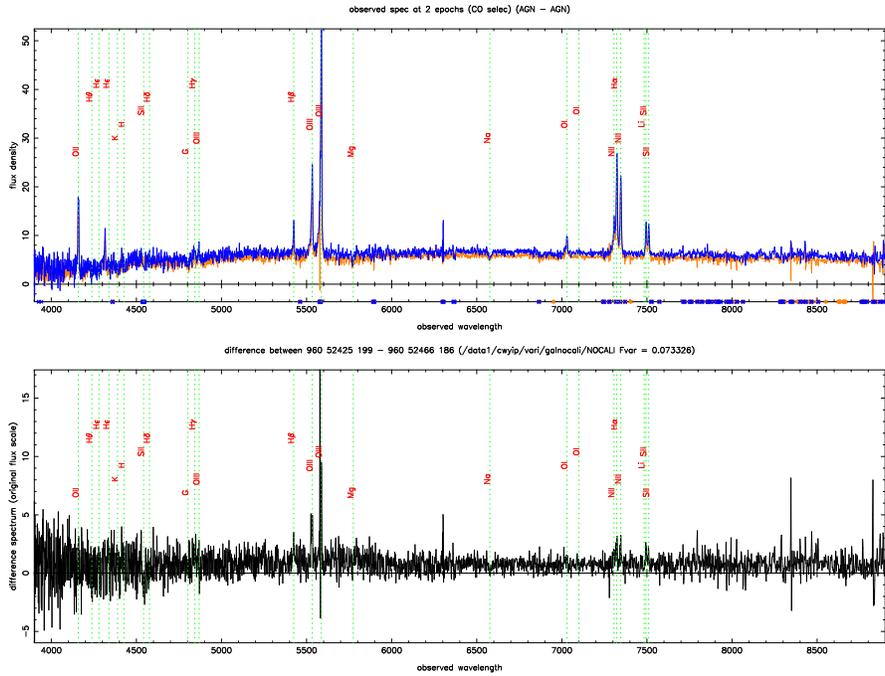
Figure 4.6 The same AGN from the pipeline (the scale of the y-axis is different). There are some residuals in the continuum.
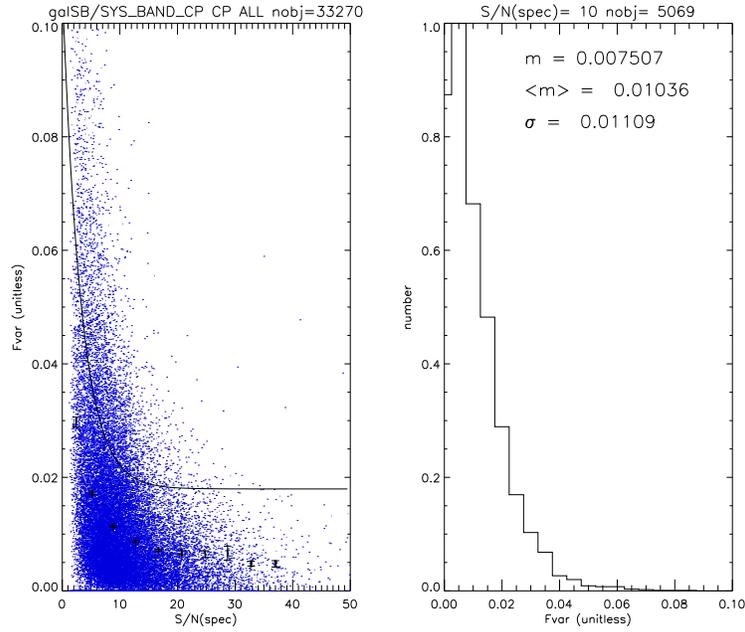
Figure 4.7 $F_{\mathrm{v}ar}$ vs. S/N of galaxy spectra in the parent sample. Re-calibration method: SYS+BAND, using all galaxy spectral types in calculating the calibration spectrum. Black marks are $(\langle F_{\mathrm{v}ar}\rangle, \langle S/N\rangle)$ with the error bars showing the 1σ of the mean.
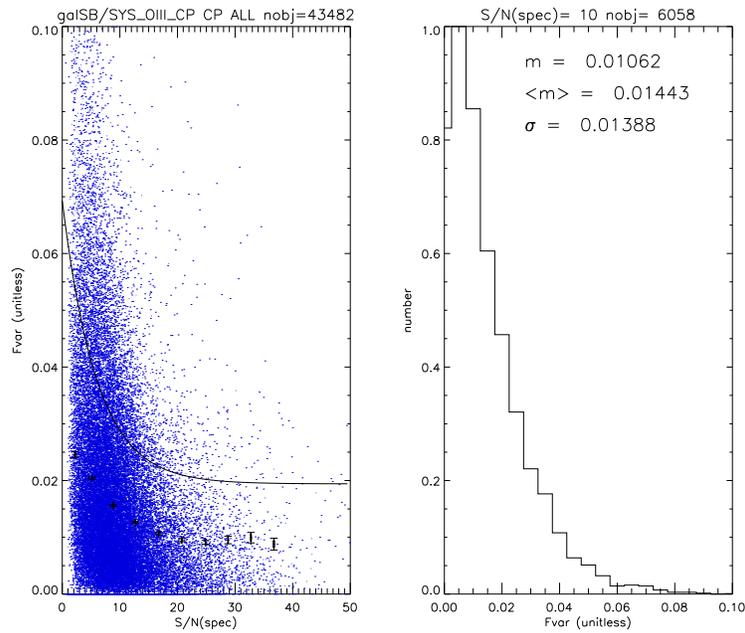
Figure 4.8 Same as Fig. 4.7 but the re-calibration method is the SYS+OIII. Scatter of $F_{\mathrm{v}ar}$ at S/N = 10 is slightly larger than the case of SYS+BAND.
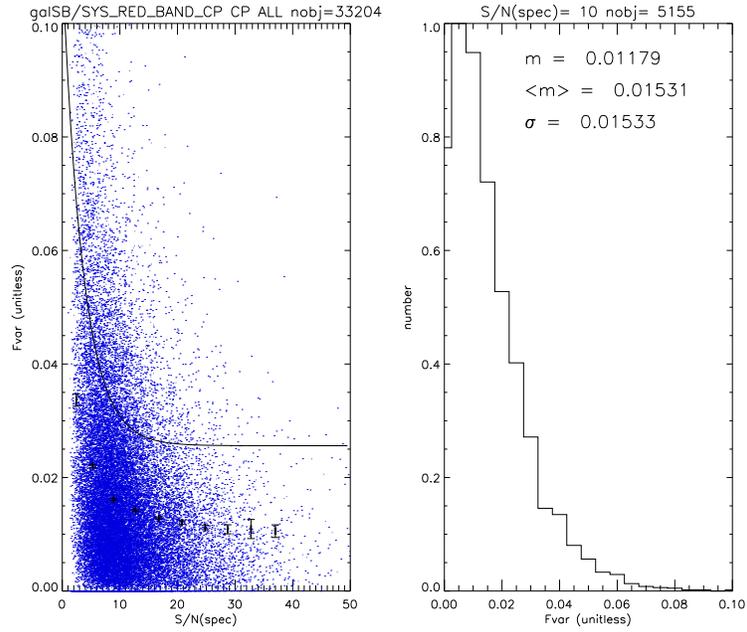
Figure 4.9 Same as Fig. 4.7 but using only red galaxies in calculating the calibration spectrum. Scatter of F_{var} at $S/N = 10$ is slightly larger than the previous cases.
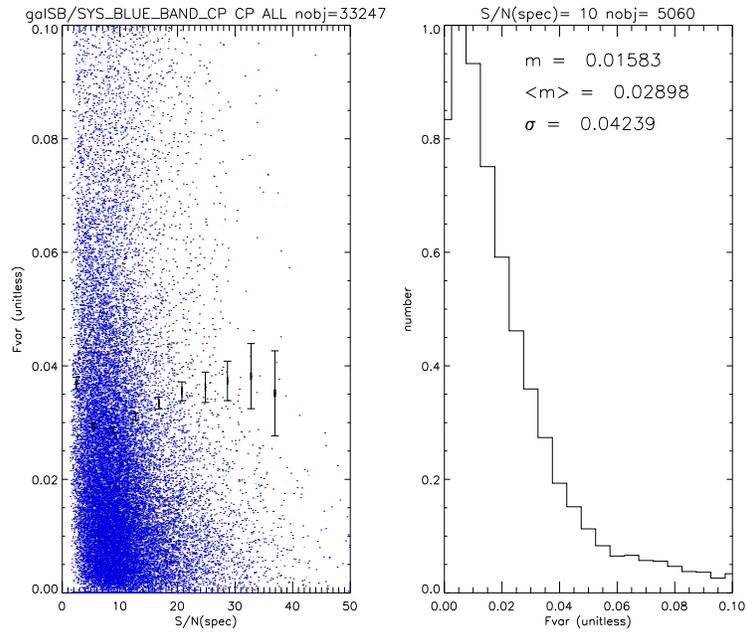
Figure 4.10 Same as Fig. 4.7 but using only blue galaxies in calculating the calibration spectrum. Scatter of F_{var} at $S/N = 10$ is substantially larger than the previous cases.

the calibration spectrum is calculated using only red (Fig. 4.9) and blue galaxies (Fig. 4.10), which is in turn used to re-calibrate galaxies of all spectral types on the same plate. The purpose behind is to determine whether the re-calibration method SYS+BAND is better, in the sense mentioned above, when the calibration spectrum is constructed from a specific spectral type. We find that using a specific spectral type does not cause the variability statistics to decrease. Using only red and only blue galaxies the mean of variability increases compared with using all spectral types by $\sim 48\%$ and $\sim 180\%$ respectively. Calibration using blue galaxies even causes a divergent behavior in F_{var} vs. S/N (Fig. 4.10). We therefore fix the re-calibration method to SYS+BAND and use all galaxy spectral types in calculating the calibration spectrum.

4.3   VARIABILITY AMPLITUDE: F_{VAR}

We adopt the dimensionless variability measure, F_{var}, for N-epoch repeated observations (Rodriguez-Pascual et al. 1997; Peterson 2001). It is a fractional root mean square variability amplitude defined as

$$F_{var} = \frac{\sqrt{\sigma^2 - \delta^2}}{\langle f \rangle}, \tag{4.3}$$

where σ^2 is the variance of the flux, δ^2 is the mean square uncertainty of the flux, and $\langle f \rangle$ is the mean flux

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (f_i - \langle f \rangle)^2, \tag{4.4}$$

$$\delta^2 = \frac{1}{N} \sum_{i=1}^{N} \delta_i^2, \tag{4.5}$$

$$\langle f \rangle = \frac{1}{N} \sum_{i=1}^{N} f_i. \tag{4.6}$$

The numerator is termed the "excess variance" (Vaughan et al. 2003), in the sense that the variability is subtracted from the flux uncertainty. In the calculation of the continuum F_{var} the spectral regions corresponding to the positions emission lines (using the SDSS line list $\pm$ 280 km s^{-1}) are excluded. The mean square uncertainty for each spectrum is calculated

by quadrature summation of the pipeline flux uncertainty per pixel, over all valid pixels. A valid pixel is defined to be a pixel that is not flagged as bad. The flags are listed in §4.2.2.

If the excess variance $\sigma^2 - \delta^2 < 0$, we assume the variability is zero for the object, which will still be included in the analysis. In this work, $N = 2$.

4.3.1 Defining variables

Following Wilhite et al. (2005b) we define variable candidates as objects having

$$F_{\mathrm{var}}(S/N) > b_0 \exp\left(\frac{S/N}{b_1}\right) + b_2\,. \tag{4.7}$$

The signal-to-noise of each spectrum, S/N, is calculated in the bright phase using all pixels. The constants b_0, b_1, b_2 are determined by least-squares fitting the above exponential form to the binned data $(\langle S/N \rangle, \langle F_{\mathrm{var}} \rangle + \sigma(F_{\mathrm{var}}))$, where $\sigma(F_{\mathrm{var}})$ is one standard deviation of the variability in bin.

For comparison we have also defined variables as objects located in a fixed upper percentile of F_{var} in the plot F_{var} vs. S/N. We find that the above mentioned 1σ cut is approximately similar to a 10% upper percentile for a given S/N.

4.3.2 Plate-to-plate difference

There are plate-to-plate differences seen in the above-mentioned definitions of variable and non-variable sources, that is, the fitting coefficients in Eqn. 4.7 are in general different for different plate. This fact is illustrated in Fig. 11(a), in which some of the objects defined on a plate-to-plate basis (red crosses) are crossing below the line defined using the plate-averaged coefficients b_0, b_1 and b_2. However, without a further treatment on refining the calibration from a given plate to another different one, we define our galaxy variable and non-variable samples on the plate-to-plate basis for this work. We can tell from Fig. 11(b) that the non-variable sample could be contaminated by potentially variable ones, which of course are not uniquely defined in the first place.
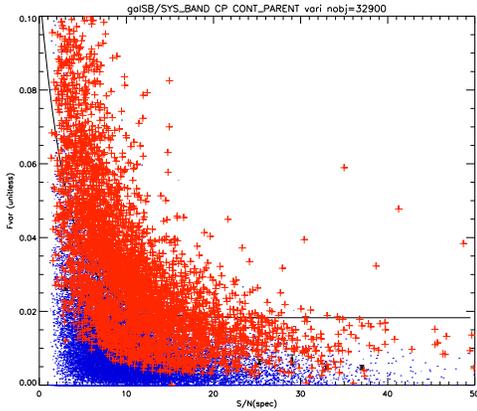
(a) Blue dots: F_{var} vs. S/N for all galaxies in the final sample, red crosses: for variable candidates defined on plate-to-plate basis. Black line is the mean + 1σ averaged over all plates (for eye-guiding).
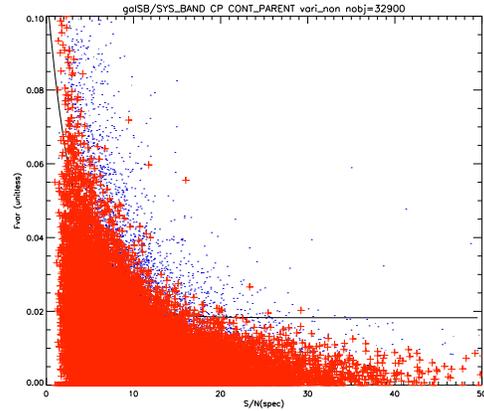
(b) Blue dots: F_{var} vs. S/N for all objects, red crosses: for non-variable candidates defined on plate-to-plate basis; of the continuum parent sample. Black line is the mean + 1σ averaged over all plates.

Figure 4.11 F_{var} vs. S/N, comparison between plate-to-plate basis and plate-averaged in defining variables.

Based on the parent sample we reject any plate in which the envelope fitting fails (either because of the small number of objects or low S/N of the objects). This results in 32,900 galaxies spanning 117 distinct plates. These objects form the final sample of our analysis.

4.4   VARIABLE STARS

In order to evaluate the performance of the re-calibration method, we re-calibrate stellar spectra in a plate using calibration spectrum from galaxies of the same plate. We define stellar variable and non-variable candidates using the previous galaxy envelope fitting parameters b_0, b_1, b_2 for each plate. Both candidates of stellar variable and non-variable are matched with catalogs of variable stars: the First Catalog of the QUEST RR Lyrae Survey (Vivas et al. 2004) and the General Catalogue of Variable Stars (GCVS, Kholopov et al. 1999, Vol. I-III) as an assessment to the definition of variables. The RA and DEC are both

matched to within $1''$ respectively.

In Fig. 12(a), red crosses mark the known RR-Lyrae according to the QUEST RR Lyrae first catalog, which are also photometrically classified as blue horizontal branch (BHB) stars in the SDSS and are variables by our definition. A subset of the variable stellar candidates defined in this work are real variable stars (64 out of 153 objects, $\sim$ 42%). Conversely, in Fig. 12(b), red crosses mark the non-variable stars in our definition which are known RR Lyrae. There are a substantial number of non-variable candidates are that indeed real variable stars (173 out of 722 objects, $\sim$ 24%). The percentage of known variable stars is, favorably, larger in the variable-candidate sample. When the larger catalog GCVS is used in the matching (Fig. 13(a) and 13(b)), there are 107 out of 153 variable candidates ($\sim$ 70%), and 500 out of 722 non-variable candidates ($\sim$ 69%), are real variables, so that the two percentages are comparable in this case.

