# Galaxies in redshift surveys -

## spatial distributions and interstellar dust

*Vivienne Wild*

*Institute of Astronomy*
*and Clare Hall, Cambridge*

*September 9, 2005*

A thesis submitted to the University of Cambridge in fulfilment of the requirements for
the degree of Doctor of Philosophy

# DECLARATION

This Thesis is the result of my own work carried out in the Institute of Astronomy, Cambridge, between October 2002 and October 2005. No part of this thesis has been submitted for a degree, diploma or other qualification at this or any other university. The total length of this thesis does not exceed sixty thousand words.

Parts of the work presented here have been, or are due to be published, in refereed scientific journals:

- *"The 2dF Galaxy Redshift Survey: stochastic relative biasing between galaxy populations"*, Wild V., Peacock J. A., Lahav O. et al (The 2dFGRS Team), 2005, MNRAS, 356:247, (**Chapter 3**)
- *"Peering through the OH forest: a new technique to remove residual sky features from Sloan Digital Sky Survey spectra"*, Wild V. and Hewett P., 2005, MNRAS, 358:1083, (**Chapter 4**)
- *"Evidence for dust reddening in damped Ly$\alpha$ absorbers identified through Ca* II *H&K absorption"*, Wild V. and Hewett P., 2005, MNRAS, 361:L30, (**Chapter 5**)
- *"Selecting damped Lyman-$\alpha$ systems through Ca* II *absorption I: Dust depletions and reddening at $z \sim 1$"*, Wild V., Hewett P. and Pettini M., 2005, MNRAS submitted, (**Chapters 5 & 6**)

Each chapter has benefitted from collaboration with the authors listed above. However, except where explicitly indicated in the text, the work contained in the chapters is my own.

Vivienne Wild
*September 9, 2005*

*In memory of my Mum*
- *who would have enjoyed the read*

# SUMMARY

This thesis is concerned with analyses of two of the largest optical datasets in astronomy - the 2-degree field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS). Both surveys have collected the spectra of hundreds of thousands of extra-galactic objects, but their different instrumentation and survey designs have made them particularly suited to different tasks.

The 2dFGRS was the first survey to be completed, its primary scientific goal was to quantify the distribution of galaxies over the largest scales ever probed. This thesis begins by following the same theme, comparing the relative density fields of blue and red galaxies in the 2dFGRS. In almost all cosmological results derived from redshift surveys, the relation between the galaxy and underlying matter density fields (galaxy bias) is assumed to be linear, an approximation that is unlikely to be true in detail. This thesis investigates the potential magnitude of any non-linearities by comparing the distributions of different types of galaxies. The bias between red and blue galaxies is found to deviate significantly from a linear relation and, furthermore, a considerable scatter is found in the joint distribution, over and above that expected from Poisson statistics. This “stochasticity” may constrain galaxy formation and/or evolution scenarios in future generations of simulations of the distribution of galaxies in the Universe.

The 2dFGRS contributed substantially to our understanding of the structure of the local Universe, however the spectra had limited further applications after providing redshifts. This is the real benefit of the SDSS dataset. Although not yet completed, the SDSS is already the most extensive survey of the local Universe obtained to date. One of the successes of the spectroscopic survey has been the automatic reduction pipelines which have achieved high quality and uniform data processing. However, despite vast improvement over previous surveys, problems remain, in particular with subtraction of night sky lines in the red half of the spectra. In this thesis a new technique is developed to remove the sky residuals using a principal component analysis (PCA) to reconstruct the signal. The technique substantially reduces the systematic noise in the red half of the majority of spectra, providing potential benefits, for example, for finding weak line features and creating high signal-to-noise ratio composite spectra.

The thesis then moves on from improving the quality of SDSS spectra to using the vast database for science applications. The quasar catalogue allows the detection of a large number of “quasar absorption line systems”, caused by the interstellar medium of galaxies intervening the line-of-sight between the quasar and us. Among these systems are a small number which exhibit Ca II absorption, an interstellar metal substantially neglected till now in high-redshift galaxy studies. In contrast to Damped Lyman-$\alpha$ systems (DLAs) at similar and higher redshifts, the Ca II absorbers significantly redden the background quasar spectral energy distributions suggesting they contain substantial quantities of dust. In order to improve the statistics of this reddening analysis, the technique of PCA is used to remove unusual quasars from the sample. The presence of considerable quantities of both dust and metals is highly suggestive that the Ca II absorbers have hydrogen column densities greater than the limit of DLAs. This in turn implies cold, neutral gas and suggests that Ca II absorbers may provide the missing link between chemically unevolved, dust-free DLAs selected by gas cross section, and the more metal rich, dustier and star-forming emission–selected galaxies at similar redshifts.

# CONTENTS

**List of Figures** **x**

**List of Tables** **xiii**

**1 Introduction** **1**

1.1 Galaxies, quasars, and their spectra . . . 1

1.1.1 The physics of spectra . . . 2

1.1.2 Active Galactic Nuclei . . . 3

1.1.3 Galaxy absorption spectra . . . 4

1.1.4 Galaxy evolution . . . 7

1.2 Redshift surveys . . . 7

1.2.1 Deep vs. wide-field observations . . . 8

1.2.2 A history of wide-field redshift surveys . . . 8

1.3 The large scale structure of the Universe . . . 9

1.3.1 Observations . . . 10

1.3.2 Theory of galaxy formation . . . 10

1.3.3 The cosmological concordance model . . . 11

1.3.4 Galaxy formation and evolution through simulations . . . 11

1.4 Thesis Motivation and Overview . . . 11

**2 Analysis of spectroscopic datasets** **15**

2.1 The 2dFGRS . . . 17

2.1.1 Survey design . . . 17

2.1.2 Scientific highlights . . . 19

2.2 The SDSS . . . 21

2.2.1 Survey design . . . 21

2.2.2 Scientific results . . . 24

2.3 Principal Component Analysis . . . 25

2.3.1 PCA in Astronomy . . . 26

2.3.2 The mathematics of PCA . . . 27

**3 Relative Biasing between Galaxy Populations** **33**

3.1 Introduction . . . 33

3.1.1 Describing bias . . . 34

3.1.2 Linking mass and light . . . 35

3.1.3 Measurements of relative biasing . . . 36

3.2 Modelling Relative Bias . . . 37

3.2.1 A framework for nonlinear, stochastic bias . . . 38
3.2.2 Redshift-space distortions . . . 39
3.2.3 One point probability distribution function . . . 40
3.2.4 Biasing models . . . 41
3.2.5 Including observational shot noise . . . 43
3.3 The Data: the 2dFGRS . . . 44
3.3.1 Galaxy properties in the 2dFGRS . . . 44
3.3.2 The volume limited galaxy sample . . . 46
3.4 Method: Counts-in-Cells . . . 49
3.5 Parameter fitting . . . 52
3.5.1 A maximum likelihood approach . . . 52
3.5.2 Error estimation . . . 53
3.5.3 Model comparison . . . 54
3.6 Results . . . 55
3.6.1 One point probability distributions . . . 56
3.6.2 Joint distributions and biasing models . . . 60
3.6.3 Origin of the stochasticity signal . . . 72
3.6.4 Consistency checks . . . 74
3.6.5 Direct variance estimates . . . 75
3.7 Comparison with simulations . . . 77
3.7.1 Previous work . . . 77
3.7.2 Preliminary mock comparison . . . 78
3.8 Summary and Conclusions . . . 79

**4 PCA sky-residual subtraction of SDSS spectra 81**
4.1 Introduction . . . 81
4.1.1 Sky light . . . 82
4.1.2 Sky subtraction methods . . . 82
4.1.3 Undersampling of OH lines . . . 83
4.1.4 Sky in the SDSS spectra . . . 84
4.2 Preparing the sky spectra . . . 85
4.2.1 A sample of sky spectra . . . 86
4.2.2 Poisson error normalisation . . . 87
4.2.3 Identification of sky pixels . . . 93
4.3 Creating the principal components . . . 94
4.3.1 The number of components . . . 94
4.3.2 Reconstructing sky residuals in an object spectrum . . . 97
4.4 Application of method to SDSS spectra . . . 99
4.4.1 Galaxies . . . 101
4.4.2 Quasars . . . 101
4.4.3 Stars . . . 101
4.5 Tests of method on SDSS science objects . . . 101

4.5.1 Galaxy absorption features: The Ca II triplet . . . . . . . . . . . . . . . . . . . . 105
4.5.2 Absorption features in damped Lyman-$\alpha$ systems . . . . . . . . . . . . . . . . . . . . 108
4.5.3 High redshift quasar composites . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
4.6 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111

**5 Dust in quasar absorption line systems 113**
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
5.1.1 Quasar absorption line and Damped Lyman-$\alpha$ systems . . . . . . . . . . . . . . . 114
5.1.2 Finding intermediate redshift DLAs . . . . . . . . . . . . . . . . . . . . . . . . 114
5.1.3 Dust in DLAs and obscuration bias . . . . . . . . . . . . . . . . . . . . . . . . 115
5.1.4 The aim of this chapter . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
5.2 Identifying Broad Absorption Line Quasars . . . . . . . . . . . . . . . . . . . . . . 116
5.2.1 Details of the method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117
5.3 A sample of Ca II and Mg II absorption systems . . . . . . . . . . . . . . . . . . . . 119
5.3.1 Ca II systems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
5.3.2 Mg II systems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123
5.3.3 Trends of strong metal line strengths . . . . . . . . . . . . . . . . . . . . . . . 123
5.4 Selection function and total redshift path . . . . . . . . . . . . . . . . . . . . . . 124
5.5 Measuring the average reddening . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
5.5.1 Quasar reference spectra and dust extinction curves . . . . . . . . . . . . . . . . 126
5.5.2 The significance of detection . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128
5.6 Average dust content of metal absorption line systems . . . . . . . . . . . . . . . . 129
5.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 132
5.7.1 Dust obscuration bias . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
5.7.2 The number density of Ca II absorbers and DLAs . . . . . . . . . . . . . . . . . 135
5.7.3 Constraining N(H I) through dust-to-gas ratios . . . . . . . . . . . . . . . . . . 136
5.7.4 Dust in intermediate redshift galaxies . . . . . . . . . . . . . . . . . . . . . . . 136
5.8 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137

**6 The element depletion pattern of Ca II absorption systems 139**
6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
6.1.1 The effect of dust on metal abundances . . . . . . . . . . . . . . . . . . . . . . 140
6.1.2 Ca II in the local Universe . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 140
6.2 Method: obtaining relative metal abundances . . . . . . . . . . . . . . . . . . . . . 141
6.2.1 Creating composite spectra . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 143
6.2.2 Column densities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 143
6.3 Results: relative abundances and dust depletions . . . . . . . . . . . . . . . . . . . 145
6.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 147
6.4.1 Dust-to-metals ratio . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 147
6.4.2 Constraints on $N$(H I) from Zn II . . . . . . . . . . . . . . . . . . . . . . . . . 149
6.5 Summary and Conclusions: reddening and dust depletions . . . . . . . . . . . . . . . 149

**7 Further work 153**

7.1 The nature of Ca II absorbers . . . . . 153
7.1.1 Follow up observations . . . . . 154
7.1.2 Diffuse interstellar bands . . . . . 155
7.2 Emission line galaxy survey . . . . . 156
7.2.1 Star formation rates . . . . . 156
7.2.2 Previous and ongoing surveys . . . . . 157
7.2.3 The SDSS blind emission line survey . . . . . 157
7.2.4 Results so far . . . . . 158

**A Air and vacuum wavelengths 163**

**Bibliography 165**

**Acknowledgments 173**

# LIST OF FIGURES

1.1 Example spectra of an O– and a G–type star. . . . . . . . . . . . . . . . . . . . . . . . 3
1.2 Example spectra of a star forming and passive galaxy . . . . . . . . . . . . . . . . . . 4
1.3 Example spectra of an ordinary and a Broad Absorption Line quasar. . . . . . . . . . . 5
1.4 What causes a quasar absorption line system? . . . . . . . . . . . . . . . . . . . . . . 6
1.5 The large scale distribution of galaxies: the CfA stickman . . . . . . . . . . . . . . . . 9

2.1 Examples of spectra from galaxies targeted by both the 2dFGRS and SDSS . . . . . . . 16
2.2 The 2dF instrument on the AAT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
2.3 The 2dFGRS survey area . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
2.4 The large scale structure of the Universe as observed by the 2dFGRS . . . . . . . . . . 19
2.5 Using redshift surveys for cosmology: simulated galaxy distributions . . . . . . . . . . 20
2.6 The SDSS survey area . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

3.1 The distributions of spectral type and rest frame colour for all 2dFGRS galaxies . . . . . 47
3.2 The joint distribution of colour and spectral types for galaxies in the 2dFGRS . . . . . . 48
3.3 Creating volume limited samples of galaxies . . . . . . . . . . . . . . . . . . . . . . . 48
3.4 The spatial distribution of a volume limited sample of 2dFGRS galaxies with typical cell boundaries overplotted . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
3.5 The completeness distributions of cells before and after spectral classification of galaxies 52
3.6 The bivariate distributions and 1D projections of the overdensities of red and blue galaxies in cells . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
3.7 Comparison of the univariate distributions of early type galaxies with those expected from the best-fitting Poisson sampled lognormal model . . . . . . . . . . . . . . . . . 58
3.8 Comparison of the bivariate distributions of best-fitting linear, power law and bivariate lognormal models with the data for galaxies classified by colour . . . . . . . . . . . . 61
3.9 Same as Fig. 3.8, with galaxies classified by $\eta$ type . . . . . . . . . . . . . . . . . . . . 62
3.10 The bivariate counts-in-cells distributions for cells of different scale lengths . . . . . . . 68
3.11 The scale dependence of the nonlinearity and stochasticity of relative bias . . . . . . . . 69
3.12 The scale dependence of the ratio of variances and linear correlation coefficient . . . . . 70
3.13 The spatial distribution of cells identified as causing the stochasticity signal . . . . . . . 73
3.14 The colour distribution of galaxies in cells identified as causing the stochasticity signal . 74

4.1 A spectrum of night sky emission beyond 6700Å . . . . . . . . . . . . . . . . . . . . . 83
4.2 Examples of residual sky features in SDSS sky spectra after the pipeline sky subtraction procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 85

4.3 Flux rms of the sky spectra before and after normalisation by the SDSS error arrays . . . 87

4.4 Analysis of raw data from a spectroscopic plate . . . . . . . . . . . . . . . . . . . . . . . . . . 89

4.5 Example of the median noise of all spectra on an SDSS plate . . . . . . . . . . . . . . . . 90

4.6 The flux rms of sky spectra before and after the new sky-residual subtraction is applied . 91

4.7 The flux rms of the sky spectra with flux normalised by a scaled version of the SDSS error arrays . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.8 The first five sky principal components . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95

4.9 The 6th-10th sky principal components . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

4.10 The flux rms in sky and non-sky pixels as a function of the number of PCA components . 97

4.11 The total number of PCA components used during reconstruction of the residual sky features in sky, galaxy and quasar spectra . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

4.12 An example of the process of PCA sky-residual subtraction on a galaxy spectrum . . . . 100

4.13 Examples of sky-residual subtraction applied to galaxy spectra . . . . . . . . . . . . . . . 102

4.14 Same as Fig. 4.13 for quasar spectra . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103

4.15 Same as Fig. 4.13 for stellar spectra . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

4.16 Example SDSS spectra in the region of the Ca II triplet before and after sky-residual subtraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106

4.17 Results of a matched filter search for the Ca II triplet in spectra before and after sky-residual subtraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108

4.18 Composite spectrum of quasars containing Damped Lyman-$\alpha$ systems, before and after sky-residual subtraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109

4.19 Composite spectra of quasars containing Damped Lyman-$\alpha$ systems in the region of the strong Mg absorption lines, before and after sky-residual subtraction . . . . . . . . . . . 109

4.20 Composite spectra of $z \sim 4$ quasars created from spectra with and without subtracting sky residuals . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111

5.1 Examples of quasars with and without broad absorption troughs, with PCA reconstructions of the spectra overplotted . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118

5.2 Examples of Ca II systems found in the SDSS quasar sample with their associated Mg II and Fe II absorption lines . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120

5.3 The equivalent width of Mg II versus Fe II and Mg I for the Mg II and Ca II absorbers . . 123

5.4 The probability of detecting a Ca II absorption system in as a function of absorption redshift and equivalent width . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

5.5 The total redshift path of the Ca II survey calculated from Monte Carlo trials . . . . . . . 125

5.6 Equivalent width distributions and number densities of the Mg II and Ca II absorbers . . 127

5.7 The distribution of apparent *E(B−V)* values for composites created from randomly chosen samples of quasar spectra . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 129