While this indicates that the variability of stars and galaxies in this study could in principle be limited by the small number of epochs, the fact that majority of the variable candidates are real variable stars shows that the calibration refinement is successful in selecting those objects.

4.5   GALAXY VARIABILITY

4.5.1   Continuum variability per galaxy spectral type

To explore any dependence of spectral variability on spectral type of galaxies, we divide the final sample using two different spectral classification schemes: eClass (Connolly et al. 1995) and Osterbrock diagrams (Veilleux and Osterbrock 1987). These two schemes are motivated by different spectral features. The eClass is a spectral type describing the steepness of the continuum slope by using a linear combination of several eigenspectra for each galaxy spectra. The eClass parameters, ϕ_{KL} and θ_{KL}, are related to the coefficients in the eigenspectra expansion (called eigencoefficients), which are calculated by the spectro1d pipeline in the SDSS. Specifically, the angle ϕ_{KL} correlates with the Hα equivalent width (Madgwick et al.
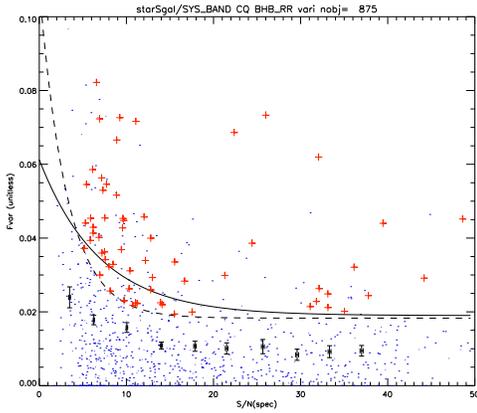
(a) Blue dots: F_{var} vs. S/N for all stars, red crosses: for stellar variable candidates, which are photometrically classified as blue horizontal branch (BHB) stars in the SDSS and also exist in the QUEST RR Lyrae First Catalog. Black line is the mean + 1σ averaged over all stars. Dashed line is that for the galaxy final sample.
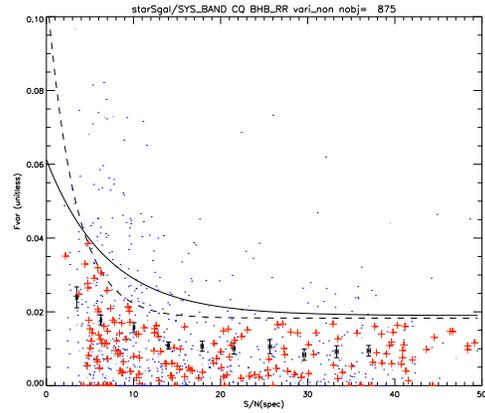
(b) Same as Fig. 12(a) but red crosses are for non-variable candidates defined in this work.

Figure 4.12 Matching stellar variables and non-varibales with the QUEST RR Lyrae first catalog.
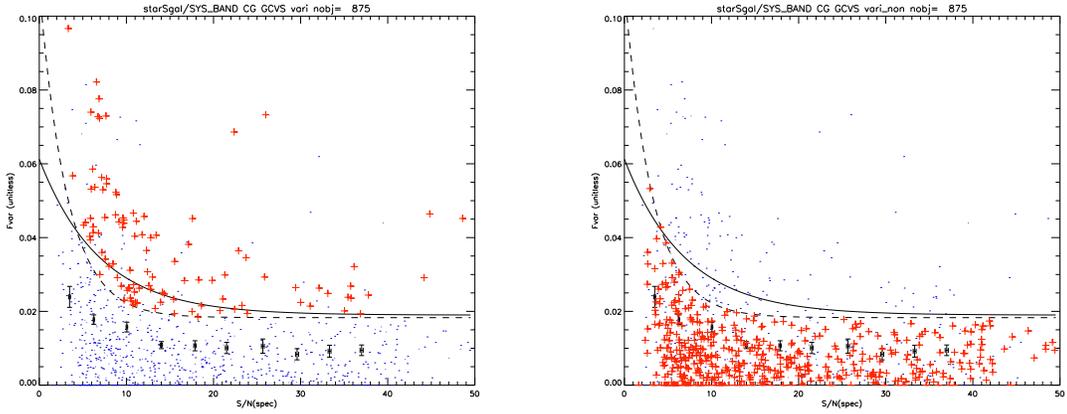
(a) Similar to Fig. 12(a) but the matching is done with the Genral Catalogue of Variable Stars. The matched objects here are not confined to be BHB stars. Similarly, a subset of the variable stellar candidates defined in this work are proved to be real variable stars.

(b) Same as Fig. 13(a) but red crosses show the stellar non-variables defined in this work.

Figure 4.13 Matching stellar variables and non-varibales with the GCVS catalog.

2003) which is an indicator of star formation rate. The angle θ_{KL} discriminates galaxy spectra exhibiting post-starburst activities (Connolly et al. 1995; Yip et al. 2004a).

On the other hand, the Osterbrock diagrams separate narrow line AGNs (powered by accretion disks) from galaxies such as H II galaxies (powered by stars) by emission line ratios which indicate the physical conditions of gas inside the galaxies such as the electron density and the mean level of ionization.

As expected, objects in these two spectral classification schemes are not mutually exclusive. An interesting future work is to discuss the relationship between the two schemes. We hereafter present results on variability in both classifications.

4.5.1.1   eClass   The spectral variability as a function of S/N per mean eClass galaxy type (defined in chapter 2) are plotted in Fig. 4.14–4.17. These are galaxy spectra each averaged over a non-trivial size of subspace of the two eClass parameters ϕ_{KL} and θ_{KL} (Yip et al. 2004a). Limited by the number of spectral pairs, only the types A, B, C and D are

considered. The continuum of the mean galaxy spectrum is reddest in the type A and bluest in the type D. In each figure, the left panel shows F_{var} vs. S/N and the right panel shows the histogram of F_{var} at S/N $= 10 \pm 1$.

Firstly, we note that the median, the mean and the scatter in the variability amplitude are different for different spectral types, with the values in the mean eClass type D being the largest. If we assume a priori that some populations of galaxies do not vary in time, a good assumption for early type galaxies, then a spectral repeatability can be assigned to the population with the smallest scatter in F_{var}, that is, the mean eClass type B. The spectral features of this class are similar to the elliptical galaxies in Kennicutt's 1992 atlas of nearby galaxies. The repeatability, taken to be the average F_{var}, is 0.68% at S/N $= 10$. The average variability for the mean eClass type D is larger and is 1.17% at S/N $= 10$.

4.5.1.2   Emission line galaxies: narrow line AGN vs. H II galaxies
We define emission line galaxy types by using the starburst modeling of Kewley et al. (2001). When applying the criteria which divides the narrow line AGNs and H II galaxies we confine ourselves to the Osterbrock diagram $\log_{10}([\mathrm{O\ III}]/\mathrm{H}\beta)$ vs. $\log_{10}([\mathrm{N\ II}]/\mathrm{H}\alpha)$, in which the EWs of emission lines are largest (hence of highest confidence) compared with those in the other diagrams. Galaxies with either one of those four lines having rest EW smaller than 1 Å are rejected. Further, the modeling line in Kewley et al. (2001) is shifted towards the 3rd quadrant by 0.1 dex, that was shown to better separate AGNs from H II galaxies in the SDSS (Miller et al. 2003; Hao et al. 2005). Stellar absorption in the related hydrogen emission lines in the Osterbrock diagram is corrected for by a constant increment of 0.7 Å (Hopkins et al. 2003; Miller et al. 2003).

In Fig. 4.18 and 4.19 we compare the variability amplitude between the H II galaxies and narrow line AGNs in the final sample. It is a little surprising that the average variability amplitude at S/N $= 10$ is 1.25% for H II galaxies, 0.51% higher than that of the narrow line AGNs.

We over-plot the variability of different galaxy types in Fig. 20(a) - 20(c). The difference in the variability amplitude is clearly seen, so that $F_{\mathrm{var}}(\text{type D}) > F_{\mathrm{var}}(\text{type A})$, $F_{\mathrm{var}}(\text{type BLUE}) > F_{\mathrm{var}}(\text{type RED})$ and $F_{\mathrm{var}}(\text{H II}) > F_{\mathrm{var}}(\text{narrow line AGN})$, for all S/N.
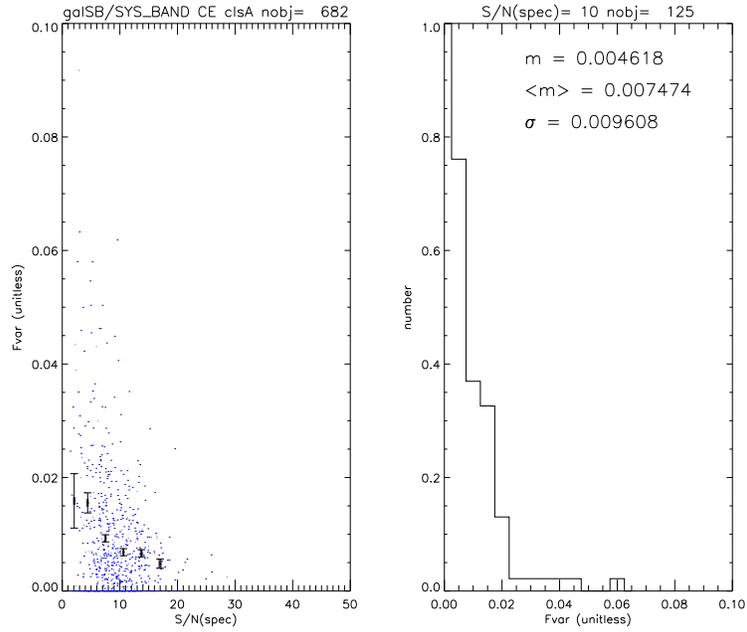
Figure 4.14 F_{var} vs. S/N for galaxies of the mean eClass type A (i.e. redder galaxies). See caption in Fig. 4.7 for re-calibration method and other details. The average F_{var} at $S/N = 10$ is below the average as shown in Fig. 4.7.
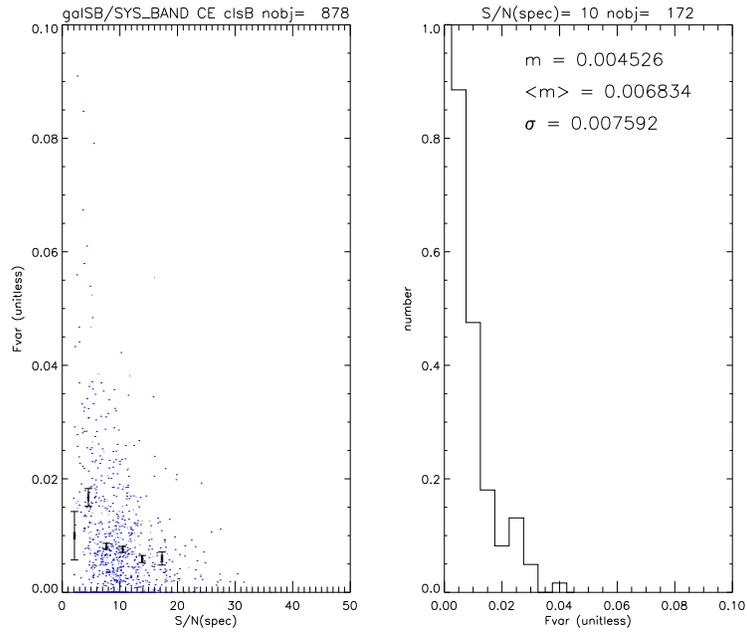
Figure 4.15 F_{var} vs. S/N for galaxies of the mean eClass type B.
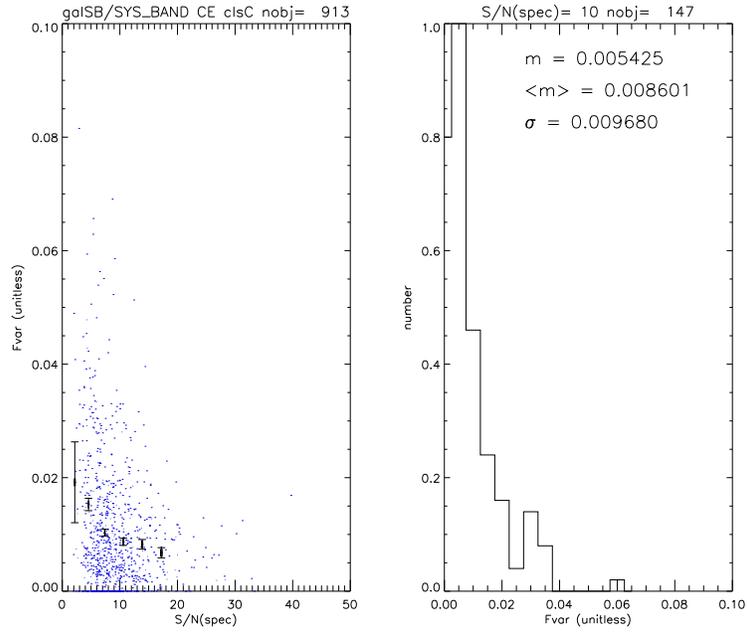
Figure 4.16 $F_{\mathrm{v}ar}$ vs. S/N for galaxies of the mean eClass type C.
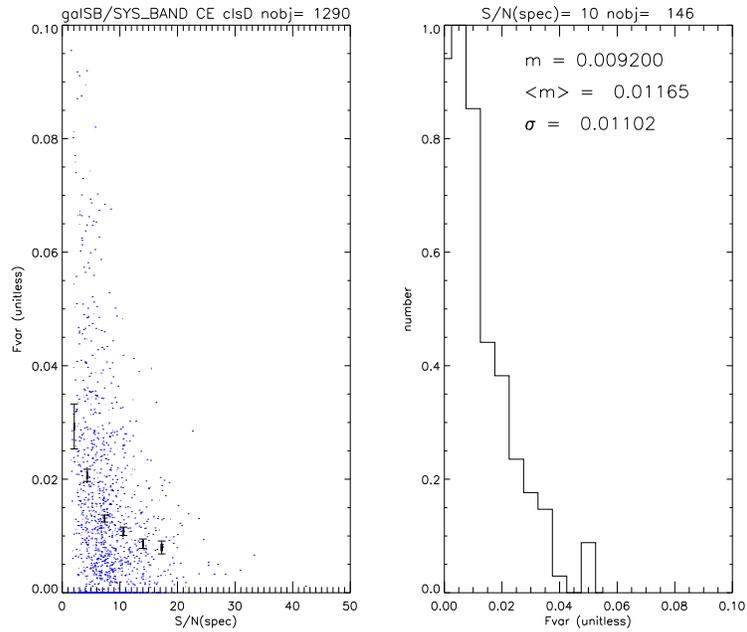
Figure 4.17 $F_{\mathrm{v}ar}$ vs. S/N for galaxies of the mean eClass type D (i.e. bluer galaxies). At S/N = 10 the distribution of $F_{\mathrm{v}ar}$ is broader than that of mean type A, B and C.
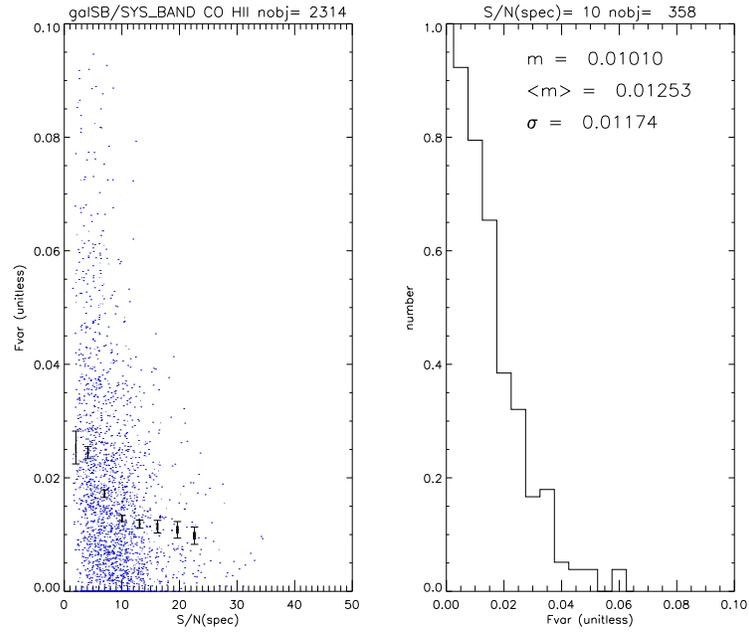
Figure 4.18 F_{var} vs. S/N for HII galaxies.
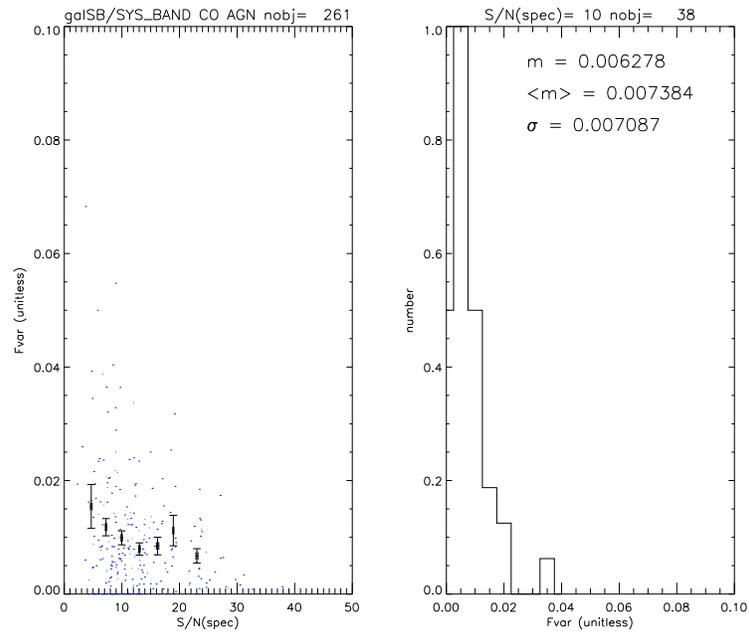
Figure 4.19 F_{var} vs. S/N for narrow line AGN.
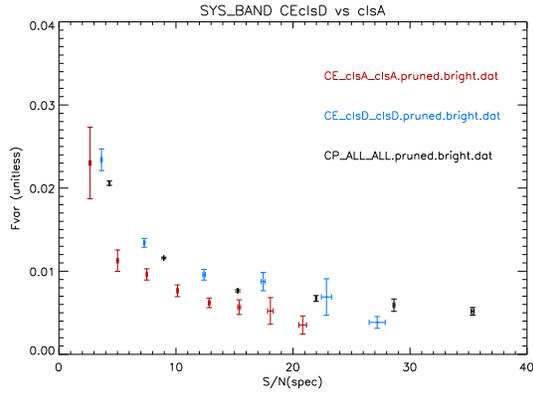
(a) The mean eClass types A vs. D.
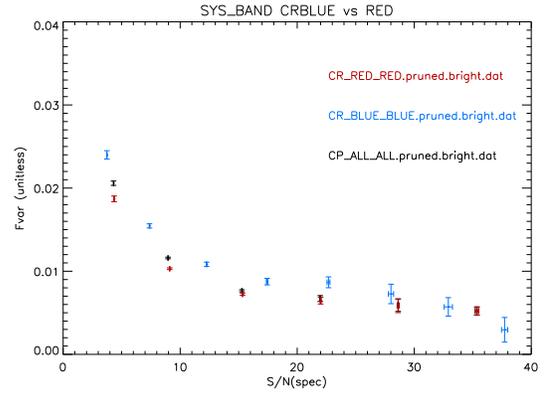
(b) The mean eClass types RED vs. BLUE.
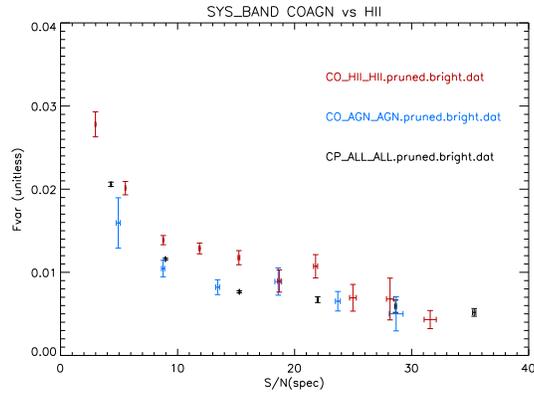
(c) The AGN vs. HII galaxies.