5.8 Composite spectrum of the 37 quasars with Ca II absorbers, after division by an ‘average’ quasar SED . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

5.9 As for Fig. 5.8 but for Mg II-selected DLA candidates . . . . . . . . . . . . . . . . . . . . . 131

5.10 Colour excess caused by individual Ca II absorbers . . . . . . . . . . . . . . . . . . . . . . 132

5.11 The effect on the total survey pathlength, were each quasar line of sight to be intercepted by an absorber with the average properties of the Ca II sample . . . . . . . . . . . . . . 133

5.12 The observed frame ($i$-band) extinction of quasar light caused by LMC type dust in an intervening galaxy as a function of $z_{\rm abs}$ . . . . . . . . . . . . . . . . . . . . . . . . . . 135

5.13 The distribution of *E(B−V)* for UV–selected galaxies at $z \sim 2$ . . . . . . . . . . . . . . 137

6.1 The composite spectrum of the 27 Ca II absorption systems . . . . . . . . . . . . . . . . 142

6.2 The abundances of refractory elements relative to Zn in DLAs and the Ca II absorbers . . 146

6.3 The abundances of refractory elements relative to Zn in the warm and neutral ISM of the Milky Way compared to the Ca II absorbers . . . . . . . . . . . . . . . . . . . . . . . 147

7.1 SDSS and INT image of a quasar with an intervening $z$=0.1 Ca II absorption line system 154

7.2 High resolution Keck spectrum in the region of the Ca II and Na I lines . . . . . . . . . . 155

7.3 Composite spectrum of ELGs detected as secondary objects in the SDSS spectroscopic catalogue . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 159

7.4 Comparison with ELGs in the primary SDSS catalogue . . . . . . . . . . . . . . . . . . 160

7.5 Average trends of line luminosities in secondary ELGs found in the spectra of SDSS galaxies . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 161

7.6 $\mathrm{R}_{23}$ and BPT diagram for the secondary emission line objects . . . . . . . . . . . . . . . 161

# List of Tables

2.1 Numbers of early/late and red/blue galaxies in the 2dFGRS catalogue . . . . . . . . . . 17

2.2 Details of the SDSS data releases . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

3.1 Completeness limits, total number of cells and average cell counts . . . . . . . . . . . . 56

3.2 The best-fitting deterministic biasing model parameters for galaxies classified by colour . 63

3.3 Same as Table 3.2 except for galaxies split by $\eta$-type . . . . . . . . . . . . . . . . . . . 64

3.4 The best-fitting bivariate lognormal model parameters . . . . . . . . . . . . . . . . . . . 65

3.5 Dependence of stochastic relative bias on redshift . . . . . . . . . . . . . . . . . . . . . . 71

3.6 Different variance estimates for cell overdensity distributions . . . . . . . . . . . . . . . 76

4.1 The dependence of the effectiveness of the sky-residual subtraction on different scalings of the SDSS noise arrays . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93

4.2 Equivalent widths of metal absorption features in quasar spectra with Damped Lyman-$\alpha$ systems, before and after sky-residual subtraction . . . . . . . . . . . . . . . . . . . . . 110

5.1 Details of each Ca II absorption line system . . . . . . . . . . . . . . . . . . . . . . . . . 121

5.2 Continuation of Table 5.1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

5.3 The average reddening introduced in quasar SEDs by the Ca II absorbers and Mg II–selected DLAs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130

6.1 The average reddening introduced in quasar SEDs by the Ca II absorbers with $z_{\rm abs} > 0.88$ 140

6.2 Average equivalent widths of metal lines in the Ca II absorbers . . . . . . . . . . . . . . 141

6.3 Ion column densities and abundances relative to Zn . . . . . . . . . . . . . . . . . . . . 144

6.4 Gas-to-dust ratios, metal abundances and derived dust-to-metals ratios of the Milky Way, LMC and SMC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 148

7.1 The number of secondary emission line galaxies with $z < 0.4$ identified in the SDSS DR4 spectroscopic catalogues . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 158

A.1 Air and vacuum wavelengths of transitions commonly seen in nebula emission . . . . . . 163

A.2 Air and vacuum wavelengths of transitions commonly seen in absorption . . . . . . . . . 164

*The sun and you me and all the stars*
*that we can see, are moving at a*
*million miles a day In an outer spiral*
*arm, at 40,000 miles an hour of the*
*Galaxy we call the Milky Way*

*Our Galaxy itself contains 100 billion stars*
*It's 100,000 light years side to side,*
*It bulges in the middle, 16,000 light*
*years thick but out by us it's just 3,000*
*light years wide*

*We're 30,000 light years from Galactic*
*central point we go round every 200*
*million years and our Galaxy is only*
*one of millions or billions in this*
*amazing and expanding universe*

From "The Galaxy Song", Monty Python

# 1 Introduction

Each of the principal chapters of this thesis begin with detailed, referenced discussions of the background material relevant to the work presented within them. This introductory chapter has been written for the purpose of providing general scientists with a history and non technical overview of the topics covered by the thesis. It also aims to draw together the different areas of work presented in subsequent chapters and place them within a common context of applications for spectra from redshift surveys. The chapter begins by introducing some of the physics behind astronomical spectra, giving examples of different optical spectra encountered in astronomy. It goes on to describe the history of wide-field spectroscopic galaxy and quasar surveys - the primary source of spectra used in this thesis. Finally, the chapter introduces the topic of the large scale distribution of galaxies in the Universe, a major application for spectra obtained through wide-field surveys. A detailed description of the two surveys used in this thesis is then provided in Chapter 2.

## 1.1 Galaxies, quasars, and their spectra

Fuzzy patches of light on the sky have been observed by humans for hundreds of years. However, the exact nature of these "nebulae" remained unclear until the early part of the 20th century. From observations made by Edwin Hubble they were confirmed to be systems of stars lying far outside the extent of our Galaxy (Hubble 1926) and this finding represented a fundamental turning point in the understanding of mankind's place in the Universe. Similar to when Galileo toppled the Earth from its pinnacle at the centre of the Universe in the 1600's by observations that proved it to be orbiting the Sun, so the confirmation of the "island universe" theory, in which the nebulae are entire star systems beyond the Milky Way, dramatically changed our perceptions of the Universe in which we live.

The Milky Way, the galaxy in which our Sun and the solar system resides, contains roughly 100 billion stars, together with reservoirs of gas and dust, all of which can be observed by the light they emit and absorb. By their gravitational effects on the stars and gas, we also believe that the centre of the Galaxy harbours a massive black hole and the majority of the matter that makes up the galaxy is "dark"– does not emit or absorb light. All these constituents are held together simply by the force of gravity. Within the Universe there are billions of similar such galaxies: gravitationally bound systems containing stars, gas, dust, dark matter and often central black holes. Galaxies range vastly in size, mass and exact composition. They are not static in time: they grow old as they use up the fuel needed to form stars; they become more metal rich[1] as the succeeding generations of stars explode, releasing the metals they have created through nuclear fusion during their lifetimes; they respond to external forces such as collisions with other galaxies.

Uncovering the story of how the galaxies form and evolve over the age of the Universe is currently one of the most active research areas in modern astrophysics. Why is one galaxy rapidly forming stars, causing it to appear blue and presenting beautiful spiral arms, whilst a neighbour has run out of gas and its stars are simply growing older, redder and cooler? Were they formed at different times? Were their initial constituents fundamentally different? Were they influenced by the properties of the environment in which they lived, for example how crowded it was? These are some examples of the many interweaving questions we have about the physical processes involved in the formation and evolution of galaxies. As with many large problems in science, different lines of enquiry will slowly fill in the gaps in our understanding.

### 1.1.1 The physics of spectra

Astrophysics is unique as a science in its almost total reliance on a single method of data collection - that of observing the light emitted by collections of atoms and molecules. For all objects except those in the very local neighbourhood (planets and comets) the distances involved are too vast to contemplate sending probes to investigate them first hand: 28 years since the spacecraft Voyager 1 was launched it recently reached the boundaries of our solar system, and it would take a probe four years traveling at the speed of light to reach the nearest known star to the Sun. The development of astronomy into the branch of physics known as astrophysics in the main came about through the first measurements of the spectra of astronomical objects (e.g. nebulae by William Huggins in the 1860s). By drawing upon our knowledge of the physics behind how atoms work and how photons and matter interact, we can infer substantial amounts of information about astronomical objects. All that we know about the Universe beyond the solar system is derived from the simple collection of photons.

Fig. 1.1 shows example spectra, in the optical waveband, of two types of star. The overall shape ("colour") of spectra and the strengths of the absorption features are determined by the physical properties of the star such as mass, temperature, surface gravity and abundance of metals. For galaxies, several constituents contribute light to a galaxy's spectrum: stars of different types; gas clouds composed of mainly hydrogen (H) gas with temperatures varying from tens to millions of Kelvin; and dust grains

[1] In astronomy any element heavier than Helium is called a "metal".

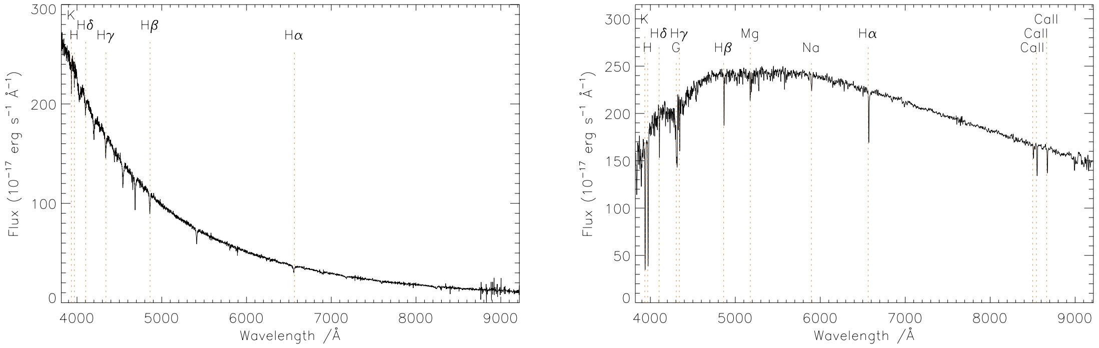

**Figure 1.1:** Example spectra of Left: an O–type star: a strong ultra-violet continuum and relatively weak hydrogen lines. Right: A G–type star (the Sun is a G–type star): strong Calcium II (Ca II) lines (K and H are also Ca II transitions) and G–band (CH molecule); neutral metals such as Sodium (Na) and Magnesium (Mg) are also present.

made up of heavy elements such as iron, silicon and carbon. These latter two make up the "interstellar medium" (ISM) of galaxies and all three components must be accounted for when analysing an observed galaxy spectrum. Fig. 1.2 shows two examples of galaxy spectra: from a blue, star forming, spiral galaxy and an old, red, elliptical galaxy. Alongside are example images of these types of galaxies.

The intense UV radiation of forming and newly formed massive stars ionises[2] nearby gas in the galaxy; these hot, ionised gas clouds emit light strongly at certain well-defined wavelengths as re-captured electrons cascade down the energy levels of atoms, leading to so-called "nebula emission lines". For example, in the top panel in Fig. 1.2 lines corresponding to transitions in hydrogen, oxygen, nitrogen and sulphur atoms can be seen. Cold, neutral atomic and molecular hydrogen gas is the most massive component of the ISM of galaxies. Cold atomic gas can generally only be observed in absorption, when the hydrogen and trace metals in the gas capture photons at wavelengths corresponding to their atomic transitions. A background light source is needed, which could be O and B stars (the hottest types of stars) in our own and very local galaxies, or light from a quasar (see Section 1.1.3). Atomic hydrogen is also visible in emission through H 21cm radiation. Molecular hydrogen observations are more difficult and carbon monoxide is often relied upon as a tracer of molecular gas.

The presence of dust in galaxies has three main effects on the spectra we observe: it obscures a significant proportion of light; causes an overall reddening of the spectrum by absorbing light across a wide range of optical wavelengths and re-emitting the light at longer (red or infra-red) wavelengths; and selectively incorporates different elements onto the dust grains, taking the atoms out of the gas-phase in which they are visible through their interactions with photons.

### 1.1.2 Active Galactic Nuclei

At the centres of some massive galaxies reside "active nuclei" which can be more than 100 times as luminous as all the stars in the galaxy combined. The radiation is caused by accretion of matter onto supermassive black holes and these objects appear in many different guises: in the optical waveband as quasars and Seyfert galaxies, as radio galaxies and as X-ray sources. A typical optical spectrum of a quasar is shown in Fig. 1.3. Their spectra are dominated by the high energy physical processes occuring

[2] Electrons are stripped away from the atoms.

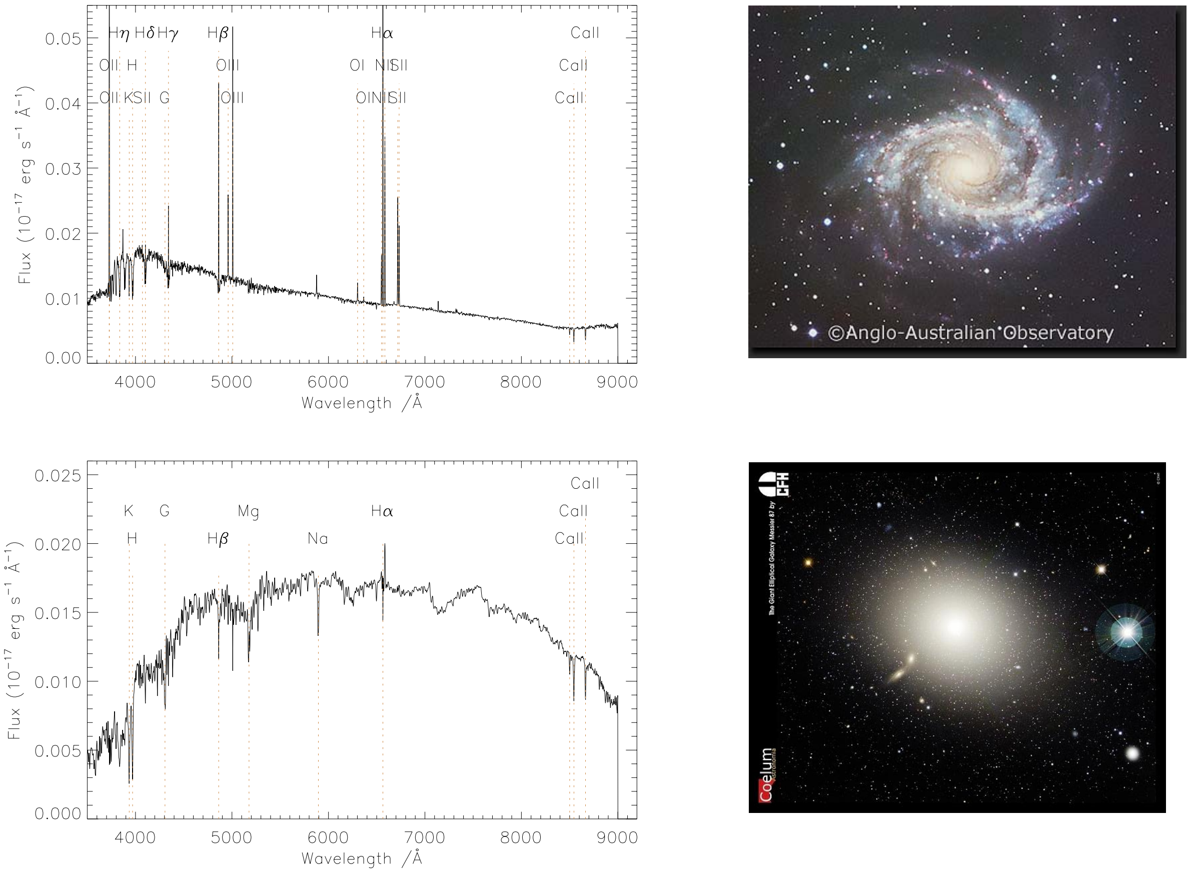

**Figure 1.2:** Top: The spectrum of a typical star forming galaxy. The star light is dominated by young, massive O– and B–type stars. Strong nebula emission lines indicate the presence of hot ionised gas. On the right is an optical image of a galaxy which would have this type of spectrum. Bottom: The spectrum of a typical passive galaxy. The star light is dominated by older, less massive stars. On the right is the image of M87, a typical elliptical galaxy that would have this type of spectrum.

in the centre of the galaxies, for example, the broad emission lines are caused by fast moving gas orbiting the central black hole. The physical processes going on in quasars which lead to their distinctive spectra is an active research area. The lower panel of Fig. 1.3 is a spectrum of a Broad Absorption Line (BAL) quasar, an interesting class which makes up around 15% of all optically identified quasars. The origin of the broad absorption troughs on the blue side of the emission lines is little understood.

Due to their phenomenal brightness the spectra of quasars can be observed from quasars at very great distances, allowing us to study the physics of the early Universe. The Sloan Digital Sky Survey (Section 2.2) has allowed the identification of 12 quasars with redshifts ($z$) above 5.7, which is equivalent to saying the Universe was only $\sim$1 Gyr old[3].

### 1.1.3 Galaxy absorption spectra

One method by which we can learn about the properties of galaxies over a large fraction of the age of the Universe is through observing their “absorption spectra”. Light from a bright source is absorbed by gas and dust in the galaxy’s ISM and the absorption lines allow the accurate measure of elements that make

[3] The Universe is thought to be $\sim$14 Gyr old today.

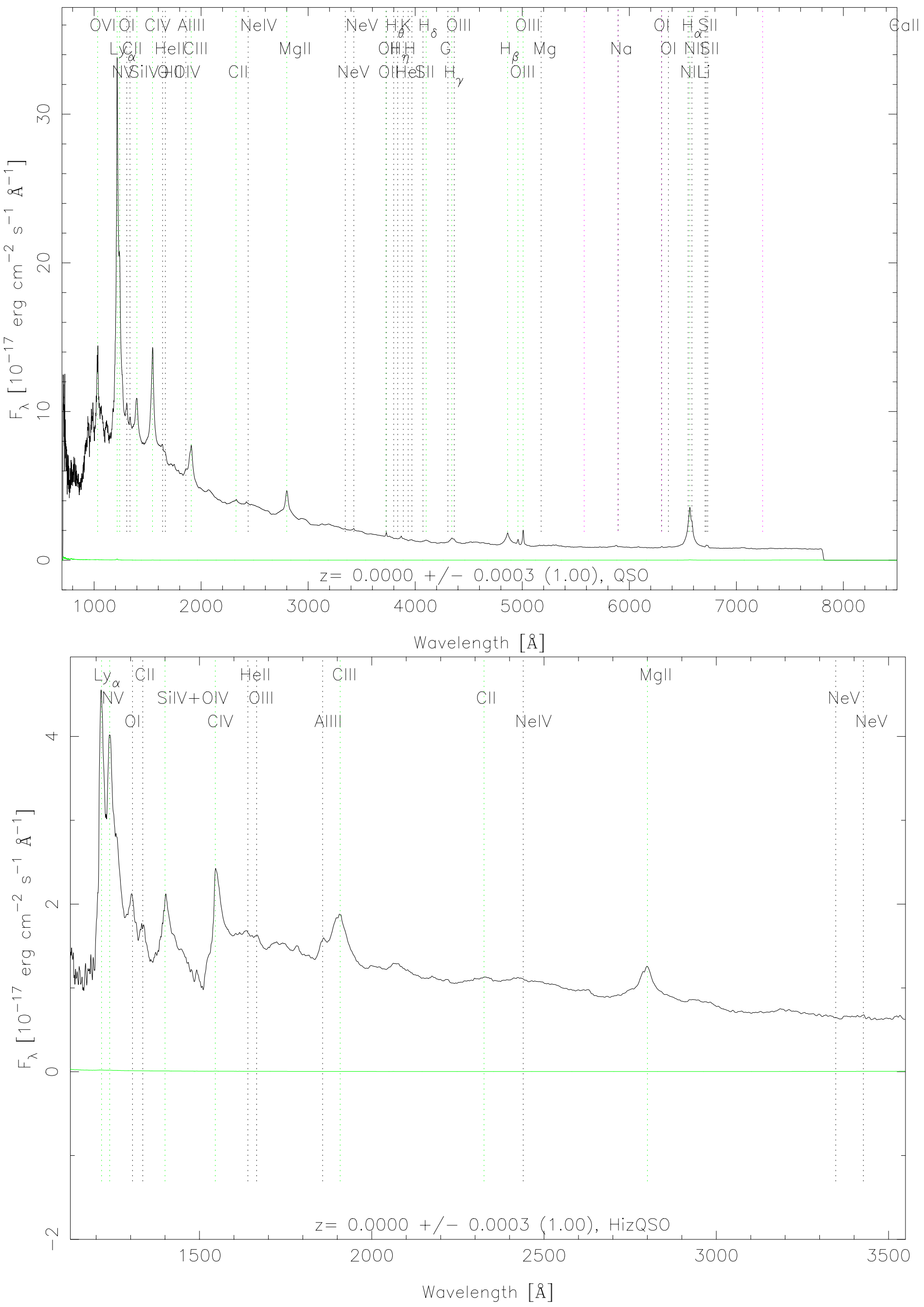

**Figure 1.3:** An example spectrum of: Top a typical quasar; Bottom a Broad Absorption Line quasar.

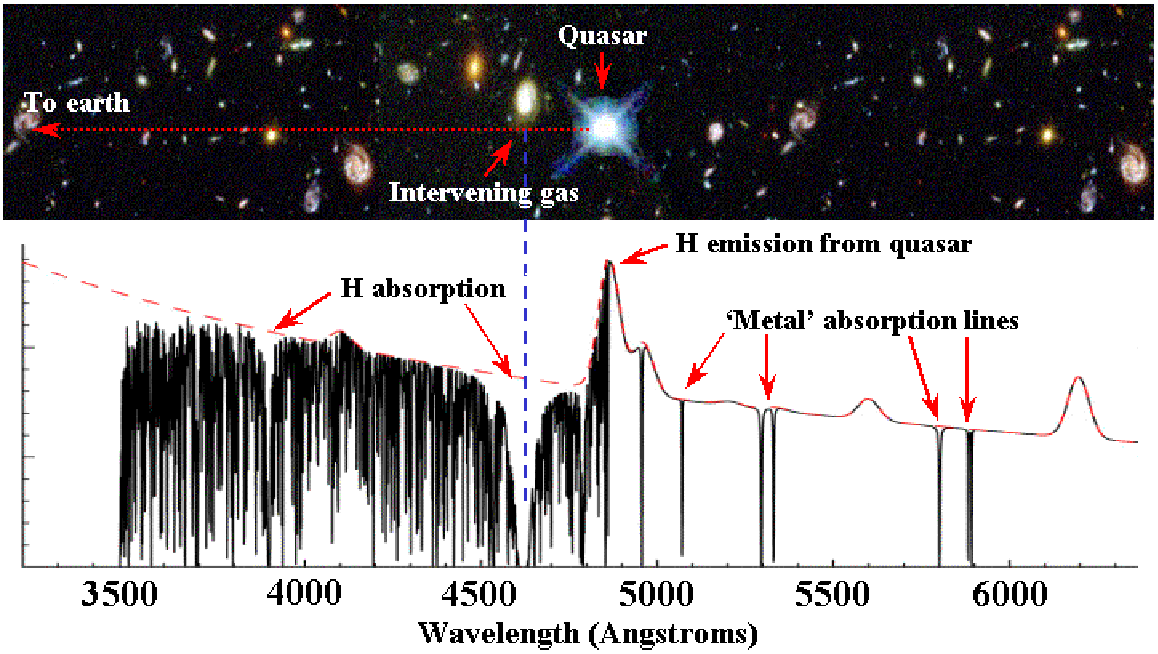

**Figure 1.4:** A quasar absorption line system arises from a galaxy intervening the line-of-sight between a quasar and us. Figure courtesy of John Webb.

up the ISM. Fig. 1.4 provides a pictorial representation of this effect. For many years, before 8 m class telescopes, these observations provided our only access to the properties of high redshift galaxies.

There are many classes of quasar absorption line systems, classified by the elements which are observed in their spectra. Primarily these classes relate to the column densities of hydrogen in the absorbers, i.e. the number of atoms of hydrogen absorbing photons along the line-of-sight between the quasar and us. Those with the largest column densities of hydrogen (greater than $2 \times 10^{20}$ atoms/cm$^2$) are termed "Damped Lyman-$\alpha$ systems" (DLAs), because of the shape of the hydrogen Lyman-$\alpha$ absorption line. These can be observed in the optical wavelength range above a redshift of $\sim$1.8, when the Lyman-$\alpha$ line, with a rest wavelength of 1216Å, is shifted into the optical wavelength range by cosmological redshift (see Section 1.2). In Fig. 1.4 the Lyman-$\alpha$ absorption line can be seen at about 4600Å. Other well studied classes of absorbers are those that show Mg II and C IV[4] metal lines. While these lines are also found in the absorption spectra of DLAs, they can be identified in spectra of objects with much lower column densities of hydrogen.

The main difference between DLAs and any other class of absorber is that the gas they contain is predominantly neutral, which implies that they are composed of cold gas clouds in the host galaxy's ISM. This fact is crucial, as cold gas is required for star formation, which suggests DLAs are important for understanding galaxy formation and evolution. This has led to the idea that DLAs provide the reservoir of gas at high redshift required for star formation over the age of the Universe. Indeed calculations have indicated that the DLAs contain most of the neutral gas in the early Universe.

Aside from the hydrogen content of high redshift galaxies, DLAs provide important measurements of metal and dust abundances in galaxy interstellar media at high redshift, not easily obtained by other methods. As described above, metals are produced by nuclear fusion in stars, and released into the ISM

[4] II means the atoms are singly ionised, IV means they are triply ionised (have lost 3 electrons).

of galaxies at the end of the stars lifetime, some of these metals then combine to form dust grains. Metal and dust abundances are therefore an important probe of how chemically evolved a galaxy is; thus, DLA observations over a range of different epochs can contribute to a picture of how galaxies have evolved from primordial gas clouds into the giant spirals and ellipticals of the present day Universe.

### 1.1.4 Galaxy evolution

Over recent years, through observations, theory and numerical simulations, we have pieced together a basic understanding of how galaxies evolve over time. Stars are formed throughout the galaxy's life when gas is available; the resulting stellar winds and supernovae as the stars explode release metals into the interstellar medium, increasing the overall metal content of the galaxy with time. As the gas is used up, no new stars are formed and the stellar population begins to age. The merger of galaxies can trigger periodic, intense bursts of star formation and it is thought that most, if not all, massive galaxies undergo a period of nuclear activity. The length, strength and properties of this activity are determined primarily by the gas reservoir at the centre of the galaxy.

Many questions remain. For example: Why does the local Universe contain such a diverse range of galaxies? How are galaxy properties related to the environment in which the galaxies live and what is the relative importance of merging on the evolution of the galaxy population? In part answering questions such as these is made difficult by intrinsic degeneracies between spectral features and physical properties. For example, older stellar populations have redder spectra, but so do young, dusty galaxies. Such degeneracies are often broken by better theoretical models and observations. Observational constraints can also prevent us from seeing the overall picture, with different classes of galaxies and quasars detected through independent methods and no single observational technique encompassing all objects. For example, the metal abundance of galaxies detected through the emission of light from stars and gas is difficult to measure if they are faint, making the history of their star formation, and how evolved they are, difficult to quantify. Conversely, galaxies detected through the absorption of light from a background quasar by the dust and gas in their ISM have precisely determined metallicities, however, the properties of their stellar populations are not well known. How the two classes of galaxies relate to one another is unclear.

## 1.2 Redshift surveys

The origin of the majority of the vast number of spectra of galaxies and quasars available for analysis nowadays are "redshift surveys", which observe the spectra of hundreds of objects in a night. The purpose of these surveys is not simply obtaining spectra for understanding the physics of galaxies and quasars however, but for creating three dimensional maps of the Universe.

Determining the distance to galaxies is not a straightforward problem because they do not have standard sizes, shapes or brightnesses. However, the cosmological expansion of the Universe provides us with a more practical alternative by relating distance ($D$) to redshift ($z$) through Hubble's law:

$$cz = H_0 D \tag{1.1}$$

where $H_0$ is Hubble's constant and the law is valid for small $z$. Redshift is simply a measure of the shift in the spectrum of an object, relative to that observed in the laboratory, and is easy to measure. It is caused by the expansion of the Universe stetching the wavelength of light over time. Hubble's law arises from the fact that, as the Universe expands, galaxies that are further away from us appear to be traveling away at a faster rate – just like stickers on a balloon that is being inflated. They therefore exhibit greater redshifts. This relation provides a useful distance estimator at large distances, when the "peculiar motions" of galaxies, caused by their gravitational pull on one another, are small compared to their redshift due to cosmological expansion.

Over the last two decades the advent of galaxy and quasar "redshift surveys", measuring the redshifts of large samples of objects, has had a dramatic impact on our understanding of both cosmology (the study of the nature and origin of our Universe) and galaxy evolution. These surveys come in two different flavours. "Deep" surveys choose a small region of sky and use large telescopes to look as far away as they can. "Wide-field" surveys are generally performed on smaller, $<$4 m class telescopes, covering a large patch of sky but not to such depth; it is data from this type of survey that are used in this thesis. Both types of surveys have been made possible by the advent of "multi-object spectrographs", instruments which allow the acquisition of the spectra of hundreds, or even thousands of objects in one go.

### 1.2.1 Deep vs. wide-field observations

There are primarily two different methods for understanding the evolution of galaxies over the age of the Universe. The first is to directly observe galaxies a long way from us, when the Universe, and the galaxies it contains, were much younger. With the advent of 8 m class telescopes, galaxies can now be observed from when the Universe was less than 10% of its current age. However the observations are difficult and obtaining spectra even more so; as a consequence the samples are biased towards the brightest objects in the Universe at any particular distance.

Alternatively, we can take a representative census of the galaxies in the local Universe, the easier observations making possible uniform spectroscopic samples of thousands of galaxies. Their spectra contain a "fossil record" of each galaxy's star formation history and this can be used, in a statistical manner, to follow the star formation rate in galaxies over a large fraction of the age of the Universe. Their distribution in space provides further insight into the formation and evolution of galaxies, with the distribution of galaxies with old stellar populations differing from those with younger populations and more current star formation. Because of the brightness of quasars, spectroscopic quasar surveys to the same limiting brightness as the galaxy surveys sample a much greater proportion of the age of the Universe. Indirectly this allows a third method of observing galaxies as mentioned above: the light of the quasars can sometimes be absorbed by a galaxy intervening the line-of-sight between us and the quasar.

This thesis makes use of data from two of these wide-field spectroscopic surveys of thousands of galaxies and quasars to investigate the properties of galaxies over the local and more distant Universe. The following subsection takes a brief look at the history of galaxy redshift surveys.

### 1.2.2 A history of wide-field redshift surveys

By 1976 the redshifts of $\sim$2 700 galaxies were known. About 600 of these had been measured primarily in order to verify Hubble's law (Humason et al. 1956). Small redshift surveys provided tentative con-

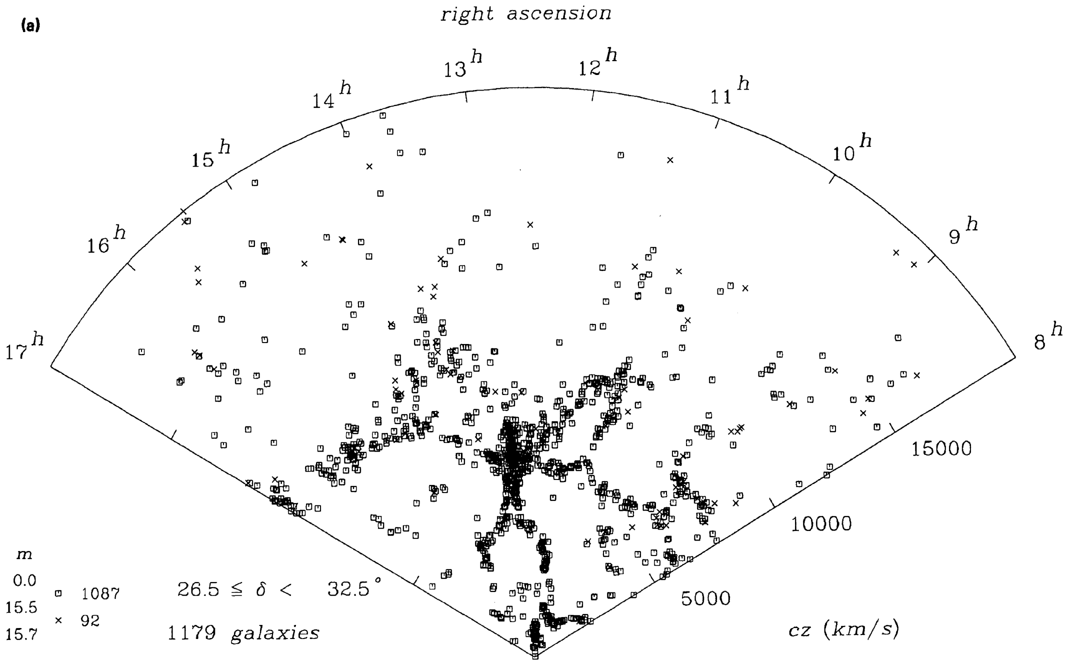

**Figure 1.5:** The CfA stickman (Geller & Huchra 1989)

firmation that the distribution of galaxies was not homogeneous on large scales, however cosmologists generally made do with the projected distribution of galaxies on the sky (e.g. Lick galaxy counts, Shane & Wirtanen 1954). The development of new technologies has always played a major role in advances in astronomy: the CfA redshift survey was the first survey to take advantage of the increased efficiency of new electronic detectors over photographic plates. With a sample of around 2 000 galaxies the CfA survey showed for the first time the filamentary distribution of galaxies through space.