Figure 4.20 F_{var} vs. S/N per galaxy spectral type. Error bar is represented by 1 $\sigma(mean)$ in bin.

The possible physical reasons behind the results in both §4.5.1.1 and §4.5.1.2 are discussed in §4.5.3 when we present the difference spectra.

4.5.1.3 Aperture effect The SDSS $3''$ diameter fiber spectroscopy does not cause substantial aperture bias in classifying galaxy spectra (Miller et al. 2003; Yip et al. 2004a). In addition, under the typical sky conditions in the SDSS observation (the median PSF width $= 1.4''$ in the r band) and given the diameter of the fiber, the aperture correction factor is close to unity (Wanders et al. 1992), meaning that no correction is needed to the total flux of narrow emission lines. Veilleux et al. (1995) have argued the importance of using linear aperture spectroscopy in general line diagnoses, that should be a minor concern for us as we are primarily interested in the continuum variability in this work.

4.5.2 Structure functions per galaxy spectral type

We investigate the structure function (SF), the variability amplitude F_{var} as a function of time lag, for the whole final sample and per galaxy spectral type.

The restframe SF, F_{var} vs $\Delta\tau_{\mathrm{rest}}$, of all galaxies in the final sample is illustrated in Fig. 4.21, where $\Delta\tau_{\mathrm{rest}} = \Delta\tau/(1+z)$, and z being the redshift of the galaxy. The time lag $\Delta\tau$ is the difference between two MJDs of observation, $|\mathrm{MJD(epoch1)} - \mathrm{MJD(epoch2)}|$.

In Fig. 22(a) - 22(c) we show restframe SF per galaxy spectral type. As a function of time lag, the dependency of variability on spectral type is similar to those seen in F_{var} vs. S/N (§4.5.1). No obvious time scale is seen in the above restframe SFs for 400 days sampled, for all galaxy spectral types considered in this work. There appears to be a slight increase in the variability with time lag for the types BLUE, RED, H II and AGNs. However, we make no attempts to fit functional forms to the SF due to the short time baseline considered here.

This lack of a timescale may be either due to the short time baseline, or may be of physical origin, such as relating to the difference in underlying mechanisms driving the spectral variability in galaxies and in broad line AGNs. For comparison, QSOs were found to vary on the time scale of months to years in the optical. If a similar time scale of variability exists in narrow line AGNs, a baseline longer than 1 year is needed.
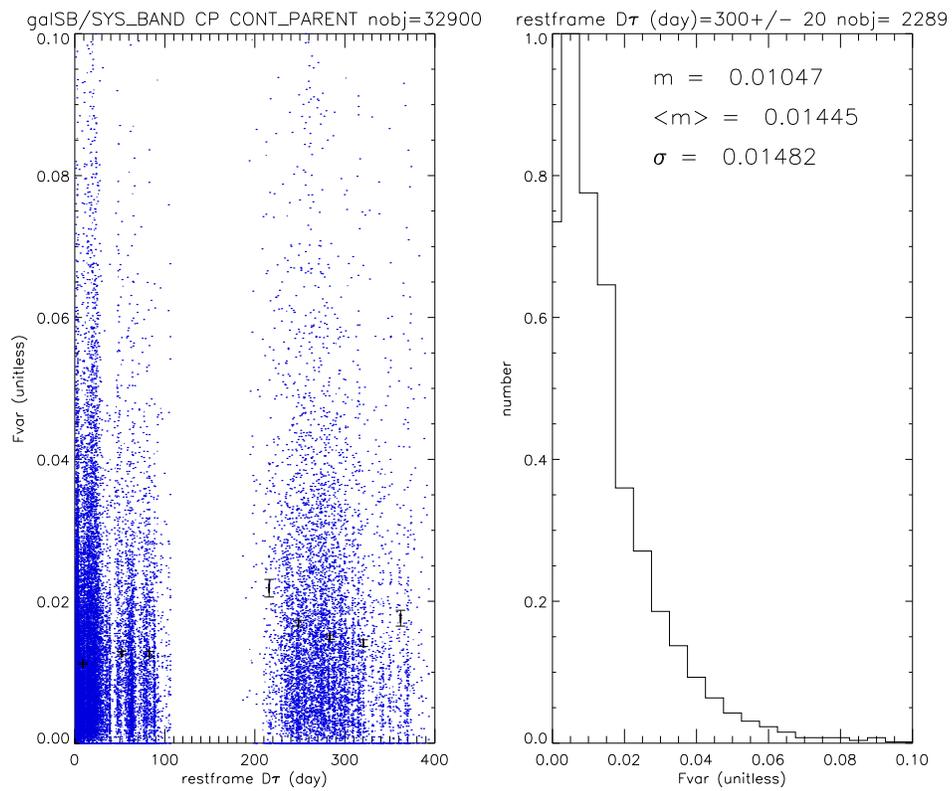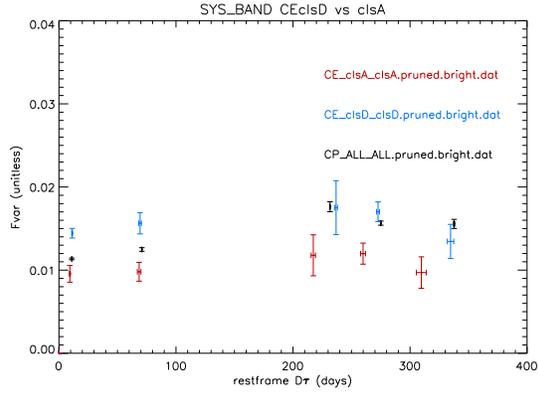
Figure 4.21 Restframe structure function for all galaxies in the final sample.

(a) The mean eClass types A vs. type D.
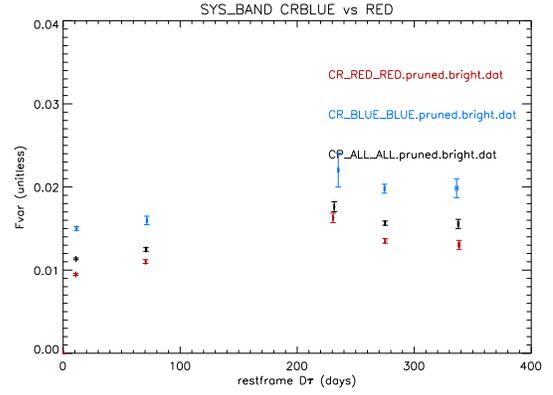
(b) The mean eClass types RED vs. BLUE.
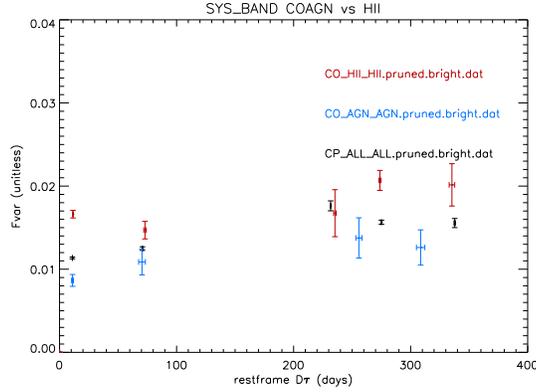
(c) The AGN vs. HII galaxies.

Figure 4.22 Restframe structure function per galaxy spectral type. Error bar is represented by 1 $\sigma(mean)$ in bin.

4.5.3 Difference spectra

Simple thin disk models for AGN predict the relationship between disk temperature and continuum slopes. Hence a fluctuation in temperature can cause the continuum slope to change. In order to explore any AGN component in a galaxy, we study the galaxy difference spectrum between two epochs. The difference spectrum is defined as f_λ(bright phase) $-$ f_λ(dim phase). The fractional difference spectrum is defined as $(f_\lambda$(bright phase) $-$ f_λ(dim phase))$/f_\lambda$(bright phase). In the calculations we use Gap-Correction (Connolly and Szalay 1999) to increase the signal-to-noise. Each difference spectrum is calculated as a median composite (Vanden Berk et al. 2001).

In Fig. 23(a)–24(d) we show the average difference spectra between two epochs for different mean eClass types. For eClass A, the difference spectrum is not well fit by a power law of a negative spectral index α_λ. This result suggests that there is probably no spectral variability due to nuclear activity in this type of galaxies, as expected.

In Fig. 25(a)–26(d) we show the difference spectra of emission line galaxy types defined by the Osterbrock diagrams. Assuming each different spectrum can be described by a spectral energy distribution of the form $f_\lambda \propto \lambda^{\alpha_\lambda}$, and fitting over the restframe wavelengths $3450 - 6400$ Å for the variable candidates, we find that for narrow line AGNs, $\alpha_\lambda = -1.79$, for H II galaxies; $\alpha_\lambda = -2.59$; and for eClass D, $\alpha_\lambda = -1.56$.

In Table 4.1 we summarize the spectral index of the power law least-squares fit to the difference spectrum per spectral type, and the respective χ^2. For a given spectral type, generally the variable candidates are better fit by a power law than the non-variables with the χ^2 approximately a factor of 10 smaller.

In Table 4.2 we show the apparent AB magnitudes in g, r and i bands of variable candidates per spectral type in both bright and dim phases. The magnitudes are calculated by convolving the SDSS filters using the transmission curves for extended sources at an airmass of 1.3. The difference in the g band magnitudes of the two phases ranges from $0.03 - 0.1$ depending on the spectral type.

While the spectral indices of variable candidates in narrow line AGN, eClass and H II galaxies respectively and coincidently agree well with the QSO difference spectrum in the
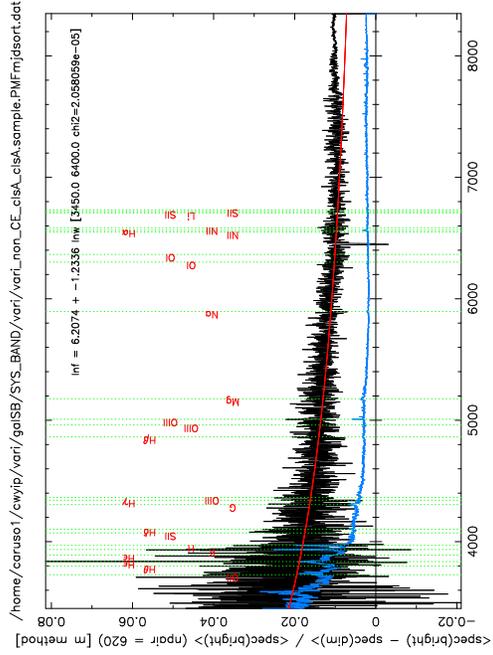
(a) The mean eClass type A, non-variable.
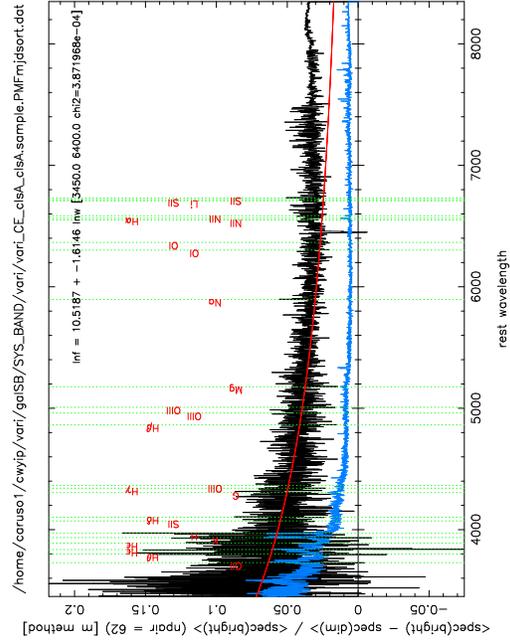
(b) The mean eClass type A, variable.