The era of modern redshift surveys truly began with the extended CfA survey. This showed the full extent of the large scale structure in the galaxy distribution with more than 15 000 galaxy redshifts. The galaxies were found to map out great walls, which surrounded large low density holes (voids) in the distribution. Figure 1.5 shows the famous "stickman" of the first CfA survey slice completed in 1986. The development over the last two decades of efficient and reliable wide-field "multi-object spectrographs" which use optical fibres to transfer the light of many galaxies between the telescope and the spectrographs has resulted in enormous advances in our ability to compile large samples of high-quality astronomical spectra. Modern instruments can now observe more than twice the number of objects in the original CfA survey in a single night. The two current state-of-the-art surveys are the 2dF Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS), both of which provide the observational data analysed in this thesis; further details of these surveys are provided in Chapter 2.

## 1.3 The large scale structure of the Universe

What has the collection of hundreds of thousands of galaxy and quasar redshifts allowed us to understand about the Universe in which we live? There are two main topics on which research has focused. Traditionally the precise quantification of the galaxy distribution has received the majority of the attention, through which much has been learnt about cosmology in the last decade. More recently research has focused on the properties of the spectra of galaxies and quasars, particularly with the advent of the high quality spectra of the SDSS survey. In this thesis both topics are touched upon. In previous sections the physical information about galaxies and quasars that can be obtained from their spectra was introduced. It remains to set the context for the work on galaxy distributions presented in Chapter 4.

Measuring the distribution of galaxies in the local Universe has contributed to our understanding of the physics that lies behind the observable Universe, both in terms of the overall dynamics of the Universe and the processes involved in galaxy formation. Broadly, three techniques are applied to understanding this problem: theoretical modelling, computer simulations which track non-interacting particles over the history of mock universes, and observations. Here a brief overview of this extensive and well advanced area of research is presented.

### 1.3.1 Observations

When we look at pictures of the Universe painted for us by the thousands of galaxies observed in galaxy surveys we see that the galaxies trace out a distinct pattern of clusters, walls and voids (the “CfA-stickman” in Fig. 1.5). What causes this pattern? Furthermore, different types of galaxies cluster differently: red, passive, elliptical galaxies prefer to live in more crowded neighborhoods than blue, star-forming, spiral galaxies. This has been known since Hubble & Humason (1931) and reconfirmed many times through extensive observations. Whether this effect is caused by the physical processes involved when the galaxies originally formed, or the environment in which the galaxy subsequently finds itself, or a mixture of the two, remains a key question in astronomy today.

Many statistics are used to quantify the galaxy distribution. However, theories and models are needed with which to compare the results in order to gain insight into the physical reasons behind the observed distributions.

### 1.3.2 Theory of galaxy formation

What factors determine where a galaxy will form? In the “high peak scenario” the efficiency of galaxy formation is assumed to be greater in regions of high matter density. This simple idea seems intuitively reasonable as the baryonic particles that are the building blocks of the luminous parts of a galaxy will collect in the regions of the dark matter distribution with the highest density, subsequently forming the galaxies we see. Over the course of time, the galaxies will continue to map out the peaks in the dark matter distribution.

Once the galaxies have formed, the change in their distribution with time is predicted to be primarily driven by the force of gravity. Following this evolution in general involves numerical computer simulations, because the dynamical equations can not be solved exactly. However, some analytical predictions have been made and the hierarchical clustering distribution predicted by “gravitational instability theory” has been confirmed by the 2dFGRS survey.

The high peak scenario has been replaced in recent years by the “halo model”, a description that links the galaxy distribution to that of the dark matter via “dark matter halos”. As a phenomological framework the halo model has successfully explained the form of the galaxy correlation function – the likelihood of a galaxy having a neighbour at a given distance. This function has two separate components which the halo model explains as arising, respectively, from galaxies that live within the same halo, and galaxies that live in different halos. The difference in clustering of red and blue galaxies is explained simply by a difference in mass of the haloes surrounding the different subclasses.

### 1.3.3 The cosmological concordance model

By studying the distribution of galaxies, we can begin to understand the properties required in a region of space for a galaxy to form. However, one of the main reasons for carrying out galaxy redshift surveys is their very important role in cosmology. By combining the distribution of galaxies seen in the local Universe with measurements of the Cosmic Microwave Background (CMB), the relic radiation left over from the Big Bang, galaxy redshift surveys constrain models for the dynamics of the Universe (Spergel et al. 2003). In the current "concordance cosmological model" the Universe is made up of 70% dark energy and 30% matter. Of this 30% matter, 3% is in the form of baryonic particles – that make the ordinary matter we see – the remainder is "dark matter". Hubble's constant (equation 1.1) is measured to be $70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Without galaxy redshift surveys such firm constraints would not be possible.

### 1.3.4 Galaxy formation and evolution through simulations

Because analytic theory is limited to approximate solutions and phenomenological models, simulating the evolution of the dark matter distribution in large volumes of the Universe using either N-body or hydrodynamic codes on massive supercomputers is now common practice. Taking as input the *cosmological concordance model* – the equations which describe the underlying nature of the Universe, for example how fast it is expanding – the dark matter particles are followed through time as they cluster under the force of gravity. Adding galaxies into these simulations remains a difficult problem, in part due to our basic lack of understanding of the highly nonlinear processes involved in the formation of galaxies, hence creating mock Universes with which we can directly compare all observations is not yet possible. However, these simulations have proved useful for the development of statistical tools for application to understanding the large scale distribution of galaxies in real surveys. They have also been used to study the physical processes involved when galaxies form, such as the need for some sort of "feedback" – when gas is removed from galaxies as they form, possibly by supernovae or quasar activity.

## 1.4 Thesis Motivation and Overview

This introduction has, due to the range of topics in the thesis, necessarily covered a large number of subjects associated with extra-galactic astronomy, and spectroscopic galaxy and quasar surveys. Hopefully it has made clear why modern redshift surveys play a vital role in helping us understand the physics underlying the dynamics of the Universe and the creation of the galaxies contained within it. Through the absorption lines caused by atomic transitions, and the precise redshifts afforded, spectra are crucial to the work presented here. Modern redshift surveys provide hundreds of thousands of spectra, allowing the exact quantification of the distribution of galaxies in space, the accurate determination of average properties of objects and the collection of statistical samples of rare objects.

One way to extract intelligible results from the huge quantities of data now available is through statistical techniques. Whether used to decompose vast matrices of thousands of spectra each a few thousand pixels long, to improve the quality of the final data product, to find rare objects, or to study the large scale structure of the Universe, statistical methods, when understood and applied appropriately, contribute enormously to modern astronomy. The use of such methods is a theme of this thesis.

The thesis makes use of the two largest local spectroscopic galaxy surveys of our time: the 2dFGRS and the SDSS. While both were conceived at similar times, their ultimate (and continuing) results represent a recent change in the most active research areas in astronomy. In the 1980's and 90's the question of the distribution of galaxies was foremost in many astronomers minds, now it is the question of galaxy formation and evolution. The 2dFGRS was designed and operated as a "redshift machine" - the task of simply collecting spectra and determining the redshifts was expertly achieved. Obtaining galaxy properties from the spectra proved difficult and only basic comparisons of different populations is possible. The SDSS, on the other hand, provides high quality spectra that contain a wealth of physical information on the objects surveyed and the era of simply making 3D maps has been left behind. We are now in a position where traditional techniques for understanding the physics behind the observed spectral energy distributions can be combined with the power of statistics.

In Chapter 2 further details are provided of both of these surveys, along with a brief overview of some of the specific science goals they have achieved. The remainder of this introduction outlines the four remaining chapters in the thesis.

### *The large scale distribution of red and blue galaxies*

The complete 2dFGRS is ideally suited to studying the distributions of galaxies in space, with a well quantified selection function embodied in specially designed random catalogues, contiguous regions of space surveyed, and semi-analytic models purposefully designed with which to compare the 2dFGRS results.

In Chapter 3 the galaxies are treated purely as points tracing the matter distribution in three dimensions, the sample is split only by spectral type (passive and star-forming) and colour (red and blue). The joint distribution of these different types of galaxies is fitted with analytical models, and a similar analysis carried out on mock galaxy distributions. The form of the joint distribution may relate to different formation scenarios, although the next generation of N-body and hydrodynamic simulations will be required to fully understand the observations. In the future the SDSS will allow more extensive analyses of this type, using different galaxy properties to better understand the physical processes involved in the joint distributions.

### *The problem of the sky....*

The SDSS represents a huge leap forward in data quality compared to previous photometric and spectroscopic surveys; combined with the huge area of sky covered it will be an invaluable resource for many years to come. However, problems remain in the reduction of such large quantities of data, which unnecessarily restricts some science applications.

While telescopes and their instruments take spectra of faint objects, the majority of the light that lands on the detectors is from the night sky. When an object is faint in comparison to the sky background, small errors in sky subtraction can result in large systematic errors in the final spectrum of the object. Optical fibre systems, which have made modern spectroscopic surveys viable by allowing the collection of hundreds of spectra in a single observation, have now been available for a couple of decades. However, the problem of removing sky light from the data signal without drastically reducing the efficiency of the instrument remains a far more difficult problem than in traditional spectroscopy. In Chapter 4 this problem is looked at in detail and a new statistical technique is developed to automatically remove artifacts remaining after the subtraction of the night sky in many SDSS spectra.

### *Dust in quasar absorption line systems*

With the high quality SDSS spectra many new scientific investigations have been made possible. The large spectroscopic quasar survey is ideal for locating large samples of quasar absorption line systems. A unique advantage over traditional quasar absorption line surveys is the availability of large numbers of quasars without intervening absorbers with which to create “control” samples for comparisons.

Chapter 5 presents two samples of quasar absorption line systems found in the SDSS quasar catalogue with $0.84 < z < 1.3$. The first is a new sample of Ca II absorption line systems, a class of absorbers largely unstudied till now; the second is defined by the strength of Mg II and Fe II lines in such a way that a large number are expected to be DLAs. The average dust content of the absorption–selected galaxies is measured by comparing the colour of the quasar spectra with and without absorption line systems.

The Ca II absorption line systems may prove to fill an important gap in our understanding of the metallicity and dust evolution of galaxies. Their dust contents suggest significant column densities of hydrogen, greater than the nominal limit for classical DLAs. If this were the case Ca II–selected DLAs would provide an important method by which DLAs can be detected without the need for ultra–violet spectroscopy, no longer available with further Space Shuttle missions to the Hubble Space Telescope grounded.

In order to detect dust reddening we must observe a region of the galaxy spectrum where dust causes a large enough gradient in the background quasar spectral energy distribution. This places the redshifted Ca II lines in the region of poor sky subtraction in the SDSS spectra. The techniques of Chapter 4 are employed to reduce the systematic noise in the spectra before searching for Ca II lines.

### *Metals in Ca* II *absorption line systems*

As discussed in this introduction, dust does not only cause obscuration and reddening of light. Dust grains are composed of metals, and different metals are more or less susceptible to being incorporated into grains. This means that the absorption line spectrum of the Ca II systems can also be used to infer their dust content. With the extensive wavelength range of the SDSS spectra a subsample of the Ca II absorbers have spectra which cover the region of absorption lines caused by the Zn II ion. This is important because Zn has little affinity for dust grains, thereby providing a firm reference against which the dust reddening and other dust depleted elements can be compared.

In Chapter 6 a full analysis of the weak metal lines available in the Ca II absorption line spectra is carried out, and column densities are obtained for Zn II, Cr II, Ti II, Fe II and Mn II. These are compared with metals seen in the Milky Way ISM, and high redshift DLAs. When combined with the dust reddening analysis, the results suggest the Ca II–selected galaxies are more chemically evolved than DLAs at high redshift. Ca II absorption line systems may provide the missing link between dust free, metal–poor DLAs and massive, chemically evolved, star forming galaxies detected through emission of light from their stars, gas and dust at similar redshifts.

### *Notes on conventions*

- Following the convention employed in the SDSS, vacuum wavelengths are used throughout. Appendix A lists vacuum and air wavelengths for lines relevant to this thesis which fall in the optical and ultra-violet wavelength ranges.
- The concordance cosmological model assumed is $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $h \equiv H_0/100\,\mathrm{km\,s^{-1}\,Mpc^{-1}} = 0.7$.

# 2

# Analysis of spectroscopic datasets

*This thesis presents work using data from two of the largest wide field spectroscopic surveys ever carried out: the 2-degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS). A summary and comparison of these two surveys is presented in this chapter. The development of new techniques with which to analyse the large numbers of spectra efficiently, from these and other redshift surveys, is assuming an increasingly important role in modern astronomy. One such technique, principal component analysis (PCA), has been used extensively throughout this thesis. Here the mathematical details of PCA and associated techniques are presented.*

In this thesis two datasets are used, the 2dFGRS and SDSS, the former to parameterise the large scale galaxy distribution, and the latter to obtain physical information about galaxies themselves. The reason behind this choice of datasets is a matter of survey designs and the details of each of these surveys, their advantages and limitations, are provided in this chapter. Both are wide field spectroscopic galaxy and quasar surveys, but each was designed and carried out with differing scientific outcomes in mind. The 2dF instrument is a “redshift machine” designed to obtain many galaxy redshifts quickly in order to map out the distribution of local galaxies in three dimensions. The survey selection function and completeness as a function of position on the sky were carefully calculated but less attention was paid to the flux calibration and continuum properties of the spectra themselves. For these reasons the main scientific contributions of the 2dFGRS have been to the quantification of the large scale galaxy distribution, although significant results pertaining to our understanding of the galaxy population were also obtained.

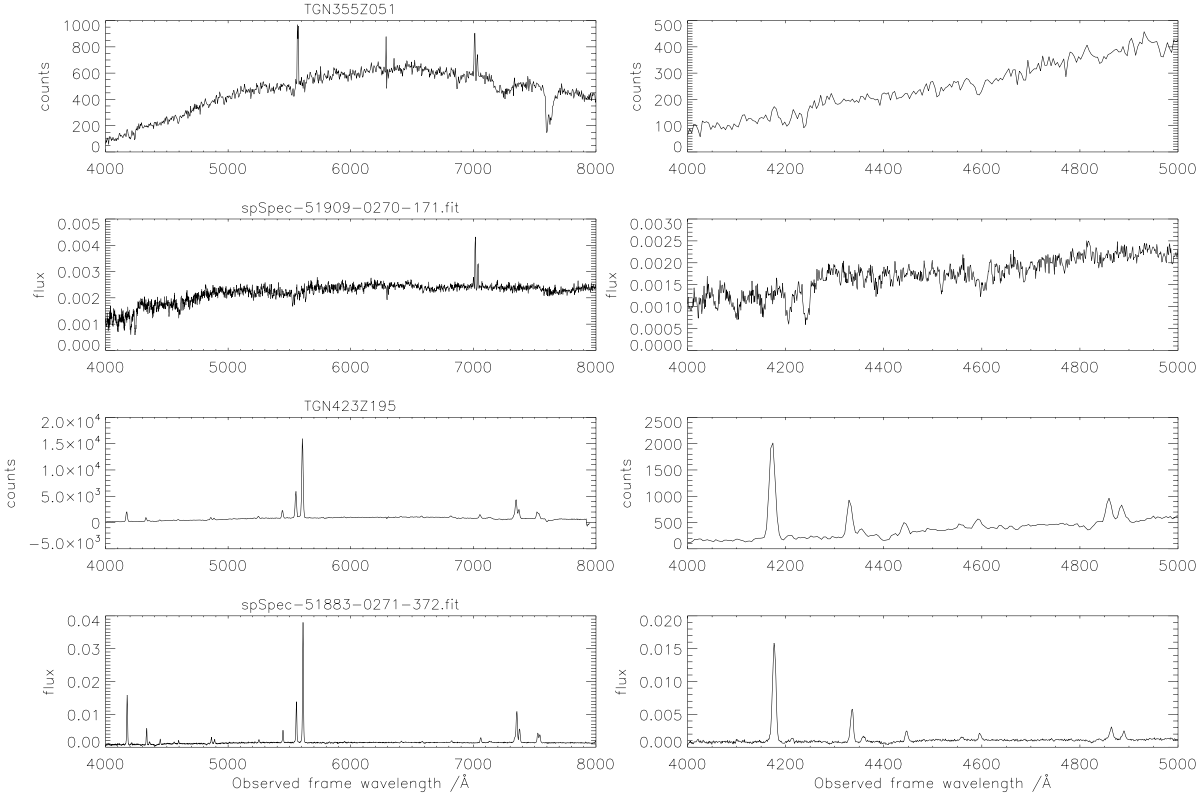

**Figure 2.1:** Left: Two examples of spectra from objects targeted by both the 2dF (upper) and SDSS (lower) with 2dF name and SDSS file identifier given in the titles. The wavelength range of the SDSS spectra have been reduced to match that of the 2dF spectra, and the flux/ Å term of the SDSS spectra multiplied out for better comparison. Right: a small section of the wavelength range is reproduced for better appreciation of the improved resolution of the SDSS spectra.