(c) The mean eClass type D, non-variable.

(d) The mean eClass type D, variable.

Figure 4.23 Difference galaxy spectrum for the eClass types A and D (black). Blue line is the mean spectrum of uncertainity. Red line is the best-fit power law. Spectra are in restframe.

(a) The mean eClass type A, non-variable.

(b) The mean eClass type A, variable.
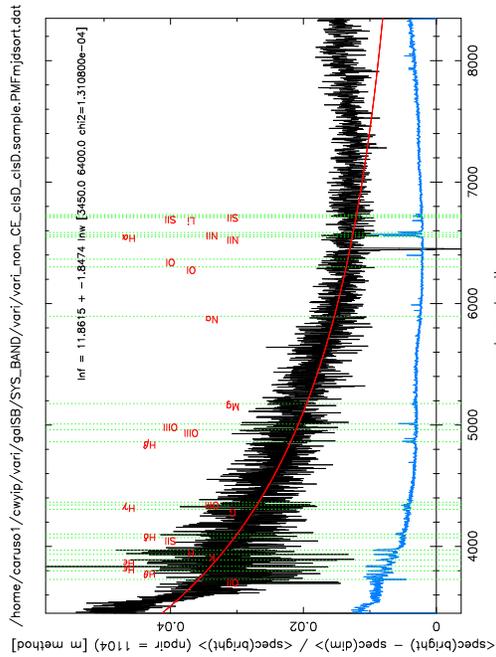
(c) The mean eClass type D, non-variable.
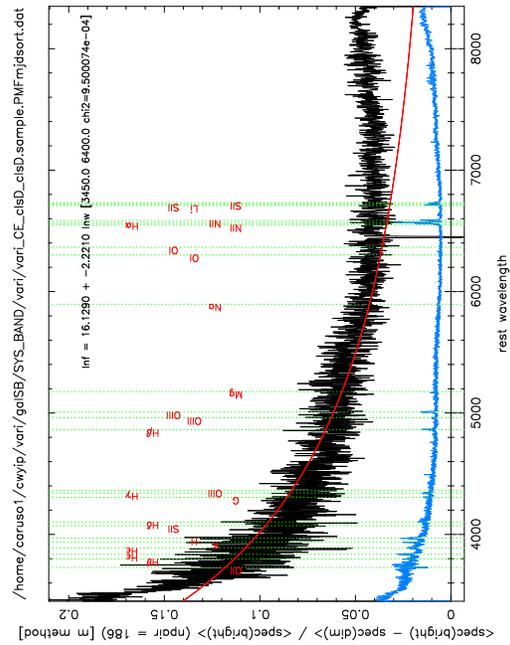
(d) The mean eClass type D, variable.

Figure 4.24 Fractional difference galaxy spectrum for the mean eClass types A and D.
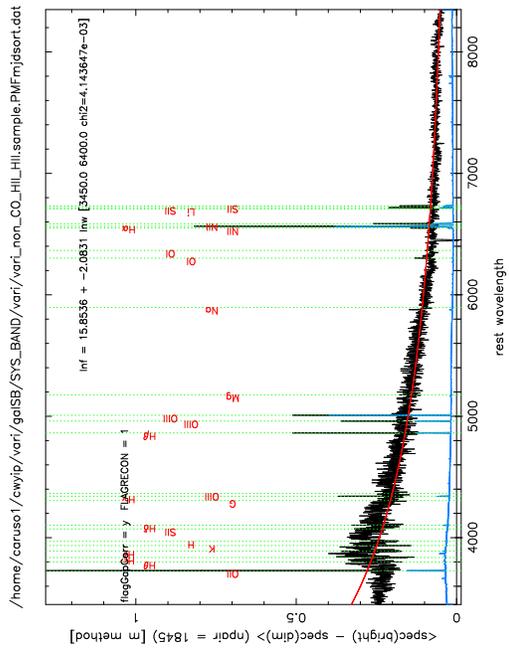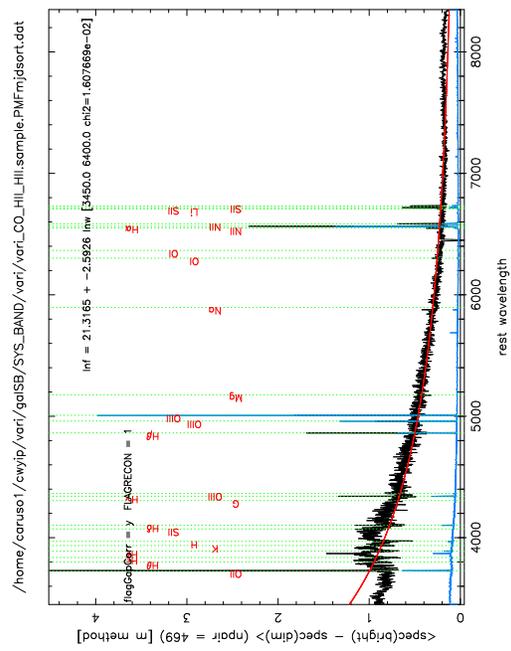
137

(a) The HII galaxies, non-variable.

(b) The HII galaxies, variable.
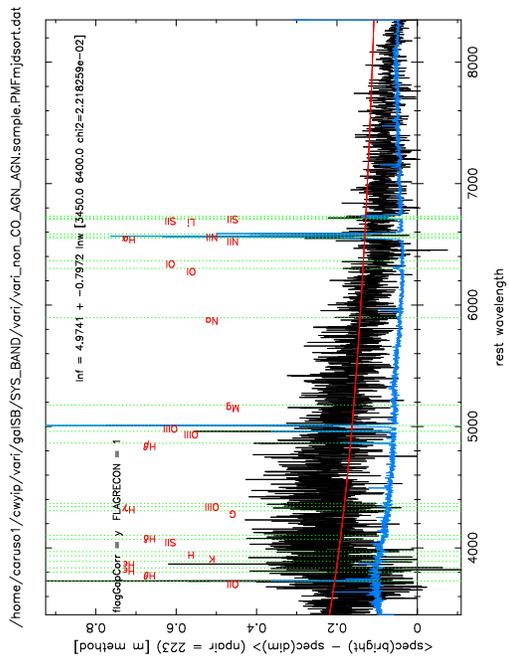
(c) Narrow-line AGNs, non-variable.
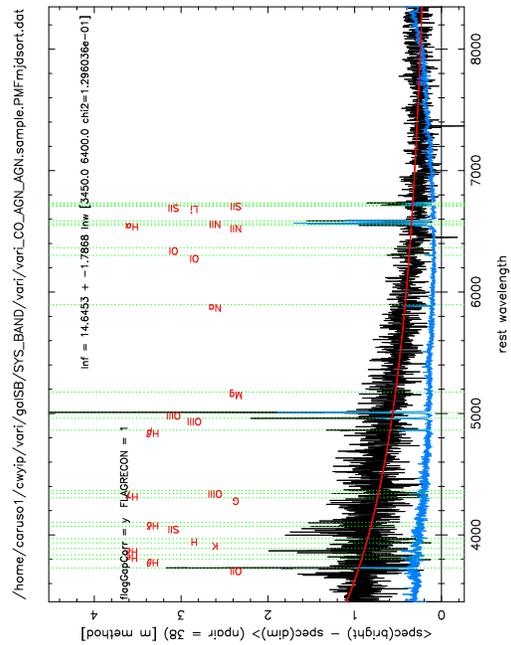
(d) Narrow-line AGNs, variable.

Figure 4.25 Difference spectrum for HII galaxies and narrow line AGNs.
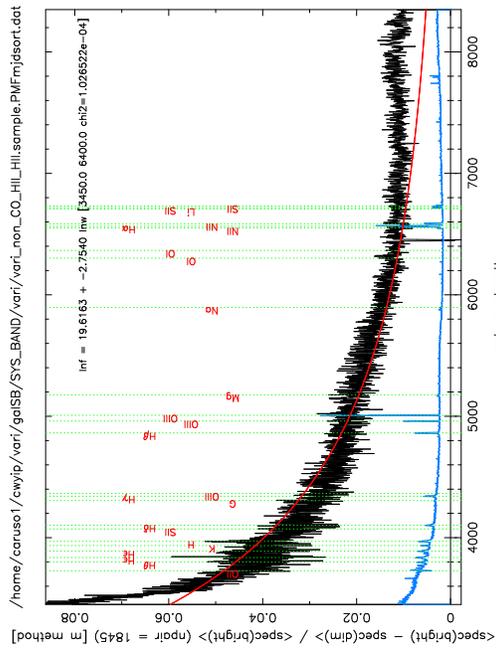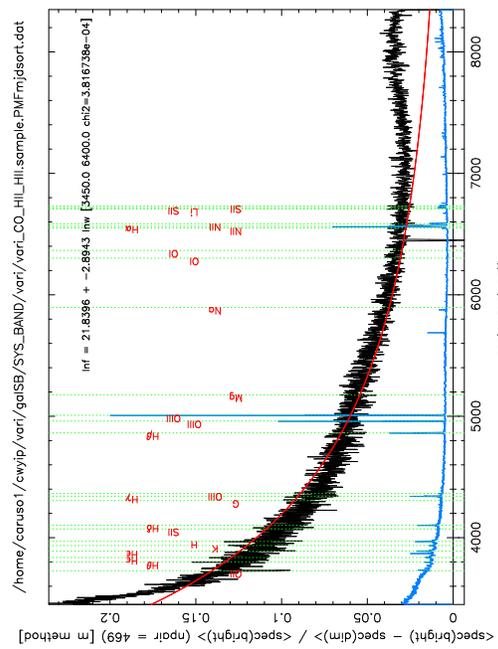
138

(a) The HII galaxies, non-variable.

(b) The HII galaxies, variable.
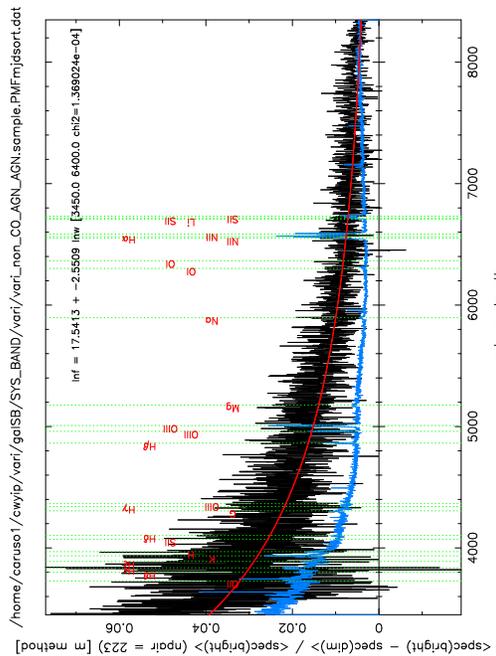
(c) Narrow-line AGNs, non-variable.
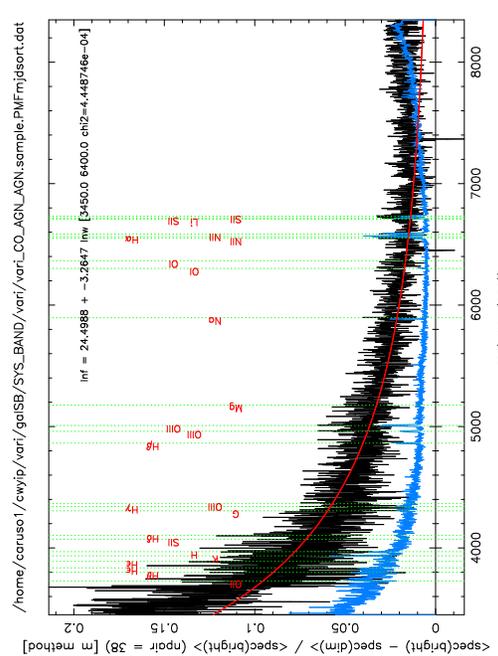
(d) Narrow-line AGNs, variable.

Figure 4.26 Fractional difference galaxy spectrum for HII galaxies and narrow line AGNs.

Table 4.1 Spectral indices of galaxy difference spectra.

	eClass A	eClass D	AGN	HII
non-variable candidates α_λ[a]	1.53	-1.56	-1.79	-2.59
χ^2[b]	2.87e-2	4.05e-2	1.30e-1	1.61e-2
variable candidates α_λ	1.85	-1.16	-0.80	-2.08
χ^2	3.54e-3	6.08e-3	2.22e-2	4.14e-3

-----

[a]Assuming a power law spectrum $f_\lambda \propto \lambda^{\alpha_\lambda}$ in restframe $3450 - 6400$ Å.

[b]Chi-square, weighted by unity per pixel and discarding emission and absroption lines.

Table 4.2 AB synthetic magnitudes of variable candidates.

	g	r	i
AGN, bright	18.69	17.97	17.68
dim	18.72	17.99	17.69
HII, bright	19.40	18.96	18.81
dim	19.48	19.01	18.85
eClass A, bright	19.33	18.50	18.15
dim	19.37	18.54	18.18
eClass D, bright	19.57	19.06	18.85
dim	19.67	19.11	18.90

SDSS in which $\alpha_\lambda = -2.00$ (Wilhite et al. 2005b), the possible physical mechanism driving the variabilities of these galaxy types maybe different from that in the QSO variability. Traditionally the H II galaxies are considered as passive, in contrast with AGNs in which nuclear activities are expected. So it may seem surprising to see the power law component in the H II galaxies. The nature of this variability, however, may not be related to an accretion disk at all. Other kinds of outburst events such as starbursts or supernova explosions can in principle be causing the spectral variability seen, so that flux densities at shorter wavelengths vary with a larger amplitude than the longer wavelengths.

4.5.4 Starburst modeling

To further investigate the difference spectrum of H II galaxies, we fit the spectra using the publicly available Starburst 99 synthesis models (Leitherer et al. 1999). This library of spectra is optimized to reproduce properties of galaxies which exhibit active star formation in wavelengths 91 Å to 160 μm, and is also applicable to less extreme star forming galaxies. The star formation laws are assumed to be either an instantaneous burst of star formation or a continuous star formation of rate 1 $M_\odot$ yr^{-1}.