The SDSS was conceived at a similar time to the 2dFGRS, but developed over a considerably longer timescale. For studying the properties of individual and populations of objects the SDSS spectra are far superior to 2dF spectra with an extended wavelength range, greater resolution and accurate flux calibration. Combined with the five band ($u, g, r, i, z$) CCD photometry covering 1/4 of the sky by completion of the project, the SDSS represents the most impressive dataset in optical astronomy today. However the survey geometry is currently complicated and completeness not totally understood making analysis of the galaxy distributions, such as carried out on the 2dFGRS dataset, rather more difficult. Although the spectroscopic catalogues have allowed the average spectral energy distributions (SEDs) of quasars to be studied in greater depth than ever before, samples suitable for the calculation of statistical properties, such as luminosity functions, have not been released to the public.

Fig. 2.1 shows two examples of spectra of objects observed by both the SDSS and 2dF spectrographs. The superior resolution and sky subtraction of the SDSS spectrum is clear. The differing overall shapes of the spectra are caused by the lack of spectrophotometry of the 2dF spectra. Due to the very different instrumentation and survey designs, the scientific emphasis of the two surveys has differed, as reflected by the chapters in this thesis. One further difference is in the organisation of the teams carrying out the surveys. The 2dFGRS was designed, carried out and the data analysed prior to public release by a core group of about 30 scientists in Australia and the UK. Conversely, the SDSS data is aimed to be released to the public within a year to 18 months of first being collected and the core team is large, comprising over

**Table 2.1:** Numbers of early/late and red/blue galaxies in the 2dFGRS catalogue with good quality spectra ($Q >= 3$).

| | red/early | blue/late | Total |
|---|---|---|---|
| colour | 77 120 | 144 292 | 221 414 |
| $\eta$ | 74 548 | 118 424 | 192 979 |

150 scientists in several countries. The following two sections summarise the survey designs, properties of each dataset, and some of the scientific results to emerge so far, concentrating in particular on those relevant to this thesis.

## 2.1 The 2dFGRS

The 2-degree Field Galaxy Redshift Survey was carried out between 1997 May and 2002 April on the Anglo-Australian Telescope in New South Wales, Australia. A total of 221 414 good quality galaxy spectra were obtained with a median redshift of 0.11. Table 2.1 gives the respective numbers of early- and late-type galaxies (spectral classification) and red and blue galaxies (photometric classification) in the final publicly released catalogue. The companion 2dF quasar survey (Boyle et al. 2000), was carried out alongside the 2dFGRS, but has not been used in this thesis.

### 2.1.1 Survey design

The target galaxies for the survey were selected from a revised and extended version of the APM galaxy catalogue (Maddox et al. 1990), with a nominal extinction corrected limiting magnitude of $B_J = 19.45$. The 2dF instrument is capable of obtaining up to 400 spectra at a time over a two degree diameter field using optical fibres to transfer the signal to two spectrographs. A robotic fibre positioner and tumbling mechanism allows each survey field to be configured in situ as the previous field is exposed and the entire instrument is mounted at prime focus on the AAT. Fig. 2.2 shows a photograph of the 2dF instrument on the AAT. The wide, two degree field is achieved with a corrector lens, however, due to design constraints, significant chromatic dispersion of images remains, varying as a function of radius from the field center. It is this effect that makes obtaining absolute spectrophotometry for 2dF spectra difficult.

The main survey area comprises two rectangular strips of sky in the Southern and Northern Galactic caps covering an area of 1500 square degrees; one hundred random fields were also selected around the southern strip. Fig. 2.3 shows the region of sky covered by the 2dFGRS. At the median redshift, the physical length of the survey strips is $375\,h^{-1}\,\mathrm{Mpc}$ and the Southern and Northern Galactic Pole (SGP and NGP) regions have widths of $75\,h^{-1}\,\mathrm{Mpc}$ and $37.5\,h^{-1}\,\mathrm{Mpc}$ respectively (assuming standard $\Lambda$CDM cosmology[1], Peacock 2003). Fig. 2.4 shows slices through the two rectangular survey strips, note the difference in number of objects and distance reached when compared to the similar plot from the earlier CfA survey (Fig. 1.5).

Further details on the survey can be found in Colless et al. (2001; 2004) and on the web at http://msowww.anu.edu.au/2dFGRS/.

[1] $\Omega_m = 0.3, \Omega_\Lambda = 0.7, h = 0.7$

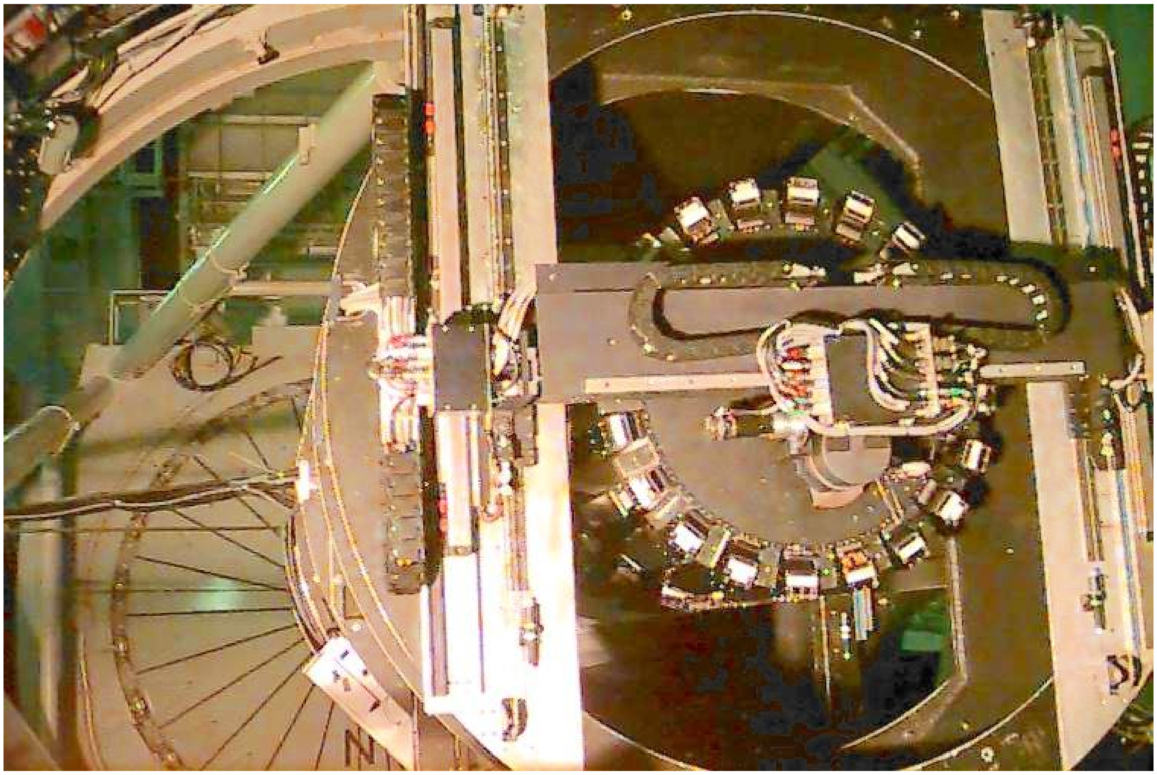

**Figure 2.2:** The 2dF instrument mounted at prime focus on the Anglo-Australian Telescope. The robotic positioner can be seen moving the optical fibres ready for the next exposure.

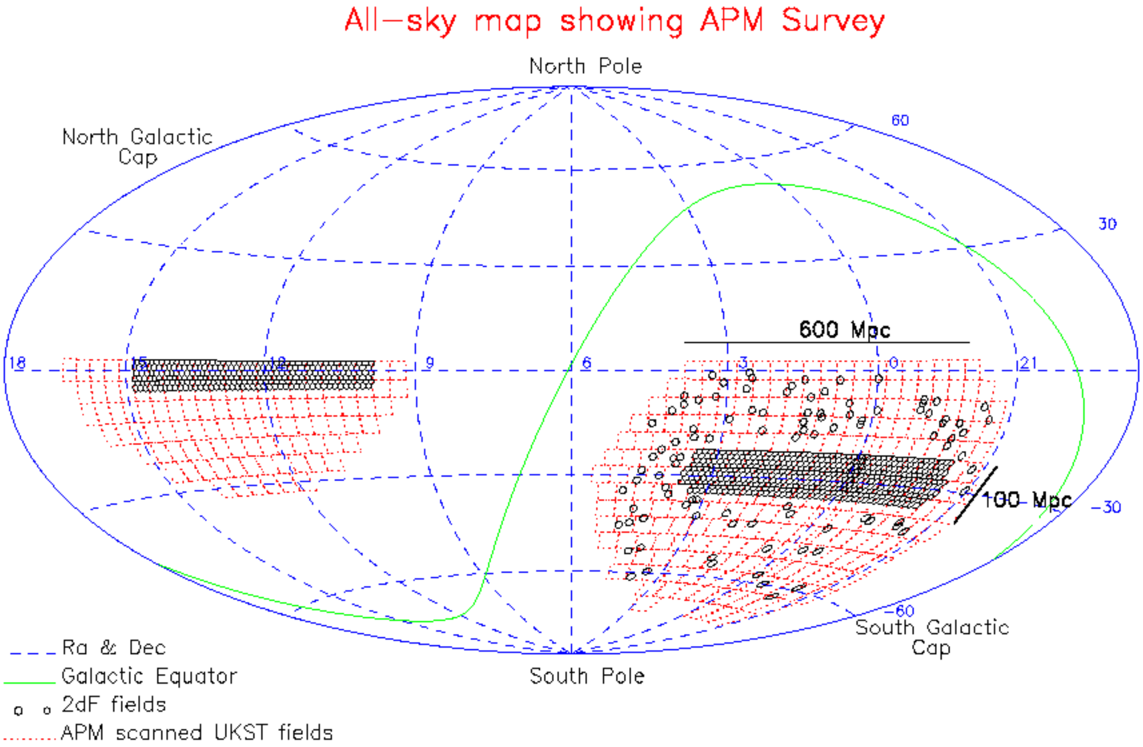

**Figure 2.3:** The 2dFGRS survey area. The red squares indicate the APM survey fields and the black circles the 2-degree regions observed by the 2dF.

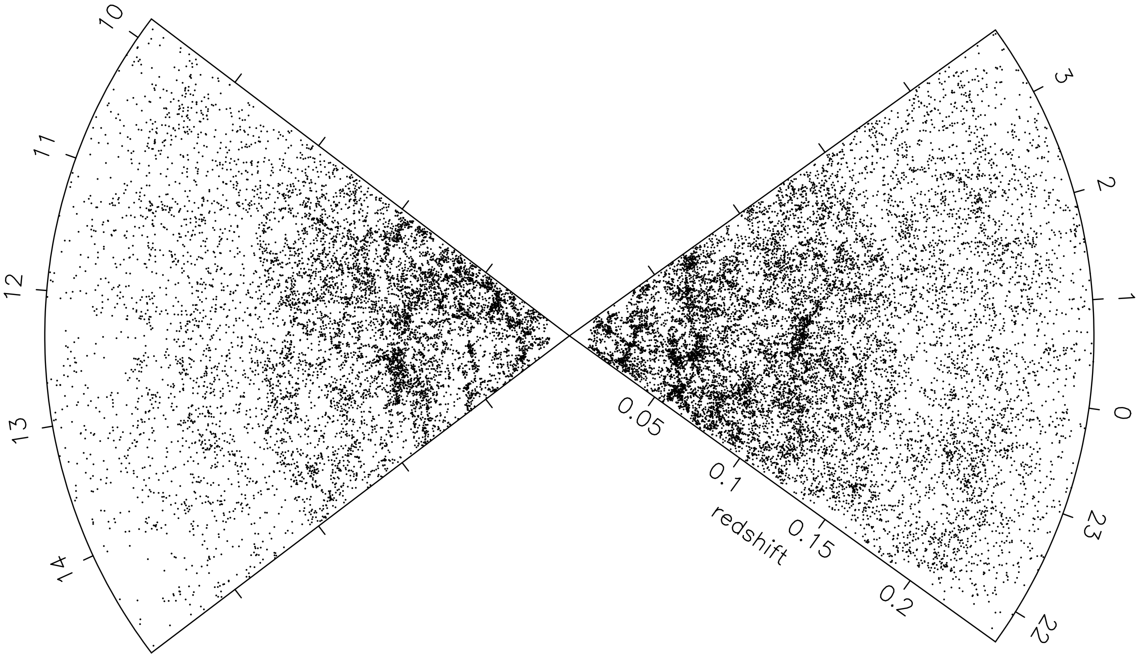

**Figure 2.4:** The distribution of 40 000 randomly selected galaxies in the 2dFGRS, with declination projected onto the page. Right Ascension is given on the left and right axes and redshift increases from our position at the center of the figure.

### 2.1.2 Scientific highlights

The primary aim of the 2dFGRS was to quantify the distribution of galaxies on 10-100 Mpc scales, i.e. between the non-linear regime and scales probed by CMB experiments. The motivation behind the creation of large galaxy surveys for this purpose is clear from Fig. 2.5 which shows the effect of different underlying cosmological models on the distribution of galaxies. Many studies of the large scale structure of the galaxy distribution have been undertaken with the 2dFGRS, recently culminating in confirmation of “baryon wiggles” in the power spectrum of the 2dF dataset by Cole et al. (2005). These are caused by fluctuations in the radiation field, which dominates the content of the early Universe, on the distribution of baryons, with the implication that a high fraction of baryons must have existed at that time in order for the wiggles to be observed. A brief summary is now given of the main results from the 2dFGRS on the distribution of galaxies and cosmological parameters derived from it.

**Power spectrum:** The amplitude of fluctuations in the distribution of galaxies as a function of scale is dependent upon the total matter content and baryonic fraction of the Universe. Percival et al. (2001) and more recently Cole et al. (2005) provided the most accurate measurement of the galaxy power spectrum to date finding $\Omega_m h = 0.168 \pm 0.016$ and $\Omega_b/\Omega_m = 0.185 \pm 0.046$. By combining with results from the CMB the cosmological constant can be constrained, without the need for results from high redshift supernovae, to be $0.65 < \Omega_\Lambda < 0.85$ (Efstathiou et al. 2002).

**Higher order moments:** Were galaxies to be discrete points sampling a Gaussian field, the power spectrum would encode all information on their distribution; however this is known not to be the case due to the effect of gravity. Baugh et al. (2004) showed the higher order moments of the galaxy distribution in the 2dFGRS follow the predictions of gravitational instability theory.