We adopt the tabulated reference models in which a power law stellar initial mass function (IMF) with an exponent $\alpha = 2.35$ is used, with the lower cutoff mass being 1 $M_\odot$ and higher cutoff mass being either 30 or 100 $M_\odot$, forming approximately the Salpeter IMF. As there is no gaseous component in the synthesis models, we exclude the major emission lines in the difference spectrum during the fitting. The best fit model and the difference spectrum of H II galaxies are shown in Fig. 27(a). The best-fit parameters have a higher cutoff mass = 30 $M_\odot$, metallicity (Z) = 0.02 and age = 9 Myr. The fitting is limited by the resolution in the early ages and the lower bound of the age sequence (both are 1 Myr). Since the time baseline is only 400 days in this study, the best-fit age is obviously unphysical. However, the spectral features in the difference galaxy spectrum agree well with those in the best-fit starburst synthesis model. If the starburst is one of the physical processes producing the variability seen, an extra spectral component, presumably not starburst-related, is necessary in the modeling.
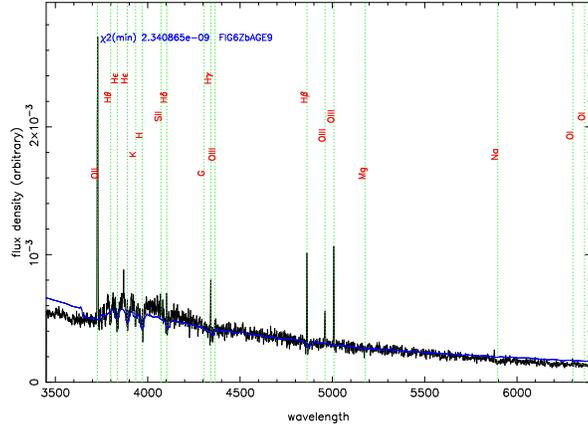
(a) The HII galaxies.
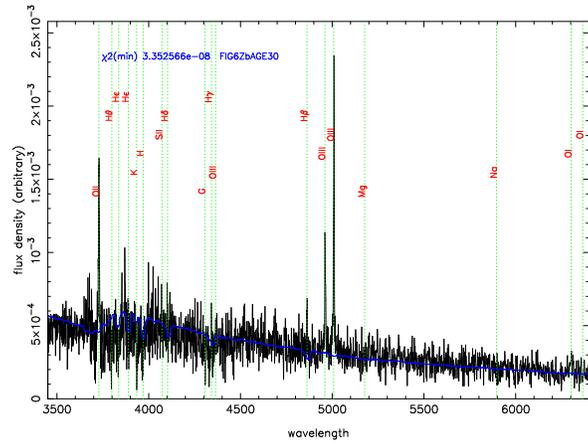
(b) The narrow line AGNs.
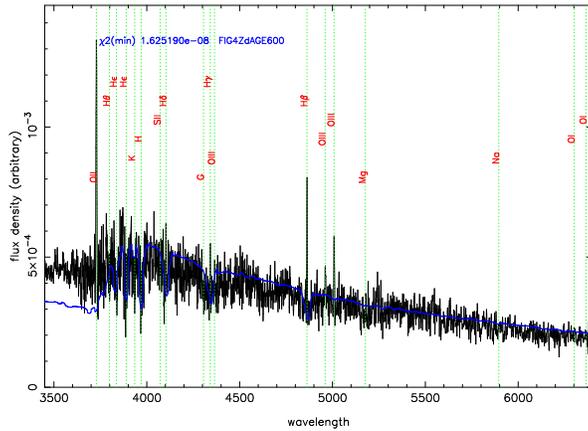
(c) The mean eClass type D.

Figure 4.27 Starburst systhesis modeling to difference spectrum of variable candidates. Spectra are in restframe.

142

In contrast, the starburst models do not fit well to the difference spectrum of narrow line AGNs (Fig. 27(b)). The χ^2 is about a factor of 10 larger in this case. Similarly for the mean eClass type D in which the χ^2 is about a factor of 10 larger.

4.5.5 Variables vs. non-variables

4.5.5.1 Effect of seeing and guiding offset To ensure that our selection method is not biased by instrumental uncertainties and observational conditions, we perform the Kolmogorov-Smirnov (KS) test between variable and non-variable samples for the parameter seeing50 (50% seeing during exposure in arcsecond) and rmsoff50 (50% rms offset of the guide fiber in arcsecond). These parameters are chosen for they represent instrumental effects and are not related to the intrinsic properties of the galaxies. They are recorded in the spSpec files by the spectro1d pipeline. Table 4.3 lists the KS statistics and probability of the tests for different galaxy types (eClass A, D, RED, BLUE; AGN and H II). For most galaxy types, the KS probability is larger than the 0.05 significance level, meaning that the cumulative distributions of the concerned parameter between the variables and non-variables are likely drawn from the same distribution. Hence, the difference in the variability is unlikely due to these instrumental and observational effects. However, our results show that statistically the variable and non-variable candidates of the eClass type RED may be subjected to both effects of seeing and guiding offset.

4.5.5.2 Mean galaxy type The AGN median composite spectra of variables and non-variables are found to locate differently on the Osterbrock diagram (Fig. 4.28). The EWs of the emission lines are calculated with a procedure described in appendix B. The mean level of ionization and temperature, as indicated by the line ratio $\log_{10}([O\ III]/H\beta)$, is higher in the variable AGN candidates than the non-variable ones. For both the H II galaxies and the mean eClass type D, no such difference is seen.

As the AGN variables and non-variables are selected by continuum variability, it is not obvious why there is a difference in the emission line diagnostics. The result may indicate that the AGN population with higher mean state of ionization tends to exhibit spectral

143

Table 4.3 KS tests between variable and non-variable samples.

		effective number[a]	KS statistics[b]	KS probability[c]
AGN	seeing50	32.47	0.0913	0.9387
	rmsoff50	32.47	0.1057	0.8407
HII	seeing50	373.94	0.0688	0.0553
	rmsoff50	373.94	0.0543	0.2143
eClass A	seeing50	56.36	0.1613	0.0958
	rmsoff50	56.36	0.1694	0.0702
eClass D	seeing50	159.18	0.0573	0.6604
	rmsoff50	159.18	0.0897	0.1466
eClass RED	seeing50	2185.77	0.0411	0.0012
	rmsoff50	2185.77	0.0328	0.0178
eClass BLUE	seeing50	1323.41	0.0367	0.0553
	rmsoff50	1323.41	0.0324	0.1210

[a]Effective number is $n_1 n_2/(n_1 + n_2)$, n_1 and n_2 being the number of objects in the two samples respectively.

[b]The Kolmogorov-Smirnov statistic. It is the maximum deviation between the cumulative distribution of the two samples being tested.

[c]If the probability is larger than the significance level (chosen to be 0.05), we do not reject the null hypothesis that two samples are being drawn from the same distribution.

variability.

4.6 SDSS SOUTHERN SURVEY: GALAXY VS STELLAR VARIABILITY

To evaluate the spectroscopic variability measurements we use the photometric co-added catalog of the SDSS Southern Stripes constructed by Ryan Scranton (private communication) to extend our analysis to a time baseline of 2,500 days in the observed frame and to a larger set of epochs of observations (about 10–20 epochs per object). As spectroscopic redshifts for many objects in the catalog are not yet available, we firstly compare the ensemble SF between galaxies and stars from this catalog in the observed frame. The separation of stars and galaxies is carried out by the SDSS photometric pipeline (Lupton et al. 2001), which compare psf and cmodel apparent magnitudes and estimates the morphology of the source (extended or point-liked). At model magnitudes $16.0 - 21.0$ the recovering fractions are very good for both stars and galaxies and are close to $0.8 - 1$. We confine our analysis to $16.0 - 21.0$ in respective magnitudes for both types of objects. The apparent magnitudes are converted to fluxes for the purpose of comparing with the spectroscopic variability.

Solely from photometry we cannot discriminate broad line AGNs, narrow line AGNs and passive galaxies, hence the galaxy photometric sample can be contaminated by the broad line ones. The fraction, however, should be < 0.001 (Osterbrock 1989).

In calculating the ensemble SF we permute the available epochs of observation for a maximum number of time lags, and the results are shown in Fig. 4.29 and 4.30. In each figure, the left panel shows the observed frame SF for g, r, i bands and the right shows $F_{\mathrm{var}}(1000$ days) as a function of effective wavelength of the bands. Each line on the left panel is a least-squares linear fit to the SF. In principle the modeling of the SF can be done with other functional forms such as power laws or asymptotic forms, which are higher-order effects in the current context of examining the wavelength dependency. Galaxy photometric variability as measured from the fiber magnitude-converted flux is found to be larger than that of stars, which uses psf magnitude-converted flux . Assuming most stars are non-variables at some level intrinsically, the variability in this case is actually the repeatability of the photometric
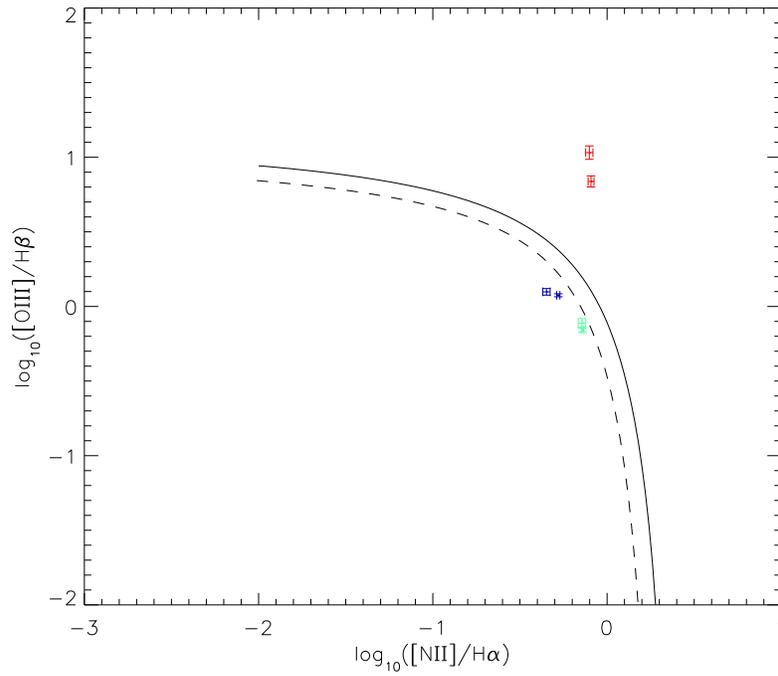
Figure 4.28 Red: narrow line AGNs; blue: H II galaxies; green: the mean eClass type D. In each duo, the point with smaller error bars represents the non-variable candidates. The AGN variables have higher mean level of ionization than the AGN non-variables.
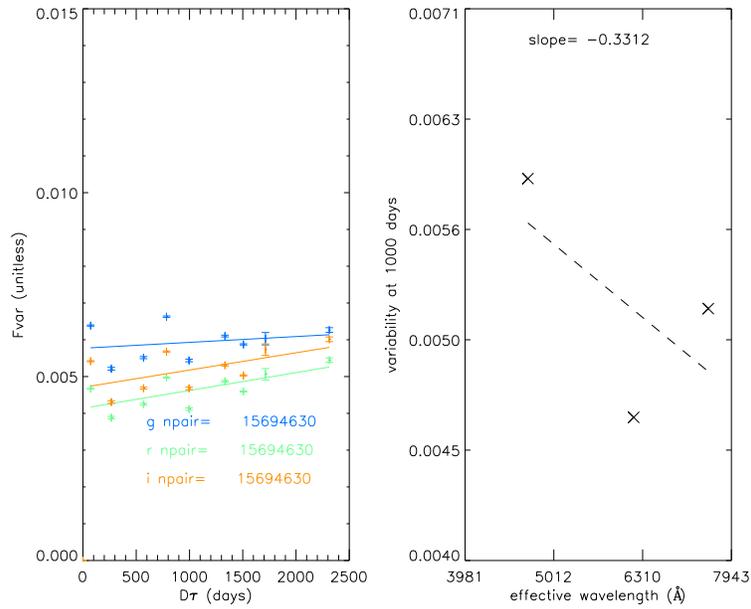
Figure 4.29 Observed frame structure function for g, r, i bands of repeatedly observed stars in the SDSS Southern Survey, using psf magnitude-converted flux. Right: variability at time lag of 1000 days as a function of effective wavelength of the bands.
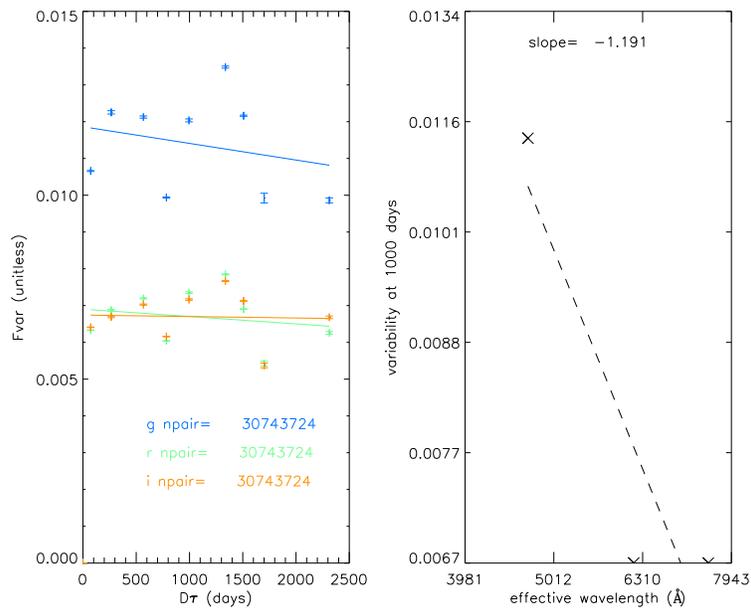
Figure 4.30 Observed frame structure function of repeatedly observed galaxies in the SDSS Southern Survey, using fiber magnitude-converted flux.
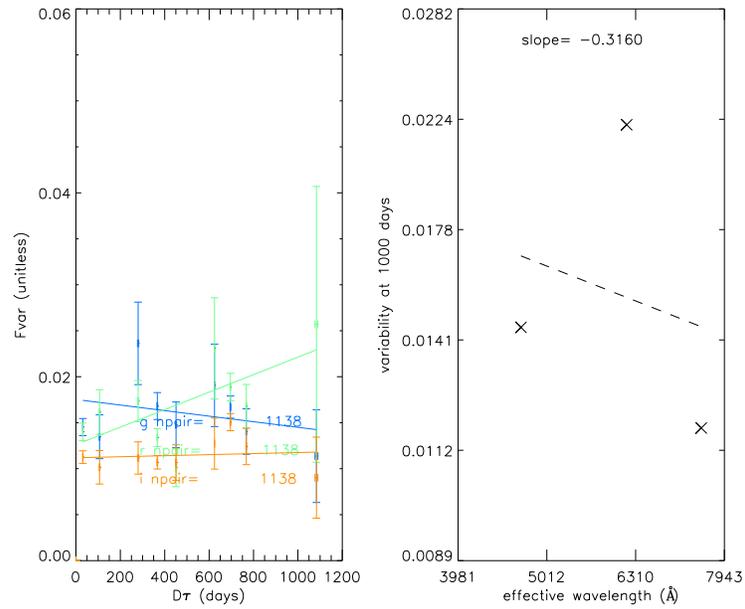
Figure 4.31 Observed frame structure function for non-variable candidates of narrow line AGN, using fiber magnitude-converted flux.
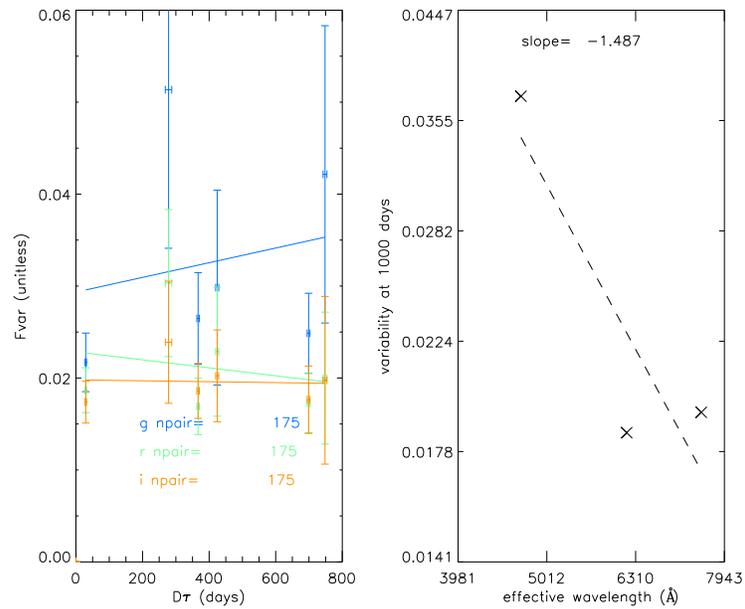
Figure 4.32 Observed frame structure function for variable candidates of narrow line AGN, using fiber magnitude-converted flux.
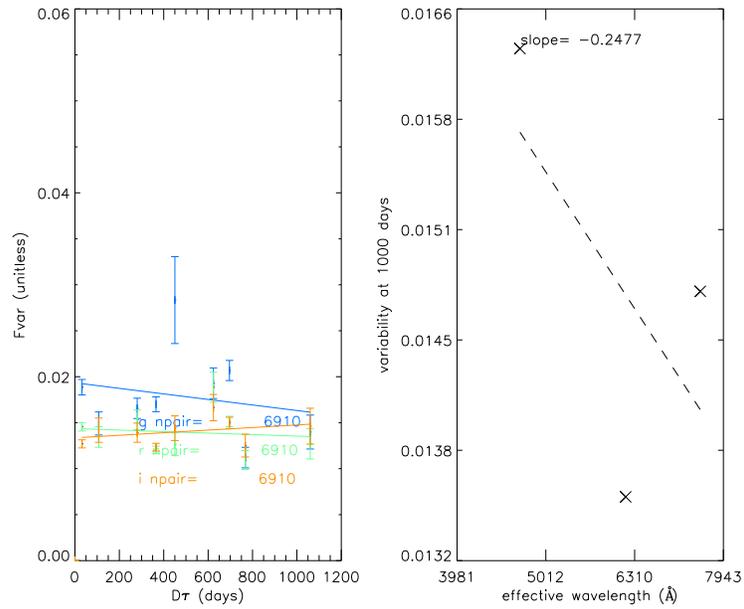
Figure 4.33 Observed frame structure function for non-variable candidates of H II galaxies, using fiber magnitude-converted flux.
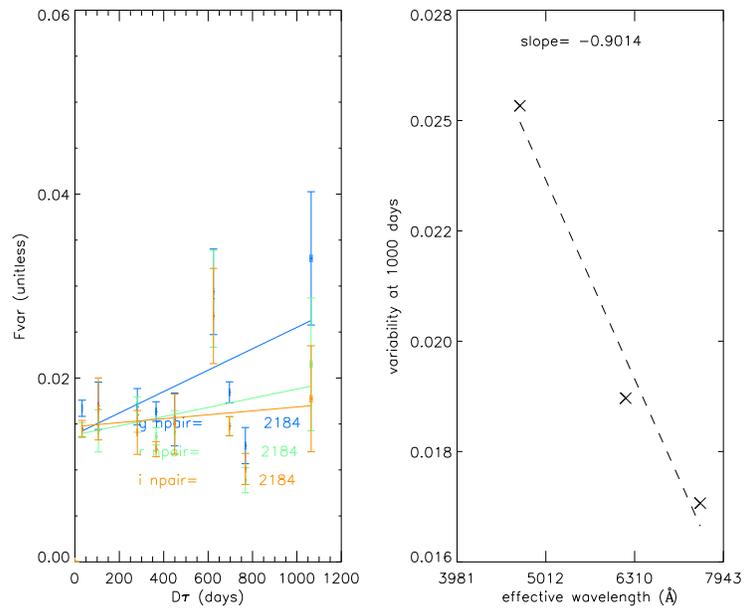
Figure 4.34 Observed frame structure function for variable candidates of H II galaxies, using fiber magnitude-converted flux.

measurements. Non-zero stellar repeatability is probability due to extrinsic effects such as seeing and scintillations, which is a small effect ($\sim 0.4 - 0.6\%$ for g, r, i bands) as shown in Fig. 4.29. For galaxies, F_{var} is approximately $0.7 - 1.2\%$ for g, r, i bands (Fig. 4.30). The result will not be affected if a substantial number of stars are intrinsically variables.

The amplitude of photometric variability from the Southern Survey, about $0.4 - 0.6\%$ for repeatability and $0.7 - 1.2\%$ for ensemble galaxy variability, respectively agree very well with the spectroscopic repeatability and variability we found earlier (§4.5.1.1).

On the wavelength dependence, the galaxies are found to vary with a larger amplitude in the g band than the r and i bands (right panel of Fig. 4.30). This confirms the spectral variability found in this work, and also numerous other studies of AGNs, in which the flux at shorter wavelengths are found to vary with a larger amplitude than at longer wavelengths.

The spectroscopic variables and non-variables defined above are matched with the co-added catalog and the ensemble SF of AGNs are shown in Fig. 4.31 and 4.32 for non-variables and variables respectively. The variability in the variable sample is much more obvious for shorter wavelengths, with $F_{\mathrm{var}}(g) > F_{\mathrm{var}}(r), F_{\mathrm{var}}(i)$, with the average variability the g band about 3%. There are unfortunately only 4 objects found in the matching. For the non-variable sample (22 objects), the SF does not show wavelength dependency. In Fig. 4.33 and 4.34 we similarly show the SFs of the matched H II-likes galaxies. The slope describing the wavelength dependency of variability is approximately a factor of 4 larger in the variable sample than in the non-variable sample.

5.0 CONCLUSIONS

But at times I wondered if I had not come a long way only to find that what I really sought was something I had left behind.

Thomas Hornbein, in Everest: The West Ridge

5.1 GALAXY TYPES

From the application of Karhunen-Loève transform, an objective classification of $\sim 170,000$ galaxy spectra in the SDSS is performed. With a quantitative convergence criteria defined, gappy galaxy spectra can be repaired and KL eigenspectra and eigencoefficients derived. For most of the galaxy types, three eigenspectra are sufficient for describing the continua and emission lines to a high degree of accuracy with a maximum error in line-reconstructions of approximately 10%. Typically ten modes are needed in the reconstruction of galaxies with extreme emission lines with errors of $15 - 25\%$ in the line fluxes. We find that a two-parameter (ϕ_{KL}, θ_{KL})-classification scheme can discriminate between spectra corresponding to all spectral types used in the current classification scheme (including galaxies with extreme emission lines). This classification is robust to repeat observations (at a level of a few degrees in the classification angles) due to the accurate spectrophotometric calibration of the SDSS data set. We find a weak dependence in the classification on the signal-to-noise of the spectra. This effect is, however, smaller than the typical dispersion between repeat observations and is negligible at signal-to-noise levels at which the SDSS spectra are defined as being of survey quality.

We find that there exists a minimum number of randomly selected spectra that are neces-

sary to statistically represent the information within the full sample (i.e., to be representative of the true distribution of galaxies). For a set of ten eigenspectra (i.e., ten eigenspectra enable the reproduction of both quiescent and active galaxies) the number of spectra required is around 3000 to 4000. This is due to the need to sample a minimum number of randomly selected galaxies in order to include galaxies with extreme emission line properties in our data set (as they comprise only 0.1% of the full galaxy sample).

We find that the bias on the spectral classification due to the fixed aperture spectroscopy is, on average, small and is negligible for all galaxies except for the reddest galaxies that are very close by ($z/z_{max} < 0.3$) and for those galaxies that are large physically (> 30 kpc) with prominent emission lines. A look-up table is constructed for the correction of this bias.

5.2   QSO TYPES

We perform KL transforms and gap-corrections on 16,707 SDSS quasar spectra. In rest-wavelengths $900 - 8000$ Å, the 1st eigenspectrum (i.e., the mean spectrum) shows agreement with the SDSS composite quasar spectrum (Vanden Berk et al. 2001), with an abrupt change in the spectral slope around 4000 Å. The 2nd eigenspectrum carries the host-galaxy contributions to the quasar spectra, hence the removal of this mode can probably prevent the obscuration of the real physics of galactic nuclei by the stellar components. Whether this eigenspectrum is the only one containing the galaxy information requires further study. The 3rd eigenspectrum mainly shows the modulation between the UV and the optical spectral slope, in agreement with the 2nd principal component of Shang et al. (2003). The 4th eigenspectrum mainly shows the correlations between Balmer emission lines. Locally around various broad emission lines, the eigenspectra from the wavelength-selected quasars qualitatively agree with those from the Large Bright Quasar Survey, the properties in the Eigenvector-1 (Boroson and Green 1992), and the anti-correlations between the FWHMs and the equivalent widths of Lyα and C IV (Francis et al. 1992). The anti-correlation between the FWHM and the equivalent width is found not to be a general property of all broad emission lines (e.g., Mg II is discrepant).

From the commonality analysis of the subspaces spanned by the eigenspectra in different redshifts and luminosities, the spectral classification of quasars is shown to be redshift and luminosity dependent. Therefore, we can either use of order 10 (M_i, z)-bin eigenspectra, or of order $(10)^2$ global eigenspectra to represent most (on average 95%) quasars in the sample.

We find that the first two modes can describe the spectral slopes of the quasars in all (M_i, z)-bins under study, which is the most significant sample variance of the current QSO catalog. The simplest classification scheme can be achieved based on the first two eigencoefficients, so that a physical sequence can be formed upon the linear-combinations of the first two eigenspectra. The diversities in the quasar spectra, and the inevitable different restframe wavelength coverages due to the nature of the survey, increase the sparseness of the data. Hence, higher-order modes enter into the construction of the broad emission lines with the eigenspectra, in contrast to the galaxy spectral classification, in which most emission lines vary monotonically with the spectral slope (Connolly et al. 1995). *The result is also a manifestation of the high uniformity of galaxy spectra compared with quasar spectra.*

We find that the BAL features do not only appear in one particular order of eigenspectrum but span a number of orders, mainly higher-orders. This may indicate substantial challenges to the classification of BAL quasars by the current sets of eigenspectra in terms of arriving a compact description, and hence a separate KL-analysis of the BAL quasars is desirable for studying the classification problem. Nonetheless, the appropriate truncation of orders of eigenspectra in reconstructing a quasar spectrum, if shown to be exist in future, in principle can lead to an unabsorbed continuum.

We find evolution of the small bump by the cross-redshift KL transforms, in agreement with the quasars from the Large Bright Quasar Survey (Green et al. 2001) and in another independent work (Kuhn et al. 2001). The Baldwin effect is detected in the first two eigenspectra of the higher redshift quasars, as well as the first eigenspectra derived with quasars with different luminosities. One implication of these redshift and luminosity effects is that they have to be accounted for in the spectral classification of quasars. This is consistent with our finding from the commonality analysis.

The high quality of the data allows us to obtain quasar eigenspectra which are generic enough to study spectral properties. Despite the presence of diverse quasar properties such

as different slopes and shapes and various emission line features known for several decades, our analysis shows that there are unambiguous correlations among various broad emission lines and with continua in different windows.

5.3 ROLE OF SPECTRAL VARIABILITY IN GALAXY CLASSIFICATION

We study spectroscopic variability of galaxies and Type 2 active galactic nuclei (AGNs) in the Sloan Digital Sky Survey (SDSS). The sample contains $\sim 33,000$ objects which are observed twice within a time baseline of 400 days. The amplitude of galaxy variability is found to depend on spectral type. In addition, a power law continuum is detected in the average difference spectrum in the relatively blue galaxies (type D, defined by $\phi_{KL} \ll 0^o$ in eClass), the Type 2 AGNs and maybe a little surprising, the H II-liked galaxies, which are not believed to harbor nuclear activity as suggested by the Osterbrock diagrams. Spectral fittings to the difference spectra using starburst synthesis models suggests those the variability in the H II-liked galaxies can be partly due to star formation; and that in the AGNs and the eClass type D are probably not related to starbursts. The narrow line AGN variables are found to have higher mean level of ionization than the non-variables. By a comparison with variability in broad line AGNs, the accretion disk inclination in narrow line AGNs is shown to be $\sim 85^o$, nearly edge-on. The ensemble variability of galaxies is confirmed with repeated imagings in the SDSS Southern Stripes.

There was no successful attempt in examining the galaxy classification from the point of view of variability. A recent work by Brunzendorf and Meusinger (2001) suggested variability is an efficient method to find narrow line AGN. However, this promising result was not subsequently confirmed by the authors (Meusinger and Brunzendorf 2002). The H II-liked galaxies are found in this work to vary with an average amplitude of ~ 1.7 larger than that of the narrow line AGNs. This may seem counter-intuitive, however a direct comparison is difficult because the underlying physical mechanisms driving the variability seen in both types of objects are likely different, as shown in the starburst modeling (§4.5.4) and by conventional wisdoms (from the Osterbrock diagrams and various spectral modelings).

The variability exhibits in the eClass BLUE type galaxies is larger than in the RED type galaxies, that can be partly understood by the fact that the star formation rate, indicated by the classification angle ϕ_{KL}, is larger in the BLUE galaxies. Starburst synthesis models alone are shown to be insufficient to explain the difference spectrum of the type D galaxies, implying an extra physical mechanism, probably other than starburst, in the bluer galaxies.

Our study also shows there is an variable spectral component in narrow line AGN. And that the AGN variables have higher mean level of ionization than the non-variables. These results put constraints on accretion disk modeling.

5.4   APPLICATIONS BY OTHER RESEARCHERS

The galaxy eigenspectra are being used in supernovae searching in the local universe (Krughoff et al. 2005).

By applying the QSO eigenspectra, York et al. (2005) have recently shown that the average extinction curve of the SDSS QSOs mimics the typical SMC (Small Magellanic Cloud) extinction curves, suggesting a possibility that our Galaxy is special in its characteristics of dust extinction.

The QSO eigenspectra are also used in decomposing broad line AGN spectra into AGN and host galaxy components (Vanden Berk et al. 2005), such that the relation between AGN and its host galaxy can be probed.

5.5   FUTURE WORK

Results from this thesis suggest that a useful classification should take account of global spectral space, which should be investigated in the future. Phenomenological models can be exploited to complement the spectral classification by the eigenspectra approach. A long term direction is to link the spectral diversity to various astrophysical problems and to study how that propagates to the physical parameters of the systems.

The morphological segregation of galaxies (chapter 1) makes us question whether different eClass types also cluster differently in the rich cluster environment and in elsewhere.

Boroson (2002) recently claimed that the diversities seen in several important QSO parameters (for example, the radio loudness, the strength of iron emissions, the QSO luminosity) are driven by the mass of the central supermassive black hole. It will be interesting to examine this suggestion by performing the analysis in the spectral space (i.e. the spectral KL transform) which does not require parameter identification.

An interesting project using the QSO eigenspectra is to relate the eigenspectra approach to the radio properties of the quasars, so that further discriminations of intrinsic and extrinsic properties can be achieved, for example, the orientation effects on the observed spectra (e.g., Richards et al. 2002b). In addition, the cross-projections can be applied to study future larger samples of quasars (e.g., $\sim 100,000$ at the completion of the SDSS) for possibly new evolution and luminosity effects.

We should include LINER (low ionization nuclear emission region, Heckman 1980a) in the variability analysis and made comparisons with variabilities in both the active and the inactive galaxies, that can give clues to the controversial nuclear activity of this type of object. Besides, we should perform further diagnostics on the variable and non-variable candidates in order to understand or discriminate their properties.

Long term monitoring of the narrow line AGN candidates defined in this work will be necessary to probe timescales in the structure function.

Another step is to explore the narrow emission line variability of AGNs, which are widely believed to be non-variable in time, commonly by the argument of a large light crossing time ($100 - 1,000$ years) in the typically low density medium of a narrow line region (Peterson et al. 1995). We can re-investigate this question and set a upper limit on the variability with the larger data sets in the SDSS or other surveys.