**Redshift-space distortions:** The correlation function $\xi(\sigma, \pi)$ measures the excess probability over ran-

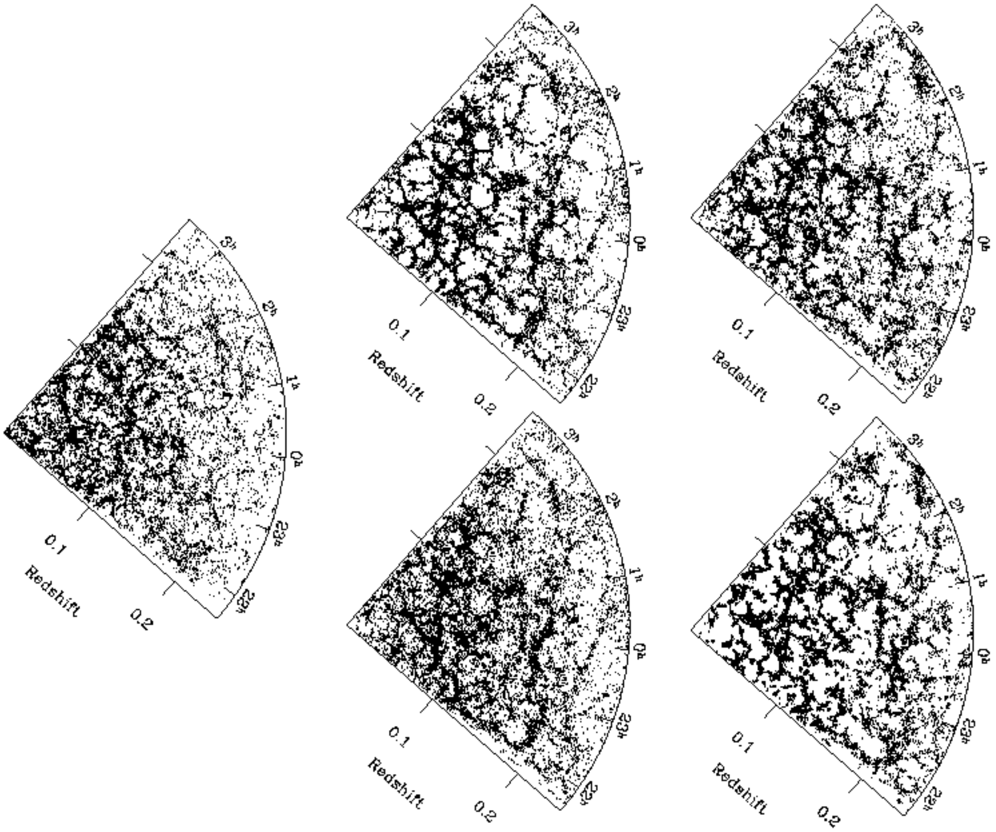

**Figure 2.5:** Galaxy surveys as a probe of cosmology. On the left the true galaxy distribution as measured by the 2dFGRS. On the right are four simulated universes of differing cosmologies: the impact of varying cosmological parameters on the observed distribution of galaxies is clear (Cole et al. 1998).

dom of finding a pair of galaxies with separation $r$ along and perpendicular to the line-of-sight. Because redshifts are used in place of distances, the peculiar velocities of galaxies and their coherent in-fall towards mass concentrations result in the correlation function becoming elongated on small scales and squashed on larger scales: the "finger-of-god" and linear $\beta$ effects. The magnitude of the latter effect depends on, and can therefore constrain, the matter density of the Universe $\Omega_m$ provided galaxy bias ($b$) is known. Peacock et al. (2001) found $\beta = \Omega_m/b = 0.47 \pm 0.07$ for the 2dFGRS.

**Bias:** The concern that galaxies may not accurately trace the underlying matter distribution was somewhat assuaged by Verde et al. (2002) and Lahav et al. (2002) both obtaining a bias parameter consistent with unity over a wide range of scales.

**Neutrino mass:** Uncovering properties of particles in the standard model was not part of the original scientific case, however Elgarøy et al. (2002) showed that the 2dFGRS could place tighter constraints on the absolute mass of neutrinos than particle physics experiments, which primarily measure relative masses of the different neutrino flavours.

Rather than using galaxies simply as tracers of the underlying density field, information can be gleaned from the survey on the physical properties of the galaxies themselves, although the scientific applications are somewhat limited by the quality of the 2dF spectra.

**Luminosity function and cosmic star formation history:** The mean current star formation in the local Universe was measured by Norberg et al. (2002a) by combining the redshifts with the APM plate $B_{\mathrm{J}}$ magnitudes. By using the fossil record of past star formation contained within the mean 2dF galaxy absorption spectrum and comparing with spectral synthesis models, Baldry et al. (2002) concluded that star formation decreased from a peak at $z \sim 1$, a fact now well established from many different observations.

**Environmental dependence of galaxy properties:** Lewis et al. (2002) found star formation rate (measured from the H$\alpha$ line) decreased rapidly for relatively low enhancements of local density. This suggests severe physical effects such as ram pressure stripping of gas can only be partly responsible for the low star formation rate of cluster galaxies. The dependence of galaxy clustering on luminosity and spectral type measured by Norberg et al. (2002b) and Madgwick et al. (2003b) has less well understood implications for galaxy formation and evolution scenarios, but certainly will provide important constraints for models in the future.

## 2.2 The SDSS

By the time of completion the Sloan Digital Sky Survey will have imaged one quarter of the sky in five bands down to a limiting magnitude of ∼22 in the $r$-band and obtained spectra for a million of the objects detected in the photometry. Table 2.2 summarises some of the survey details and numbers relevant to this thesis for the second, third and fourth data releases (DR2, DR3 and DR4). The areas of sky covered by the DR4 photometric and spectroscopic catalogues are shown in Fig. 2.6.

### 2.2.1 Survey design

The SDSS survey is being carried out on a dedicated 2.5m telescope at the Apache Point Observatory in New Mexico. Photometric data is collected on an array of 30 2048x2048 pixel CCDs arranged in five rows and six columns, each row being covered by one of the five SDSS filters, $u$, $g$, $r$, $i$ and $z$. The telescope is operated in drift scan mode, with the CCDs slowly read out at the same rate as the objects pass down the columns. It takes 54 seconds for an object to cross a CCD chip, and a further 18 seconds to cross the gap between chips. The result is a 54 second exposure in each of the five bands with a spacing of 72 seconds between the start of each. A further 24 CCDs around the central chips are used to observe reference stars for accurate astrometry. A second 20” diameter telescope on site observes photometric standards throughout the night. Photometric data is only collected on nights with good conditions (seeing $\lesssim$1.5”), spectroscopic data is collected on other nights.

The spectroscopic data is collected on four CCD chips using two spectrographs with resolution $\lambda/\Delta\lambda \approx 1800$, each with a red and blue channel, separated by a dichroic at $\sim 6150$Å. The wavelength range covered is 3800-9180Å. The spectra are fed to the spectrographs along 640 optical fibres with 3” diameter and covering a three degree field of view. At least three 15 minute exposures are taken, with more exposures taken until a signal-to-noise ratio (SNR) of more than 15 is obtained in $g$ and $i$.

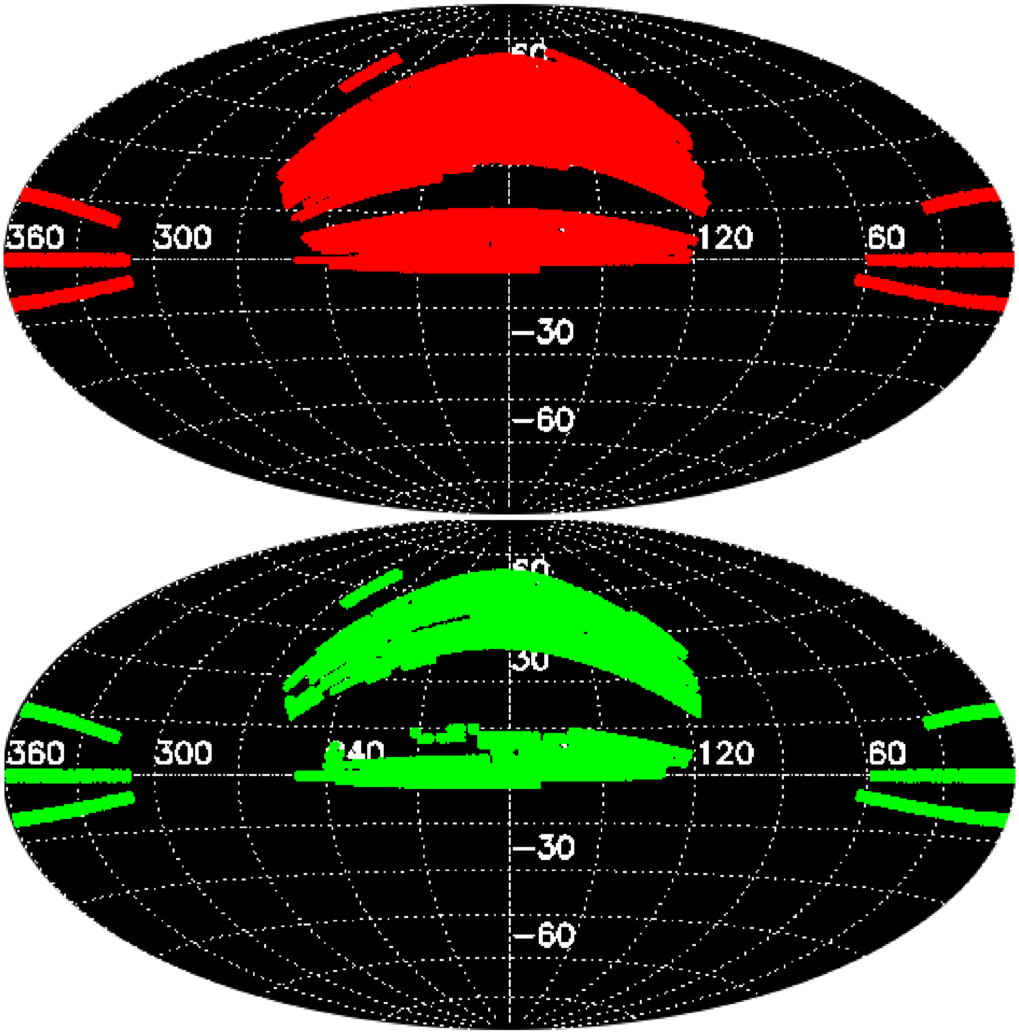

**Figure 2.6:** The SDSS DR4 survey area for the photometric (top) and spectroscopic surveys (bottom) in all-sky right ascension and declination projection.

**Table 2.2:** Some details of the SDSS data releases used in this thesis. These numbers are taken directly from the SDSS website and include repeated observations and "bonus" plates not part of the main catalogue.

| | photometric catalogue | | spectroscopic catalogue | |
|---|---|---|---|---|
| catalogue and release data | Area of sky (sq. deg) | No. of objects | Area of sky (sq. deg.) | No. of objects |
| DR2<br>release 03/04 | 3324 | 88 million | 2627 | total: 367 360<br>galaxies: 260 490<br>quasars: 36 032<br>stars: 34 998<br>sky: 18 767 |
| DR3<br>release 09/04 | 5282 | 141 million | 4188 | total: 528 640<br>galaxies: 374 767<br>quasars: 51 027<br>stars: 50 369<br>sky: 26 819 |
| DR4<br>release 06/05 | 6670 | 180 million | 4681 | total: 849 920<br>galaxies: 565 715<br>quasars: 76 483<br>stars: 102 714<br>sky: 44 363 |

### *Spectroscopic target selection*

Three main classes of objects are targeted for spectroscopic follow up after the photometric data has been reduced: galaxies, quasars and luminous red galaxies. Galaxies are selected based on magnitude and morphology: the difference between point-spread-function (PSF) magnitudes and model magnitudes $> 0.3$ mag in $r$, $r < 17.77$, Petrosian half-light surface brightness $< 24.5$ mag arcsec$^{-2}$ and a bright limit of 15, 15 and 14.5 mag is set respectively in $g$, $r$ and $i$ to prevent fibre cross talk. The SEDs of quasars differ substantially from blackbody-like stellar spectra and, for SDSS filters, quasars are well separated from the stellar colour locus for all but $z \sim 2.7 - 2.8$. The main quasar sample is selected as those objects which lie more than $4\sigma$ away from the stellar locus in the $u - g$, $g - r$, $r - i$ colour cube and have $15 < i < 19$. Further cuts are applied to regions in the colour cube known to contain outlying stars and objects that are blended are discarded. High-redshift quasars are selected in the same way from the $g - r$, $r - i$ and $i - z$ colour cube and to have $i < 20$.

### *Information sources*

Technical details of the SDSS hardware and software are given by York & et al. (The SDSS Collaboration), (2000) and in the SDSS Project Book which can be found online. The data reduction pipelines, details of the catalogue parameters and search facilities can be found in the Early Data Release paper (Stoughton et al. 2002) and each data release is accompanied by a brief summary of progress (Abazajian et al. 2003; 2004; 2005; Adelman-McCarthy et al. 2005). Further details on the spectroscopic target selection for the main galaxy sample are given by Strauss et al. (2002). Finally, the SDSS website (http://www.sdss.org) provides further information of interest, including precise details of algorithms used when deriving parameters from the final photometric and spectroscopic datasets.

### 2.2.2 Scientific results

The range and quantity of results to come from the SDSS to date are numerous and any selection of “important” results is bound to be subjective. This subsection therefore summarises results from studies with a bias towards topics broadly relevant to this thesis.

#### *Physical properties of galaxies*

Comparison with spectral synthesis models (e.g. Bruzual & Charlot 2003) has allowed the measurement of “hidden” physical properties of galaxies, such as stellar masses and star formation histories (SFHs). While errors on parameters for individual galaxies remain substantial, the large number of spectra allow reliable distributions and trends to be obtained.

The stellar mass distribution appears to be bimodal, reminiscent of the bimodality seen in galaxy types and colours (Kauffmann et al. 2003b). Similar to previous results from much smaller surveys, many physical parameters are found to be correlated with stellar mass such as stellar age, star formation rate (SFR), dust content and AGN activity (Kauffmann et al. 2003a; 2003c). Tremonti et al. (2004) find metallicity to be strongly correlated with stellar mass over three orders of magnitude in stellar mass and 10 in metallicity and Brinchmann et al. (2004) find a strong correlation between SFR estimated by nebula emission lines and stellar mass in star-forming galaxies. They also find the $z \sim 0.1$ Universe to be forming stars at $\lesssim 1/3$ the past averaged rate. Using a very different method which makes use of the entire stellar spectrum rather than specific indicators, Panter et al. (2003) calculated SFRs as a function of lookback time from the SFHs of the SDSS galaxies, extending as far back as $z \sim 1$. They find star formation to have declined by a factor of $\sim 30$ since $z \sim 1$. While all these results are in qualitative agreement with previous, smaller studies, the large, homogeneous SDSS samples allow the precise distributions of parameters to be quantified which provide a firm basis for constraining theoretical models.

#### *Environmental effects*

An obvious approach to take in the hope of gaining insight into the reasons behind the observed trends in galaxy properties is to study their dependence on local environment. The SDSS allows in-depth analysis of the well known morphology-density relation over a wide range of galaxy densities. Results show that while, for example, SFR, colour and stellar mass distribution correlate strongly with environment, properties relating to galaxy structure do not (Hogg et al. 2003; Gómez et al. 2003; Blanton et al. 2003; Kauffmann et al. 2004), suggesting that galaxy properties are not determined primarily by mergers and galaxy harrasment. However, the root causes of the trends remain unclear.

#### *Properties of quasars*

A major contribution to understanding the early Universe has been made by the discovery of a significant number of high redshift ($z > 5.7$) quasars in the SDSS catalogue (e.g. Fan et al. 2004). Regions with no flux blueward of the Lyman-$\alpha$ line (complete Gunn-Peterson trough) are found in all quasars at $z > 6.1$, possibly indicating that the Universe was not fully reionised by this time.

The large spectroscopic catalogues of SDSS quasars - 46 420 objects in DR3 (Schneider et al. 2005) - are allowing mean and intrinsic variation in properties to be quantified as never before. For example, high SNR composite spectra provide a reference for many studies (vanden Berk et al. 2001), the occurrence of rare objects such as Broad Absorption Line quasars (BALs) can be precisely quantified (Reichard et al. 2003a), and objects showing reddening due to dust found (Richards et al. 2003).

*Quasar absorption line systems*

The SDSS quasar survey also contains a wealth of information on the ISM of galaxies which intervene the line-of-sight to the quasars. Its principle contributions have been towards calculating the incidence of absorbers and discovery of objects for follow-up observations, both crucial to uncovering the physical nature of the systems hosting the absorbers.

At the highest column density end of the distribution, SDSS DR3 has allowed the quantification of the number density of Damped Lyman-$\alpha$ systems (DLAs) (Prochaska et al. 2005). Prochaska & Herbert-Fort (2004) find the cosmological mass density of neutral hydrogen to decrease by a factor of $\sim 3$ between $z = 3.5$ and 2.25. By comparing with the large number of “ordinary” quasars, Murphy & Liske (2004) show DLAs to be relatively dust free at high redshift (see Chapter 5).