6.0   GLOSSARY

4000 Å break – defined by Gustavo Bruzual in 1983, it describes the discontinuity (a sharp opacity edge) in a galaxy spectrum in the vicinity of 4000 Å, which is due to the presence of a large number of spectral lines in a narrow wavelength region. The lines are usually due to ionized metals., 10

BADARC – a bad arc solution, 115

BADFLAT – low counts in fiberflat, 115

BADTRACE – bad trace from the routine TRACE20CRUDE, 115

Blue horizontal branch – population II stars (B3 – A0) in the galactic halo, characterized by strong, sharp hydrogen lines, large Balmer jumps and weak lines of all other elements (taken from the Extragalactic Articles)., 124

BRIGHTSKY – sky level is greater than observed flux plus ten times the observed flux error., 115

COMBINEREJ – the pixel is rejected in the combine B-spline, 115

Fiber magnitude – one of the magnitudes adopted by the SDSS, it is the magnitude of an object calculated from the flux in its central 3 arcsecond., 145

FULLREJECT – the pixel is fully rejected in extraction, 115

HII galaxies – inactive galaxies with emission lines resemble those always found in the HII regions., 126

Host galaxy – in this thesis it is the galaxy which harbors the QSO., 55

LOWFAT – the flat field is less than 0.5, 115

Moore's Law – an observation made by Gordon Moore in 1965, that the number of transitors per square inch on integrated circuits had doubled every year since the integrated circuit was invented., 6

NEARBADPIX – the bad pixel is within 3 pixels of trace, 115

NODATA – no data available in the combine B-spline, 115

Nuclear activity – usually referred to the activity that is due to the presence of black hole in the center of a galaxy, 2

Petrosian magnitude – the SDSS has adopted a modified form of the Petrosian (1976) system, measuring galaxy fluxes within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile (taken from the website www.sdss.org/dr4/)., 12

Post-starburst activity – or "K+A" activity, characterized by an enhancement of Balmer absorption lines and an decrease of Hα emission line., 26

PSF magnitude – a common magnitude for stars, it measures the magnitude calculated from the point-spread-function modeling of a stellar image, 145

Quasar – "quasi-stellar radio source". When it was discovered, it was referred to the center of a galaxy which luminosity far exceeds that of the underlying galaxy (a stellar appearance on the photographic plate as a results), and is radio loud. Subsequently majority of the quasars were found to be radio quiet, but the notation persisted. Nowadays, the word "quasar" is used interchangeably with the acronym "QSO", the quasi-stellar object., 1

RR Lyrae – one kind of periodic variable stars with period less than one day, and of spectral type A to F giant., 123

SCATLIGHT – scatter light is significant., 115

Seeing – usually given as the FWHM of the intensity of an stellar image in arcsecond, it describes the blurring of a point-source image due to the Earth's atmospheric turbulence., 117

Star formation rate (SFR) – the rate of star formation in a galaxy. It is usually expressed in units of M$_\odot$ yr^{-1}, 2

APPENDIX A

KARHUNEN-LOÈVE TRANSFORM

In astronomy, the applications of Karhunen-Loève transform (or Principal Component Analysis, PCA) to studying multivariate distributions have been discussed in detail (Efstathiou and Fall 1984; Murtagh and Heck 1987). The basic idea in applying the KL transforms in studying spectra is to derive from them a lower dimensional set of *eigenspectra* (Connolly et al. 1995), from which the essential physical properties are represented and hence a compression of data can be achieved. The eigenspectra are the principal components in the KL transform. Each spectrum can be thought of as an axis in a multidimensional hyperspace, $f_{\lambda_k i}$, which denotes the flux density per unit wavelength at the k-th wavelength in the i-th quasar spectrum.

For the moment, we assume that there are no gaps in each spectrum; we discuss the ways we deal with missing data in §2.2. From the set of spectra we construct the correlation matrix

$$C_{\lambda_k \lambda_l} = \hat{f}_{\lambda_k i} \hat{f}_{i \lambda_l} \; , \tag{A.1}$$

where the summation is from $i = 1$ to the total number of spectra, N, and $\hat{f}_{\lambda_k i}$ is the normalized i-th spectrum, defined for a given i as

$$\hat{f}_{\lambda_k} = \frac{f_{\lambda_k}}{\sqrt{\sum_{\lambda_k} f_{\lambda_k} f_{\lambda_k}}} \; . \tag{A.2}$$

The eigenspectra are obtained by finding a matrix, U, such that

$$U^T C U = \Lambda \; , \tag{A.3}$$

where Λ is the diagonal matrix containing the eigenvalues of the correlation matrix. U is thus a matrix whose i-th column consists of the i-th eigenspectrum $e_{i\lambda_k}$. We solve this eigenvalue problem by using Singular Value Decomposition.

The observed spectra are projected onto the eigenspectra to obtain the eigencoefficients. In these projections, every wavelength bin in each spectrum is weighted by the error associated with that particular wavelength bin, σ_λ, such that the weights are given by $w_\lambda = 1/\sigma_\lambda^2$. The observed spectra can be decomposed, with no error, as follows

$$f_{\lambda_k} = \sum_{i=1}^{M} a_i e_{i\lambda_k} \; , \tag{A.4}$$

where M is the total number of eigenspectra, and a_i are the expansion coefficients (or the *eigencoefficients*) of the i-th order. It is straightforward to see that, if the number of spectra is greater than the number of wavelength bins, M equals the total number of wavelength bins in the spectrum.

APPENDIX B

CONTINUUM ESTIMATION AND EQUIVALENT WIDTH CALCULATION

The continuum of the galaxy spectra at a given wavelength is estimated by the average flux density in a sliding window of 300 pixels ($\sim$ 21,000 km s^{-1}), with ±280 km s^{-1} centered around each emission line lie inside the window masked. A gaussian profile is fitted to each continuum-subtracted emission line by subroutines in the Common Mathematical Library SLATEC (version 4.1). Our subroutines are applied to the observed galaxy spectra in the SDSS and the emission line EWs are compared with the SDSS pipeline measurements. We find good agreement in the two EWs, with about 4% on average for each emission line in the Osterbrock diagrams. The emission line [O III]λ5008 is found less in agreement with the SDSS value, with an average of 9% difference. This difference is suspected to be due to the difference in the continua in the two, as such the continuum calculated in the SDSS pipeline have more small scale structures in the spectral region around the 4000 Å break and the [O III]λ5008 compared with us.

For the de-reddened and re-calibrated spectra in chapter 4, we find that their EW measurements do not differ substantially from those of the original observed spectra, with their difference within the discrepancy in the EW measurement by us and by the SDSS pipeline. This is partially expected because $EW = \int (f_\lambda^{spec} - f_\lambda^{cont})/f_\lambda^{cont}\,d\lambda$, where f_λ^{spec} is the spectral flux densities and f_λ^{cont} is the continuum flux densities, so even a multiplication of a λ-dependent constant to the spectral fluxes (as in the re-calibration process) do not alter the EW. However, strictly speaking, the above changes slightly the continuum estimation at a given pixel which uses fluxes within a spectral window, which in turn change the EW. In

this thesis, $M \sim 4,000$ for the galaxies, and $\sim 600 - 2,500$ for the QSOs.

BIBLIOGRAPHY

K. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, et al. The First Data Release of the Sloan Digital Sky Survey. *AJ*, 126:2081–2086, October 2003.

I. Aretxaga, R. Cid Fernandes, and R. J. Terlevich. QSO variability: probing the starburst model. *MNRAS*, 286:271–283, April 1997.

J. N. Bahcall, S. Kirhakos, D. H. Saxe, and D. P. Schneider. Hubble Space Telescope Images of a Sample of 20 Nearby Luminous Quasars. *ApJ*, 479:642–+, April 1997.

C. A. L. Bailer-Jones. Neural Network Classification of Stellar Spectra. *PASP*, 109:932–+, August 1997.

C. A. L. Bailer-Jones, M. Irwin, and T. von Hippel. Automated classification of stellar spectra - II. Two-dimensional classification with neural networks and principal components analysis. *MNRAS*, 298:361–377, August 1998.

I. K. Baldry, K. Glazebrook, C. M. Baugh, et al. The 2dF Galaxy Redshift Survey: Constraints on Cosmic Star Formation History from the Cosmic Spectrum. *ApJ*, 569:582–594, April 2002.

J. A. Baldwin. Luminosity Indicators in the Spectra of Quasi-Stellar Objects. *ApJ*, 214: 679–684, June 1977.

M. R. Blanton, H. Lin, R. H. Lupton, et al. An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey "Tiling" Algorithm. *AJ*, 125:2276–2286, April 2003.

G. Bono, D. Trevese, and M. Turatto. Synergies in Variability Studies of Stars, Supernovae, and Active Galactic Nuclei. *Memorie della Societa Astronomica Italiana*, 74:1004–+, 2003.

T. A. Boroson. Black Hole Mass and Eddington Ratio as Drivers for the Observable Properties of Radio-loud and Radio-quiet QSOs. *ApJ*, 565:78–85, January 2002.

T. A. Boroson and R. F. Green. The emission-line properties of low-redshift quasi-stellar objects. *ApJS*, 80:109–135, May 1992.

J. Brinchmann, S. Charlot, S. D. M. White, et al. The physical properties of star-forming galaxies in the low-redshift Universe. *MNRAS*, 351:1151–1179, July 2004.

B. C. Bromley, W. H. Press, H. Lin, and R. P. Kirshner. Spectral Classification and Luminosity Function of Galaxies in the Las Campanas Redshift Survey. *ApJ*, 505:25–36, September 1998.

J. Brunzendorf and H. Meusinger. A QSO survey via optical variability and zero proper motion in the M 92 field. I. QSO candidates and selection effects. *A&A*, 373:38–55, July 2001.

G. Bruzual and S. Charlot. Stellar population synthesis at the resolution of 2003. 344: 1000–1028, October 2003.

F. J. Castander, R. C. Nichol, A. Merrelli, et al. The First Hour of Extragalactic Data of the Sloan Digital Sky Survey Spectroscopic Commissioning: The Coma Cluster. *AJ*, 121: 2331–2357, May 2001.

A. J. Connolly, T. Budavári, A. S. Szalay, et al. An Orthogonal Approach to Photometric Redshifts. In *ASP Conf. Ser. 191: Photometric Redshifts and the Detection of High Redshift Galaxies*, pages 13–+, 1999.

A. J. Connolly and A. S. Szalay. A Robust Classification of Galaxy Spectra: Dealing with Noisy and Incomplete Data. *AJ*, 117:2052–2062, May 1999.

A. J. Connolly, A. S. Szalay, M. A. Bershady, et al. Spectral Classification of Galaxies: an Orthogonal Approach. *AJ*, 110:1071–+, September 1995.

A. Dressler. Galaxy Morphology in Rich Clusters: Implications for the Formation and Evolution of Galaxies. *ApJ*, 236:351–365, March 1980.

A. Dressler and J. E. Gunn. Spectroscopy of galaxies in distant clusters. II - The population of the 3C 295 cluster. *ApJ*, 270:7–19, July 1983.

G. Efstathiou and S. M. Fall. Multivariate analysis of elliptical galaxies. *MNRAS*, 206: 453–464, January 1984.

D. J. Eisenstein, D. W. Hogg, M. Fukugita, et al. Average Spectra of Massive Galaxies in the Sloan Digital Sky Survey. *ApJ*, 585:694–713, March 2003.

R. Everson and L. Sirovich. Karhunen-loève transform procedure for gappy data. *J.Opt.Soc.Am.A*, 12:1657, 1994.

X. Fan, M. A. Strauss, D. P. Schneider, et al. A Survey of z>5.7 Quasars in the Sloan Digital Sky Survey. II. Discovery of Three Additional Quasars at z>6. *AJ*, 125:1649–1659, April 2003.

M. Fioc and B. Rocca-Volmerange. PEGASE: a UV to NIR spectral evolution model of galaxies. *A&A*, 326:950–962, 1997.

S. Folkes, S. Ronen, I. Price, et al. The 2dF Galaxy Redshift Survey: spectral types and luminosity functions. *MNRAS*, 308:459–472, September 1999.

S. R. Folkes, O. Lahav, and S. J. Maddox. An artificial neural network approach to the classification of galaxy spectra. *MNRAS*, 283:651–665, December 1996.

P. J. Francis, P. C. Hewett, C. B. Foltz, and F. H. Chaffee. An objective classification scheme for QSO spectra. *ApJ*, 398:476–490, October 1992.

M. Fukugita, T. Ichikawa, J. E. Gunn, et al. The Sloan Digital Sky Survey Photometric System. *AJ*, 111:1748–+, April 1996.

P. L. Gómez, R. C. Nichol, C. J. Miller, et al. Galaxy Star Formation as a Function of Environment in the Early Data Release of the Sloan Digital Sky Survey. *ApJ*, 584:210–227, February 2003.

G. Galaz and V. de Lapparent. The ESO-Sculptor Survey: spectral classification of galaxies with $z > 0.5$. *A&A*, 332:459–478, April 1998.

P. J. Green, K. Forster, and J. Kuraszkiewicz. Quasar Evolution and the Baldwin Effect in the Large Bright Quasar Survey. *ApJ*, 556:727–737, August 2001.

M. Guainazzi, G. Matt, L. A. Antonelli, et al. The X-ray spectrum and variability of the Seyfert 2 galaxy NGC 7172. *MNRAS*, 298:824–830, August 1998.

J. E. Gunn, M. Carr, C. Rockosi, et al. The Sloan Digital Sky Survey Photometric Camera. *AJ*, 116:3040–3081, December 1998.

P. B. Hall, S. F. Anderson, M. A. Strauss, et al. Unusual Broad Absorption Line Quasars from the Sloan Digital Sky Survey. *ApJS*, 141:267–309, August 2002.

F. Hamann, M. Dietrich, B. Sabra, et al. Probing High Redshift Galaxy Evolution with Quasar Elemental Abundances (invited review). In *Astronomical Society of the Pacific Conference Series*, pages 236–+, 2003.

F. Hamann and G. Ferland. Elemental Abundances in Quasistellar Objects: Star Formation and Galactic Nuclear Evolution at High Redshifts. *ARA&A*, 37:487–531, 1999.

L. Hao, M. A. Strauss, C. A. Tremonti, et al. Active Galactic Nuclei in the Sloan Digital Sky Survey. I. Sample Selection. *AJ*, 129:1783–1794, April 2005.

M. R. S. Hawkins. Gravitational Microlensing Quasar Variability and Missing Matter. *Nature*, 366:242–+, November 1993.

A. F. Heavens, R. Jimenez, and O. Lahav. Massive lossless data compression and multiple parameter estimation from galaxy spectra. *MNRAS*, 317:965–972, October 2000.

T. M. Heckman. An optical and radio survey of the nuclei of bright galaxies - Activity in normal galactic nuclei. *A&A*, 87:152–164, July 1980a.

T. M. Heckman. The broad line region in active nuclei and quasars - Correlations with luminosity and radio emission. *A&A*, 88:311–316, August 1980b.

D. W. Hogg, D. P. Finkbeiner, D. J. Schlegel, and J. E. Gunn. A Photometricity and Extinction Monitor at the Apache Point Observatory. *AJ*, 122:2129–2138, October 2001.

A. M. Hopkins, C. J. Miller, R. C. Nichol, et al. Star Formation Rate Indicators in the Sloan Digital Sky Survey. *ApJ*, 599:971–991, December 2003.

E. P. Hubble. Extragalactic Nebulae. *ApJ*, 64:321–369, December 1926.

Ž. Ivezić, R. H. Lupton, M. Juric, et al. Quasar variability measurements with SDSS repeated imaging and POSS data. In *IAU Symposium*, pages 525–526, November 2004a.

Ž. Ivezić, R. H. Lupton, D. Schlegel, et al. SDSS data management and photometric quality assessment. *AN*, 325:583, 2004b.

M. Jang. Intranight optical variability in Seyfert 2 galaxies. *Ap&SS*, 275:209–215, February 2001.

G. Kauffmann, T. M. Heckman, S. D. M. White, et al. Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey. *MNRAS*, 341:33–53, May 2003.

T. Kawaguchi, S. Mineshige, M. Umemura, and E. L. Turner. Optical Variability in Active Galactic Nuclei: Starbursts or Disk Instabilities? *ApJ*, 504:671–+, September 1998.

R. C. Kennicutt. A spectrophotometric atlas of galaxies. *ApJS*, 79:255–284, April 1992.

L. J. Kewley, M. A. Dopita, R. S. Sutherland, et al. Theoretical Modeling of Starburst Galaxies. *ApJ*, 556:121–140, July 2001.

P. N. Kholopov, N. N. Samus, M. S. Frolov, et al. Combined General Catalogue of Variable Stars (Kholopov+ 1998). *VizieR Online Data Catalog*, 2214:0–+, April 1999.

C. S. Kochanek, M. A. Pahre, and E. E. Falco. Evidence for Inconsistencies in Galaxy Luminosity Functions Defined by Spectral Types. 0011458, November 2000.