Moving to lower gas column densities Nestor et al. (2005) present a catalogue of 1331 Mg II absorption line systems in the SDSS EDR, an order of magnitude more than in any previous survey. They found the incidence of intermediate strength absorbers to remain constant over a lookback time of some 6 Gyr ($0.366 \leq z \leq 2.269$), presenting a challenge to theories of the nature of the absorbers. The incidence of strong line systems is found to decrease with decreasing redshift, perhaps due to a decrease in major merger rate or occurrence of superwinds.

## 2.3 Principal Component Analysis

Turning now to a technique used throughout this thesis to aid in the analysis of the SDSS spectra, and that lies behind the derivation of the 2dFGRS spectral types used in Chapter 3, this section presents the concepts and mathematics behind principal component analysis (PCA), a standard multivariate statistical technique used in many quantitative science fields. While the hundreds of thousands of moderate resolution galaxy and quasar spectra in the SDSS provide a staggering amount of information on the nature and evolution of the galaxies in the local universe, extracting and analysing that information in some optimal way is increasingly becoming a challenge in computational resources and statistical methods. Traditionally our understanding of the physical processes occurring in galaxies has come from concentrating on small spectral regions, particular lines, single features or broad band colours, simply due to the limitations inherent in the observations. With high quality spectra, extending more than 5000Å in the rest frame, combined analyses of multiple features are possible. With part of the advantage of the SDSS catalogue being in its sheer size, traditional astronomical packages requiring large amounts of user interaction are obsolete for many investigations and new methods must be developed.

Modern statistical analysis techniques can study the entire spectral range and be used to uncover the "most important" regions of spectra, based either upon spectral synthesis models or correlations in the data with no need for prior physical assumptions. However, consideration must be given to the computational resources potentially required for such analyses: for example to place the 565 715 galaxy spectra of SDSS DR4 into a single floating point array would require over 8Gb of RAM.

PCA can contribute in several way to the analysis of spectroscopic datasets:

- Data compression: picking out the true variations in the data from the noise and hence reducing the dimensionality of the dataset substantially.
- Reconstruction: filling in missing pixels due to errors or recovering continua effected by physical processes in a minority of objects.
- Classification: providing an "unsupervised" method without the need for training sets or model input.

The first two applications are directly applied to spectra in this thesis, the third has been used to derive the spectral type of 2dFGRS galaxies given in the catalogue.

### 2.3.1 PCA in Astronomy

Within the last couple of decades PCA has become well known in astronomy, particularly with respect to the classification of objects (e.g. galaxies: Connolly et al. 1995; Folkes et al. 1996; Madgwick et al. 2002, 2003c; Yip et al. 2004a. Quasars: Francis et al. 1992; Yip et al. 2004b. Stars: Whitney 1983). Other authors have employed PCA for data reduction, for example the PCAz method of Glazebrook et al. (1998), and reconstruction problems such as in the Lyman-$\alpha$ forest (Suzuki et al. 2005).

#### *Advantages and limitations of PCA*

In order to know whether PCA can prove beneficial in a particular situation we must first understand the method. The mathematical formalism is presented in the following subsection, but it is simple to visualise. An array of $N$ spectra each containing $M$ pixels can alternatively be imagined as each spectrum being represented by a point in an $M$ dimensional space. PCA searches for the lines of greatest variance in the cloud of points representing the spectra. Each line is constrained to be orthogonal to all those previous, therefore a basis set is picked out in which the eigenspectra are ordered according to their relative contribution to the overall variance of the dataset.

PCA's sole objective is to identify correlations in datasets and this makes it ideally suited to the procedure developed in Chapter 4 to remove OH sky residuals from SDSS spectra. The SDSS sky residuals are highly correlated both in wavelength space and between fibres making PCA the method of choice for characterising them which in turn enables their subtraction. The fact that PCA has no knowledge of "physical information" means that physical features which only occur in a few spectra will appear in much later eigenspectra or be lost in the noise; reconstructions that make use of only the first few eigenspectra will fail to follow such features. This is exploited in Chapter 4 where BALs are identified through their poorly fitting PCA reconstructions in the spectral regions where their broad absorption troughs lie. By isolating the unquantifiable continuum variations, the unsupervised nature of PCA allowed the development of a spectral classification scheme for 2dFGRS spectra, often considered useful only for their redshifts (Madgwick et al. 2002). Until the much later measurement of colour from the original UK Schmidt Telescope plates scanned by the SuperCOSMOS machine, the 2dF spectral type (termed $\eta$) provided the simplest way by which the galaxies could be split according to their properties.

It is, however, important to realise the limitations of PCA alongside any potential advantages. As an "unsupervised" statistical method, PCA can retrieve information that we may not know is there, or that we are unable to characterise with a physical model prior to analysis of the data. However, for classification and the retrieval of physical information such as galaxy SFRs, PCA has gained a reputation in astronomy for being difficult to interpret precisely because of its lack of knowledge about physical processes. One alternative technique, MOPED (Heavens et al. 2000), overcomes this problem by inputting model spectra from the start, but the results are then immediately restricted to the physics contained in the models. In reality, there is no single ideal technique.

The "linearity" of PCA can also present a problem for galaxy, quasar and stellar classification as the physical processes contributing to the spectra are far from linear. In other words, a banana shaped cloud of points would best be represented by a single curve but PCA is confined to straight lines. One technique which lifts this restriction is Independent Component Analysis (ICA) used often in CMB analysis (e.g. Maino et al. 2002) but the method is largely untried on astronomical spectra.

### 2.3.2 The mathematics of PCA

The standard formalism of PCA presented here is compiled from Kendall (1975), Efstathiou & Fall (1984) and Murtagh & Heck (1987). In all that follows vectors are represented by a single underscore, two dimensional matrices by a double underscore.

$N$ mean subtracted spectra each $M$ pixels long are placed in a data array $\underline{\underline{X}}$ with elements $X_{ij}$, where $1 \leq i \leq N$, $1 \leq j \leq M$. The elements of the covariance matrix ($\underline{\underline{C}}$) of this data array are given by:

$$C_{jk} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{X_{ij} X_{ik}}. \tag{2.1}$$

The covariance matrix can be decomposed into an eigenbasis, described by a set of eigenvectors ($\{\underline{e}\}$, principal components or eigenspectra in the language of this thesis) each $M$ pixels long:

$$\underline{\underline{C}}\underline{e}_j = \lambda_j \underline{e}_j \tag{2.2}$$

where $\lambda_j$ are the eigenvalues and $j$ identifies the eigenvector. Note that these are orthogonal unit vectors:

$$\underline{e}_j^T \underline{e}_k \equiv \sum_i e_{ji} e_{ki} = \delta_{jk} \tag{2.3}$$

where $^T$ represents the transpose. It can be shown that $\underline{e}_1$ is the axis along which the variance is maximal, $\underline{e}_2$ is the axis with the second greatest variance, and so on until $\underline{e}_M$ has the least variance. The principal component amplitudes for each input spectrum $\underline{f}$ are given by

$$a_j = \underline{f}^T \underline{e}_j. \tag{2.4}$$

Reconstruction of the spectrum is achieved by multiplying the principal component amplitudes by their respective eigenvectors:

$$\underline{f} = \sum_{j=1}^{M} a_j \underline{e}_j . \tag{2.5}$$

In general as the variance of the eigenvectors decrease, so does the useful information contained in the spectra, hence making PCA a useful form of data compression: in equation (2.5) we would sum from $j = 1$ to $m$ where $m << M$, hence reducing the dimensionality of each spectrum from $N$ to $m$. The exact number of eigenvectors required to reconstruct the input spectrum is unconstrained, and decided upon based on the dataset and purpose of analysis. The $M$ or $m$ eigenvectors can be used as a basis set upon which to project *any* spectrum ($\underline{f}$) of the same dimensions. The reconstructed spectrum only contains information present in the eigenvectors, which may not be a fair representation of the spectrum if the spectrum were not used during creation of the eigenvectors and/or fewer than $M$ eigenvectors are used during reconstruction.

Various extensions and adaptations have been formulated for this standard procedure. For the analysis of spectra covering different rest frame wavelength regions the technique of “gappy PCA” is introduced. This routine requires a quantification of the difference between sets of eigenvectors and a convergence criteria is presented. The computational speed of PCA depends on the number of pixels in the spectra and for spectra such as in the SDSS it can become necessary to speed up the process by using an Expectation Maximisation procedure. All three methods used in this thesis are outlined below, together with derivations and proofs where appropriate[2].

### *Gappy PCA*

In general, datasets contain missing data values; for spectra of galaxies and quasars this results from spectra covering different rest frame wavelength ranges. Connolly & Szalay (1999) pioneered the technique of “gappy PCA” for astronomical spectra allowing each pixel to be weighted arbitrarily when calculating the principal component amplitudes of a spectrum from the eigenspectra.

Gappy PCA states that the optimal principal component amplitudes for a spectrum, $\underline{f}$, are given by

$$a_j = \sum_k M^{-1}_{jk} F_j \tag{2.6}$$

where

$$\begin{aligned} M_{jk} &\equiv \sum_i w_i e_{ji} e_{ki} && (2.7) \\ F_k &\equiv \sum_i w_i f_i e_{ki} && (2.8) \end{aligned}$$

where $w_i = 1/\sigma_i^2$ are the weights for each pixel based on the pixel errors, $\sigma_i$, and $w_i = 0$ in regions of missing data. The sum is over the $i$ pixels of the eigenvectors and spectrum.

The derivation of gappy PCA is based upon least squares minimisation. We wish to minimise the

[2] Many of the proofs were completed with much assistance from Daniel Mortlock.

$\chi^2$ between the input spectrum and the reconstructed spectrum with respect to the principal component amplitudes:

$$\frac{\partial \chi^2}{\partial a_j} = \sum_i w_i \frac{\partial}{\partial a_j}(f_i - \sum_k a_k e_{ki})^2 \tag{2.9}$$

$$= -2\sum_i w_i e_{ji}(f_i - \sum_k a_k e_{ki}) = 0. \tag{2.10}$$

Rearranging and substituting:

$$\sum_i w_i f_i e_{ji} = \sum_i w_i e_{ji} \sum_k a_k e_{ki} \equiv F_j \tag{2.11}$$

$$= \sum_k \sum_i w_i e_{ji} e_{ki} a_k \tag{2.12}$$

$$= \sum_k M_{jk} a_k \tag{2.13}$$

and solving for $a$

$$a_j = \sum_k M^{-1}_{jk} F_j. \tag{2.14}$$

To summarise the procedure for creating eigenvectors using gappy PCA:

1. Shift all galaxies to their rest frame.

2. Replace missing pixels by the mean of that pixel for all other spectra that contain data.

3. Perform PCA to calculate the eigenvectors, $\underline{e}_j$ (previous subsection).

4. For each spectrum use gappy PCA to calculate the principal component amplitudes, $a_j$, weighting by the errors on each pixel $w_i = 1/\sigma_i^2$ and setting $w_i = 0$ in regions of no data.

5. Use equation (2.5) to reconstruct the input spectrum from the eigenvectors.

6. Use the reconstructions to provide data values for pixels with no data and return to part 3, iterating until convergence (see following subsection).

### *Convergence criteria*

When performing the iterations required during gappy PCA it is necessary to have a measure of how different the new set of eigenvectors are compared to the previous set, and so know when the process has converged. Yip et al. (2004a), based on a similar statement in Everson & Sirovich (1995), states that two subspaces ($\underline{\underline{E}}$ and $\underline{\underline{F}}$) are in common if

$$Tr(\underline{\underline{E}}\,\underline{\underline{F}}\,\underline{\underline{E}}) = D \tag{2.15}$$

where $D$ is the dimensionality of the subspace and $\underline{\underline{E}}$ is given by the sum of the outer products of the eigenvectors:

$$\underline{\underline{E}} = \underline{e}_i \underline{e}_i^T = \sum_{i=1}^{D} e_{ij} e_{ik} \tag{2.16}$$

which is stated to be equivalent to the sum of the projection operators of the space described by $\underline{\underline{E}}$:

$$\underline{\underline{E}} = \sum_{i=1}^{D} \underline{\underline{P}}_i \tag{2.17}$$

When two subspaces are disjoint, equation (2.15) is equal to zero hence allowing the expression to be used as a convergence criterion.

As the reasoning behind these statements was far from transparent and no further clarification was to be found in the literature or from google, this subsection provides some proofs and explanations. Conceptually it can be imagined as follows. A set of $m$ eigenvectors, where $m < M$, describes a subspace of the total space spanned by the $M$ eigenvectors. For the iterations to converge we require the $m$ eigenvectors to describe the same subspace between one iteration and the next. By projecting the eigenvectors describing one subspace onto the subspace described by the second set we can see that if the eigenvectors describe different subspaces (they are disjoint) the result will be zero. It is proved below that, if the subspaces are the same, the result equals the dimensionality of the subspaces.

Firstly, a projection operator (projector) collapses a matrix along one dimension and is constructed by forming outer products of unit vectors. An important property of projectors is that upon application of the same projector no further effect is had on the matrix:

$$\underline{\underline{P}}^2 \ \underline{a} = \underline{\underline{P}}^n \ \underline{a} = \underline{\underline{P}} \underline{a} \tag{2.18}$$

and any matrix with this property is a projection operator. It can be proved that $\underline{\underline{E}}$, the sum of projection operators defined by the eigenvectors [equation (2.16)], is a projection operator as follows:

$$\begin{aligned}
\underline{\underline{E}}^2 &= \sum_l E_{jl} E_{lk} = \sum_l (\sum_i e_{ij} e_{il})(\sum_{i'} e_{i'l} e_{i'k}) && (2.19)\\
&= \sum_{ii'} e_{ij} e_{i'k} \sum_l e_{il} e_{i'l} && (2.20)\\
&= \delta_{ii'} \sum_{ii'} e_{ij} e_{i'k} && (2.21)\\
&= \sum_i e_{ij} e_{ik} && (2.22)\\
&= \underline{\underline{E}} && (2.23)
\end{aligned}$$

Finally, given that $\underline{\underline{E}}$ is a projection operator it is now trivial to prove equation (2.15) for the case that

$\underline{\underline{E}} = \underline{\underline{F}}$:

$$\begin{aligned} Tr(\underline{\underline{EEE}}) &= Tr(\underline{\underline{E}}^3) && (2.24) \\ &= Tr(\underline{\underline{E}}) = \sum_j E_{jj} && (2.25) \\ &= \sum_j \sum_i e_{ij} e_{ij} = \sum_j \delta_{jj} = D && (2.26) \\ & && (2.27) \end{aligned}$$

### *Expectation Maximisation Algorithm for PCA*

Computation of the data covariance array scales as $\sim O(NM^2)$, where $N$ is the number of data vectors and $M$ the size of the vectors, solving for the eigenvectors scales as $\sim O(M^3)$ (Hand et al. 2001). The computation of PCA for large $N$ datasets is simply a problem of memory, while for large numbers of data points another method must be found. One such method is the application of an Expectation Maximisation (EM) algorithm to solve only for those eigenvectors required without computing the covariance matrix.

A full description of the EM algorithm is beyond the scope of this thesis, but it can be visualised in the following way. We wish to find only the first $k$ eigenvectors spanning the subspace described by the data, that account for the majority of the variance. In the first instance the orientation of this subspace is chosen at random, the data are projected onto this guessed subspace and the hidden states (the principal component amplitudes we wish to recover) calculated. Using these values a reconstruction error can be calculated - the squared difference between the data and their reconstructions from the basis spanned by the subspace and the hidden states. A new subspace is then found which minimises these squared reconstruction errors and the process repeated, converging towards a set of eigenvectors which describe the subspace containing the maximum variance of the dataset.

The EM algorithm scales particularly favourably with traditional PCA algorithms when the number of eigenvectors ($k$) required is small compared to the size of the dataset, scaling as $O(kNM)$. To give an example of why this is necessary, the SDSS spectra contain of order 4000 pixels. With the large redshift range covered by the quasars and the application of gappy PCA, this number can easily be extended by 50%. Say we wish to compute the eigenvectors for 5000 quasars over 4000 pixels, but only require the first 10 eigenvectors. EM PCA would be a factor of ∼70 faster if it were to require 10 iterations to converge.

Full details of the EM algorithm used in this thesis can be found in Roweis (1997). Tipping & Bishop (1999) show that the algorithm converges to the correct solution. An online tutorial provides a useful summary of the method (H. Chen, http://www.caip.rutgers.edu/riul/research/tutorials/tutorialrpca.pdf).