J. Kormendy, L. Ho, and P. Murdin. Supermassive Black Holes in Inactive Galaxies. *Encyclopedia of Astronomy and Astrophysics*, October 2001.

J. H. Krolik. *Active galactic nuclei : from the central black hole to the galactic environment.* Princeton, N. J. : Princeton University Press, c1999., 1999.

K. S. Krughoff, A. J. Connolly, J. Frieman, et al. Detection of Type 1a Supernovae in Archival SDSS Spectra. *in prep.*, 2005.

O. Kuhn, M. Elvis, J. Bechtold, and R. Elston. A Search for Signatures of Quasar Evolution: Comparison of the Shapes of the Rest-Frame Optical/Ultraviolet Continua of Quasars at z>3 and z∼0.1. *ApJS*, 136:225–264, October 2001.

J. Kuraszkiewicz, B. J. Wilkes, W. N. Brandt, and M. Vestergaard. [O II] Emission, Eigenvector 1, and Orientation in Radio-quiet Quasars. *ApJ*, 542:631–643, October 2000.

A. Laor, J. N. Bahcall, B. T. Jannuzi, et al. The Ultraviolet Emission Properties of 13 Quasars. *ApJS*, 99:1–+, July 1995.

C. Leitherer, D. Schaerer, J. D. Goldader, et al. Starburst99: Synthesis Models for Galaxies with Active Star Formation. *ApJS*, 123:3–40, July 1999.

R. H. Lupton, J. E. Gunn, Ž. Ivezić, et al. The SDSS Imaging Pipelines. In *ASP Conf. Ser. 238: Astronomical Data Analysis Software and Systems X*, pages 269–+, 2001.

D. S. Madgwick, R. Somerville, O. Lahav, and R. Ellis. Galaxy spectral parametrization in the 2dF Galaxy Redshift Survey as a diagnostic of star formation history. *MNRAS*, 343: 871–879, August 2003.

M. A. Malkan and W. L. W. Sargent. The ultraviolet excess of Seyfert 1 galaxies and quasars. *ApJ*, 254:22–37, March 1982.

R. J. McLure and J. S. Dunlop. The cosmological evolution of quasar black hole masses. *MNRAS*, 352:1390–1404, August 2004.

R. J. McLure, J. S. Dunlop, and M. J. Kukula. Two-dimensional modelling of optical Hubble Space Telescope and infrared tip-tilt images of quasar host galaxies. *MNRAS*, 318:693–702, November 2000.

R. J. McLure, M. J. Kukula, J. S. Dunlop, et al. A comparative HST imaging study of the host galaxies of radio-quiet quasars, radio-loud quasars and radio galaxies - I. *MNRAS*, 308:377–404, September 1999.

E. Merzbacher. *Quantum Mechanics.* John Wiley&Sons, 2nd edition, 1970.

H. Meusinger and J. Brunzendorf. A QSO survey via optical variability and zero proper motion in the M 92 field. III. Narrow emission line galaxies. *A&A*, 390:439–448, August 2002.

C. J. Miller, R. C. Nichol, P. L. Gómez, et al. The Environment of Active Galactic Nuclei in the Sloan Digital Sky Survey. *ApJ*, 597:142–156, November 2003.

M. Mueller, G. M. Madejski, C. Done, and P. T. Zycki. The Power Density Spectrum of the Seyfert 2 Galaxy NGC 4945: The Relationship between X-ray Variability and Black Hole Mass. *AAS/High Energy Astrophysics Division*, 7:–+, March 2003.

F. Murtagh and A. Heck. *Multivariate data analysis.* Astrophysics and Space Science Library, Dordrecht: Reidel, 1987., 1987.

H. Netzer and B. J. Wills. Broad emission features in QSOs and active galactic nuclei. I - New calculations of Fe II line strengths. *ApJ*, 275:445–460, December 1983.

L. A. Nolan, J. S. Dunlop, M. J. Kukula, et al. The ages of quasar host galaxies. *MNRAS*, 323:308–330, May 2001.

S. C. Odewahn, R. R. de Carvalho, R. R. Gal, et al. The Digitized Second Palomar Observatory Sky Survey (DPOSS). III. Star-Galaxy Separation. *AJ*, 128:3092–3107, December 2004.

J. E. O'Donnell. R_ν-dependent optical and near-ultraviolet extinction. *ApJ*, 422:158–163, February 1994.

P. S. Osmer, A. C. Porter, and R. F. Green. Luminosity effects and the emission-line properties of quasars with $0 < z3.8$. *ApJ*, 436:678–695, December 1994.

D. E. Osterbrock. *Astrophysics of gaseous nebulae and active galactic nuclei.* Research supported by the University of California, John Simon Guggenheim Memorial Foundation, University of Minnesota, et al. Mill Valley, CA, University Science Books, 1989, 422 p., 1989.

E. A. Owens, R. E. Griffiths, and K. U. Ratnatunga. Using oblique decision trees for the morphological classification of galaxies. *MNRAS*, 281:153–157, July 1996.

P. G. Pérez-González, J. Zamorano, J. Gallego, et al. Spatial Analysis of the Hα Emission in the Local Star-forming UCM Galaxies. *ApJ*, 591:827–842, July 2003.

B. M. Peterson. *An introduction to active galactic nuclei.* Cambridge, New York Cambridge University Press, 1997 Physical description xvi, 238 p. ISBN 0521473489, 1997.

B. M. Peterson. Variability of Active Galactic Nuclei. In *Advanced Lectures on the Starburst-AGN*, pages 3–+, 2001.

B. M. Peterson, R. W. Pogge, I. Wanders, et al. Aperture Effects and Limitations on the Accuracy of Ground-Based Spectrophotometry of Active Galactic Nuclei. *PASP*, 107:579–+, June 1995.

B. M. Peterson, I. Wanders, R. Bertram, et al. Optical Continuum and Emission-Line Variability of Seyfert 1 Galaxies. *ApJ*, 501:82–+, July 1998.

J. R. Pier, J. A. Munn, R. B. Hindsley, et al. Astrometric Calibration of the Sloan Digital Sky Survey. *AJ*, 125:1559–1579, March 2003.

W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *AJ*, 187:425–438, 1974.

D. M. Rawson, J. Bailey, and P. J. Francis. Neural networks and the classification of active galactic nucleus spectra. *PASA*, 13:207, November 1996.

T. A. Reichard, G. T. Richards, D. P. Schneider, et al. A Catalog of Broad Absorption Line Quasars from the Sloan Digital Sky Survey Early Data Release. *AJ*, 125:1711–1728, April 2003.

C. Reichardt, R. Jimenez, and A. F. Heavens. Recovering physical parameters from galaxy spectra using MOPED. *MNRAS*, 327:849–867, November 2001.

G. T. Richards, X. Fan, H. J. Newberg, et al. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample. *AJ*, 123:2945–2975, June 2002a.

G. T. Richards, D. E. Vanden Berk, T. A. Reichard, et al. Broad Emission-Line Shifts in Quasars: An Orientation Measure for Radio-Quiet Quasars? *AJ*, 124:1–17, July 2002b.

P. M. Rodriguez-Pascual, D. Alloin, J. Clavel, et al. Steps toward Determination of the Size and Structure of the Broad-Line Region in Active Galactic Nuclei. IX. Ultraviolet Observations of Fairall 9. *ApJS*, 110:9–+, May 1997.

S. Ronen, A. Aragon-Salamanca, and O. Lahav. Principal component analysis of synthetic galaxy spectra. *MNRAS*, 303:284–296, February 1999.

R. R. Ross, A. C. Fabian, and S. Mineshige. The spectra of accretion discs in active galactic nuclei. *MNRAS*, 258:189–197, September 1992.

A. Sandage, M. Sandage, and J. Kristian. *Galaxies and the Universe.* Chicago, University of Chicago Press (Stars and Stellar Systems. Volume 9), 1975. 839 p, 1975.

D. J. Schlegel, D. P. Finkbeiner, and M. Davis. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. *ApJ*, 500:525–+, June 1998.

D. P. Schneider, X. Fan, P. B. Hall, et al. The Sloan Digital Sky Survey Quasar Catalog. II. First Data Release. *AJ*, 126:2579–2593, December 2003.

D. P. Schneider, P. B. Hall, T. R. Gordon, et al. The Sloan Digital Sky Survey Quasar Catalog. II. Third Data Release. 0503679, 2005.

D. P. Schneider, G. T. Richards, X. Fan, et al. The Sloan Digital Sky Survey Quasar Catalog. I. Early Data Release. *AJ*, 123:567–577, February 2002.

C. K. Seyfert. Nuclear Emission in Spiral Nebulae. *ApJ*, 97:28–+, January 1943.

Z. Shang, B. J. Wills, E. L. Robinson, et al. The Baldwin Effect and Black Hole Accretion: A Spectral Principal Component Analysis of a Complete Quasar Sample. *ApJ*, 586:52–71, March 2003.

H. P. Singh, R. K. Gulati, and R. Gupta. Stellar Spectral Classification using Principal Component Analysis and Artificial Neural Networks. *MNRAS*, 295:312–318, April 1998.

N. Slonim, R. Somerville, N. Tishby, and O. Lahav. Objective classification of galaxy spectra using the information bottleneck method. *MNRAS*, 323:270–284, May 2001.

J. A. Smith, D. L. Tucker, S. Kent, et al. The u'g'r'i'z' Standard-Star System. *AJ*, 123: 2121–2144, April 2002.

S. Snider, C. Allende Prieto, T. von Hippel, et al. Three-dimensional Spectral Classification of Low-Metallicity Stars Using Artificial Neural Networks. *ApJ*, 562:528, April 2001.

L. Sodre and H. Cuevas. Regularities in Galaxy Spectra and the Nature of the Hubble Sequence. In *ASP Conf. Ser. 114: Young Galaxies and QSO Absorption-Line Systems*, pages 149–+, 1997.

C. C. Steidel and W. L. W. Sargent. Emission-line and continuum properties of 92 bright QSOs - Luminosity dependence and differences between radio-selected and optically selected samples. *ApJ*, 382:433–465, December 1991.

C. Stoughton, R. H. Lupton, M. Bernardi, et al. Sloan Digital Sky Survey: Early Data Release. *AJ*, 123:485–548, January 2002.

M. A. Strauss, D. H. Weinberg, R. H. Lupton, et al. Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample. *AJ*, 124:1810–1824, September 2002.

J. W. Sulentic, P. Marziani, and D. Dultzin-Hacyan. Phenomenology of Broad Emission Lines in Active Galactic Nuclei. *ARA&A*, 38:521–571, 2000.

K. L. Thompson, G. J. Hill, and R. Elston. Lack of Iron Abundance Evolution in High-Redshift QSOS. *ApJ*, 515:487–496, April 1999.

B. M. Tinsley. Evolutionary synthesis of the stellar population in elliptical galaxies. II - Late M giants and composition effects. *ApJ*, 222:14–22, May 1978.

J. R. Trump, P. B. Hall, T. A. Reichard, et al. Broad Absorption Line Quasars from the Sloan Digital Sky Survey Second Data Release. *American Astronomical Society Meeting Abstracts*, 203:–+, December 2003.

D. Tytler and X. Fan. Systematic QSO emission-line velocity shifts and new unbiased redshifts. *ApJS*, 79:1–36, March 1992.

D. E. Vanden Berk, G. T. Richards, A. Bauer, et al. Composite Quasar Spectra from the Sloan Digital Sky Survey. *AJ*, 122:549–564, August 2001.

D. E. Vanden Berk, J. Shen, C. W. Yip, et al. Spectral Decomposition of Broad-Line AGNs and Host Galaxies. *SDSS Publication*, 601, 2005.

D. E. Vanden Berk, B. C. Wilhite, R. G. Kron, et al. The Ensemble Photometric Variability of ˜25,000 Quasars in the Sloan Digital Sky Survey. *ApJ*, 601:692–714, February 2004a.

D. E. Vanden Berk, C. W. Yip, A. J. Connolly, et al. Luminosity and Redshift Dependence of Quasar Spectral Properties. In *ASP Conf. Ser. 311: AGN Physics with the Sloan Digital Sky Survey*, pages 21–+, 2004b.

S. Vaughan, R. Edelson, R. S. Warwick, and P. Uttley. On characterizing the variability properties of X-ray light curves from active galaxies. *MNRAS*, 345:1271–1284, November 2003.

S. Veilleux, D.-C. Kim, D. B. Sanders, et al. Optical Spectroscopy of Luminous Infrared Galaxies. II. Analysis of the Nuclear and Long-Slit Data. *ApJS*, 98:171–+, May 1995.

S. Veilleux and D. E. Osterbrock. Spectral classification of emission-line galaxies. *ApJS*, 63: 295–310, February 1987.

M. Vestergaard. Early Growth and Efficient Accretion of Massive Black Holes at High Redshift. *ApJ*, 601:676–691, February 2004.

A. K. Vivas, R. Zinn, C. Abad, et al. The QUEST RR Lyrae Survey. I. The First Catalog. *AJ*, 127:1158–1175, February 2004.

T. von Hippel, L. J. Storrie-Lombardi, M. C. Storrie-Lombardi, and M. J. Irwin. Automated Classification of Stellar Spectra - Part One - Initial Results with Artificial Neural Networks. *MNRAS*, 269:97, July 1994.

F. Walter. Resolved Molecular Gas Emission in a QSO host galaxy at $z = 6.4$. In *ASSL Vol. 329: Starbursts: From 30 Doradus to Lyman Break Galaxies*, pages 327–+, May 2005.

I. Wanders, B. M. Peterson, R. W. Pogge, et al. Seeing and aperture effects on forbidden O III-based flux calibration of AGN spectra - NGC 3516: A case study. *A&A*, 266:72–76, December 1992.

W. B. Weaver and A. V. Torres-Dodgen. Accurate Two-dimensional Classification of Stellar Spectra with Artificial Neural Network. *ApJ*, 487:847, October 1997.

D. W. Weedman, F. R. Feldman, V. A. Balzano, et al. NGC 7714 - The prototype star-burst galactic nucleus. *ApJ*, 248:105–112, August 1981.

N. Weir, U. M. Fayyad, and S. Djorgovski. Automated Star/Galaxy Classification for Digitized Poss-II. *AJ*, 109:2401–+, June 1995.

B. C. Wilhite et al. Spectral Variability of Quasars in the Sloan Digital Sky Survey. II: The CIV Line. 461, 2005a.

B. C. Wilhite, D. E. Vanden Berk, R. G. Kron, et al. Spectral Variability of Quasars in the Sloan Digital Sky Survey. I: Wavelength Dependence. 0504309, 2005b.

B. J. Wills, H. Netzer, and D. Wills. Broad emission features in QSOs and active galactic nuclei. II - New observations and theory of Fe II and H I emission. *ApJ*, 288:94–116, January 1985.

B. J. Wills, Z. Shang, and J. M. Yuan. H$/\beta$ line width and the UV-X-ray spectra of luminous AGN. *New Astronomy Review*, 44:511–517, September 2000.

H. Winkler, I. S. Glass, F. van Wyk, et al. Variability studies of Seyfert galaxies. I - Broadband optical photometry. *MNRAS*, 257:659–676, August 1992.

C. W. Yip, A. J. Connolly, A. S. Szalay, et al. Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey. *AJ*, 128:585–609, August 2004a.

C. W. Yip, A. J. Connolly, D. E. Vanden Berk, et al. Spectral Classification of Quasars in the Sloan Digital Sky Survey: Eigenspectra, Redshift, and Luminosity Effects. *AJ*, 128: 2603–2630, December 2004b.

D. G. York, J. Adelman, J. E. Anderson, et al. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120:1579–1587, September 2000.

D. G. York, P. Khare, D. E. Vanden Berk, et al. Average Extinction Curves and Relative Abundances for QSO Absorption Line Systems at $1 < z_{abs} < 2$. *SDSS Publication*, 526, 2005.

W. Zheng and M. A. Malkan. Does a Luminosity-dependent Continuum Shape Cause the Baldwin Effect? *ApJ*, 415:517–+, October 1993.