*"There is an answer?" said Lunkwill*
*"A simple answer?" said Fook ...*
*"Yes" said Deep Thought.*
*"Life, the Universe, and Everything.*
*There is an answer. But" he added*
*"I'll have to think about it."*

Douglas Adams
The Hitch Hiker's Guide to the Galaxy

# 3
# Relative Biasing between Galaxy Populations

*It is well known that the clustering of galaxies depends on galaxy type. Such* **relative bias** *is a challenge to theories of galaxy formation and evolution and complicates the inference of cosmological parameters from galaxy redshift surveys. This chapter makes use of the 2dF Galaxy Redshift Survey, which is ideally suited to large scale structure analyses, to study the difference between the spatial distributions of red and blue, and early- and late-type galaxies.*

## 3.1 Introduction

The question of whether galaxies trace the matter distribution of the universe has many implications for cosmology and galaxy formation theories. Since Hubble & Humason (1931) it has been known that galaxies of different type cluster differently, and as such it cannot be possible for all galaxies to trace the matter distribution exactly. This observation has been reconfirmed many times, traditionally by comparisons of the correlation functions of different subgroups. For example, early type (or passive) galaxies are more strongly clustered than late type (or actively star-forming) galaxies (e.g. Davis & Geller 1976; Dressler 1980; Lahav, Nemiroff & Piran 1990; Hermit et al. 1996; Norberg et al. 2002b; Zehavi et al. 2002; Madgwick et al. 2003b) and luminous galaxies cluster more strongly than faint galaxies (e.g. Willmer, da Costa & Pellegrini 1998; Norberg et al. 2001, 2002b; Zehavi et al. 2002, 2004). The problem has been phrased in terms of "galaxy bias" describing the difference in distribution between the galaxy and matter distribution, with the implication in terminology that galaxies are biased tracers of the mass distribution.

In recent years much effort has been put into investigating galaxy bias through theory and numerical modelling, while observational results have been restricted by small survey volumes. In common with subsequent chapters, it is the advent of large galaxy redshift surveys that is making possible the quantification of galaxy properties and their spatial distribution as never before, providing detailed descriptions with which to compare theoretical and numerical models. This chapter makes use of the 2dF Galaxy redshift survey for understanding the spatial distribution of galaxies, a task for which the survey is particularly suited. In contrast to subsequent chapters, which make use of the improved spectra of the SDSS survey, each galaxy is described by only a few parameters, namely position, redshift and colour or spectral type.

Quantifying galaxy bias is of interest in relation to an understanding of galaxy formation and evolution. Bias is also a major practical source of uncertainty in deriving cosmological constraints from galaxy surveys. Some particular examples are the measurement of $\beta = \Omega_m^{0.6}/b$ (Peacock et al. 2001; Hawkins et al. 2003), where Dekel & Lahav (1999, hereafter DL99) showed that stochastic effects could explain large discrepancies between results from different methods (for a review see Dekel & Ostriker 1999, Table 7.2). Power spectrum measurements require constant bias as a fundamental assumption (Percival et al. 2001), and constraints placed on neutrino mass also assume scale independent biasing (Elgarøy & Lahav 2003). The importance of biasing has increased still further with the release of the WMAP first year results (Spergel et al. 2003). In order to combine CMB and 2dFGRS data to give tighter constraints on cosmological parameters, a model for galaxy bias is required (Verde et al. 2003).

### 3.1.1 Describing bias

It is common in cosmology to describe a density field in terms of the "overdensity perturbation", $\delta(r)$, smoothed over a given scale, $r$:

$$1 + \delta(r) = \frac{\rho}{\langle\rho\rangle} = \frac{N}{\langle N\rangle} \tag{3.1}$$

where $\rho$ is the density of the field, and $\langle\rho\rangle$ is the "expected" density for a region of that size if the Universe were isotropic and homogeneous[1]. Alternatively, for discrete objects such as galaxies $N$ is the number of objects in a region of size $r$ and $\langle N\rangle$ is the expected number in that region. Bias can then be described as the relation between two overdensity perturbation fields.

In principle the form of bias should be derivable from the fundamental physical processes involved in galaxy formation; until we understand these, bias remains a description of our ignorance. The simplest model of galaxy biasing is the linear biasing model

$$\delta_g(\mathbf{x}) = b\delta_m(\mathbf{x}) \tag{3.2}$$

where $\delta_g$ is the galaxy overdensity perturbation, $\delta_m$ the mass overdensity perturbation and $b$ a constant bias parameter. Because galaxies are discrete objects, this prescription is normally supplemented by the *Poisson Clustering Hypothesis*, in which galaxies are modelled as random events, whose expectation

[1] Strictly this should be the "ergodic" average over many Universes, however it is more practically approximated as the average over the region of the Universe available in the survey

number density is specified via $\delta_g$. This model for discreteness can only be an approximation, but there is no simple alternative. Poisson sampling is therefore assumed in what follows; for consistency, theoretical predictions are treated in the same way as the real data.

The linear biasing model is unphysical for $b > 1$, as it allows negative galaxy densities in regions with low matter densities. Alternative models in the literature fall into several basic classes: linear or non-linear, local or non-local, deterministic or stochastic. Locally biased galaxy formation (e.g. Coles 1993; Scherrer & Weinberg 1998; Fry & Gaztanaga 1993) depends only on the properties of the local environment, and the galaxy density is assumed to be a universal function of the matter density:

$$\delta_g = f(\delta_m). \tag{3.3}$$

Non-local models (e.g. Bower et al. 1993; Matsubara 1999) arise when the efficiency of galaxy formation is modulated over scales larger than those over which the matter moves, for example by effects of quasar radiation on star formation. Stochastic bias (Pen 1998, DL99) allows a range of values of $\delta_g$ for a given $\delta_m$, above the Poissonian scatter caused by galaxy discreteness. Stochasticity is a natural part of non-local models, but some stochasticity is always expected to arise from physical processes of galaxy formation (Blanton et al. 1999).

DL99 set out a general framework for describing non-linear stochastic biasing, which allows the two effects to be naturally separated and it is this framework which is followed throughout this chapter. Full details are given in Section 3.2.1.

### 3.1.2 Linking mass and light

Three independent methods have previously been used to investigate galaxy biasing in the 2dFGRS catalogue. Lahav et al. (2002) combined pre-WMAP CMB and 2dFGRS datasets to measure the average bias over a range $0.02 < k < 0.15\,h\,\mathrm{Mpc}^{-1}$, concluding that galaxies are almost exactly unbiased on these scales. Verde et al. (2002) found the bias parameter to be consistent with unity over scales $0.1 < k < 0.5\,h\,\mathrm{Mpc}^{-1}$ through measurements of the 2dFGRS bispectrum.

A more direct method of studying the relation between mass and light is to map the dark matter using gravitational lensing. This field has made great progress in recent years, and it has been possible to measure not only the absolute degree of bias, but also its nonlinearity and stochasticity (Fischer et al. 2000; Hoekstra et al. 2002; Fan 2003; Pen et al. 2003). For example, Hoekstra et al. (2002) combine the Red-Sequence Cluster Survey and VIRMOS-DESCART survey to find an average bias $b = 0.71$ and linear correlation coefficient of $r \simeq 0.57$ on scales of $1 - 2\,h^{-1}\,\mathrm{Mpc}$. However, current weak lensing measurements are dominated by non-linear and quasi-linear scales in the power spectrum, and it is not yet possible to say a great deal about bias in the very large-scale linear regime. This is of course the critical region for the interpretation of redshift surveys in terms of cosmological parameters, where we want to know the relation between the power spectra of mass and light on$\gtrsim 100\,\mathrm{Mpc}$ scales.

This question will be settled by future weak lensing surveys. In the meantime, we can address a related simpler problem: the *relative* bias between subsets of galaxies. The morphological differences between galaxies and the link to their environments has been discussed for many decades as a potential clue to the nature and evolution of galaxy clustering (e.g. Spitzer & Baade 1951; Gunn & Gott 1972;

Davis & Geller 1976; Yoshikawa et al. 2001). Modern galaxy redshift surveys allow us to split the galaxy population into a variety of subdivisions such as spectral type, colour and surface brightness. We can look at relative bias as a function of scale, and weighted by luminosity, which should yield important insights into the absolute degree of bias that may exist. Norberg et al. (2001) measured bias as a function of luminosity in the 2dFGRS, finding a bias relative to $L^*$ galaxies of $b/b^* = 0.85 + 0.15L/L^*$.

This chapter exploits the natural bimodality of the galaxy population, between red early types with little active star formation, and the blue late type population (e.g. Baldry et al. 2003). Lahav & Saslaw (1992) measured bias as a function of morphological type and scale using the UGC, ESO and IRAS catalogues. The Las Campanas Redshift Survey has already provided some observational evidence against the linear deterministic model from splitting galaxies by their spectral types (Tegmark & Bromley 1999; Blanton 2000), and this chapter presents a more extensive analysis of this type.

### 3.1.3 Measurements of relative biasing

There are several complementary methods for the measurement of galaxy clustering, although most previous studies of the relative bias between galaxy types have concentrated on a relative bias parameter defined as the square root of the ratio of the correlation functions for the types under study:

$$b = \sqrt{\frac{\xi_1^2}{\xi_2^2}} \tag{3.4}$$

where the correlation function is given by

$$\xi(\underline{r}) \equiv \langle \delta(\underline{x}) \delta(\underline{x} + \underline{r}) \rangle. \tag{3.5}$$

Madgwick et al. (2003b) used this method to measure the relative bias in the 2dFGRS, finding $b$ (passive/active) ranging from about 2.5 to 1.2 on scales $0.2\,h^{-1}\,\mathrm{Mpc} < r < 20\,h^{-1}\,\mathrm{Mpc}$. However, even within such a large survey as the 2dFGRS the correlation functions become noisy beyond about $10\,h^{-1}\,\mathrm{Mpc}$.

A second method is counts-in-cells, where cells are placed across the survey volume and the number of galaxies within each cell counted. The average number of galaxies in all the cells provides an estimator for the expected number, $\langle N \rangle$, and a value for the overdensity perturbation of each cell is calculated using equation 3.1. The moments of the counts-in-cells overdensity perturbations can be directly related to the correlation functions (Peebles 1980), for example the second moment is given by

$$\mu_2 = \langle (N - n\,V)^2 \rangle = n\,V + n^2\,V \int \xi(r)\,d^3r \tag{3.6}$$

where $N$ is the number of galaxies in a given cell, $n$ the average number density and $V$ the volume of the cell. Counts-in-cells is optimised for the study of larger scales compared to the correlation function.

The bias can be obtained either directly from the counts-in-cells via non-parametric calculation of moments and correlation coefficients, or via fits of probability distributions to the data. With either method it is necessary to account for the Poisson sampling of the overdensity field by the galaxies, and it is the latter method that is employed in this chapter. Conway et al. (2004) have also investigated the relative bias of different galaxy types using a counts-in-cells analysis of the 2dFGRS, but they use magnitude limited samples, and consider only deterministic bias models, whereas this analyses uses volume limited samples, and considers stochastic bias models. The counts-in-cells method has also been used to calculate the variance and higher order moments of galaxy clustering in the 2dFGRS (Conway et al. 2004; Croton et al. 2004a,b; Baugh et al. 2004).

Many theoretical results on biasing from numerical models have also been reported. There are two main approaches to modelling galaxy distributions: semianalytic (e.g. Kauffmann, Nusser & Steinmetz 1997; Benson et al. 2000; Somerville & Primack 1999) and hydrodynamic (e.g. Cen & Ostriker 1992; Blanton et al. 1999; Cen & Ostriker 2000; Yoshikawa et al. 2001). Comparisons are given by Helly et al. (2003) and Yoshida et al. (2002). Several studies have been made of galaxy biasing in these numerical simulations (e.g. Somerville et al. 2001a; Yoshikawa et al. 2001), but none provide results in sufficient detail to allow an easy comparison with the 2dFGRS. A large new semianalytic calculation is therefore analysed in this chapter which is capable of yielding mock results that can be compared directly to the 2dF data.

This chapter concentrates on a few aspects of relative bias mentioned above, splitting galaxies by spectral type and colour. The nonlinearity, stochasticity and scale dependence of the biasing relation is investigated through comparison with three models. Throughout, a cosmological geometry with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ is adopted in order to convert redshifts and angles into three-dimensional comoving distances.

## 3.2 Modelling Relative Bias

The simplest model for any bias (i.e. mass-galaxy, early-late, red-blue etc.) is that of *linear deterministic* bias: given a number of one type of object you can predict precisely (within Poisson errors) the number of the other type of object in the same region of space, and the relationship between the two numbers is linear. Recalling the relation for the mass/galaxy distributions $\delta_g = b\delta_m$, we can write

$$\delta_{\rm L} = b\delta_{\rm E} \tag{3.7}$$

where $\delta_{\rm E}$ ($\delta_{\rm L}$) denotes the overdensity of early (late) type galaxies in a volume of space. As described above, this empirical model can become unphysical in low density regions. Considering the complex processes involved in galaxy formation, it would be surprising to find linear deterministic biasing to be true in all cases. Any reasonable physical theory in fact predicts non-trivial mass/galaxy biasing (Cole & Kaiser 1989) and simulations can also find biasing to be a complicated issue particularly on small scales (Cen & Ostriker 1992; Blanton et al. 1999; Somerville et al. 2001a).

This chapter investigates two potential improvements to the linear deterministic model. Firstly the bias could be nonlinear, and some nonlinearity is inevitable in order to ‘fix’ the unphysical properties of the linear model. Secondly, there may exist stochasticity (scatter beyond Poissonian discreteness noise), due to astrophysical processes involved in galaxy formation. DL99 present a general framework to quantify these different aspects of biasing, and the following section summarises their results.

### 3.2.1 A framework for nonlinear, stochastic bias

The overdensity of one field can be related to that of a second field contained in the same volume of space through

$$\delta_1 = b(\delta_2)\delta_2 + \epsilon. \tag{3.8}$$

The scatter (stochasticity) in the relation is given by

$$\epsilon \equiv \delta_1 - \langle\delta_1\rangle. \tag{3.9}$$

In principle the bias parameter $b$ can be any function of $\delta_2$; a constant value of $b$ and $\epsilon = 0$ represents deterministic linear biasing.

The notation $f(\delta_{\rm E})$ and $f(\delta_{\rm L})$ is used to denote the one-point probability distribution functions (PDFs) of the overdensity perturbations of early and late type galaxies. The fields $\delta_{\rm E}$ and $\delta_{\rm L}$ have zero mean by definition, and their variances are defined by

$$\sigma_i^2 \equiv \int_{-1}^{\infty} \delta_i^2 f(\delta_i) d\delta_i \equiv \langle\delta_i^2\rangle. \tag{3.10}$$

The joint underlying probability distribution of early and late type galaxies is given by

$$\begin{aligned} f(\delta_{\rm E}, \delta_{\rm L}) &= f(\delta_{\rm E}) f(\delta_{\rm L}|\delta_{\rm E}) && (3.11)\\ &= f(\delta_{\rm L}) f(\delta_{\rm E}|\delta_{\rm L}). && (3.12)\end{aligned}$$

Both equations (3.11) and (3.12) should give the same outcome, but choosing to work with equation (3.11) avoids values of $b > 1$ and therefore unphysical linear biasing.

The natural generalisation of linear biasing is given by

$$b(\delta_{\rm E})\delta_{\rm E} \equiv \langle\delta_{\rm L}|\delta_{\rm E}\rangle = \int f(\delta_{\rm L}|\delta_{\rm E})\delta_{\rm L} d\delta_{\rm L}. \tag{3.13}$$

There are several useful statistics that can be derived from equantion (3.13) and used to investigate independently the fraction of nonlinearity and stochasticity of a model or data. Firstly the *mean biasing* is defined by

$$\hat{b} \equiv \frac{\langle b(\delta_{\rm E})\delta_{\rm E}^2\rangle}{\sigma_{\rm E}^2}; \tag{3.14}$$

the nonlinear equivalent of this is

$$\tilde{b}^2 \equiv \frac{\langle b^2(\delta_{\rm E})\delta_{\rm E}^2\rangle}{\sigma_{\rm E}^2}. \tag{3.15}$$

In each case the denominator is assigned such that linear biasing reduces to $b = \hat{b} = \tilde{b}$. The *random biasing field* is defined as

$$\boldsymbol{\epsilon} \equiv \delta_{\rm L} - \langle\delta_{\rm L}|\delta_{\rm E}\rangle \tag{3.16}$$

and the statistical character of the biasing relation can be described via its variance, the *biasing scatter function*

$$\sigma_b^2(\delta_{\rm E}) \equiv \frac{\langle \boldsymbol{\epsilon}^2 | \delta_{\rm E} \rangle}{\sigma_{\rm E}^2}. \tag{3.17}$$

The *average biasing scatter* is then given by

$$\sigma_b^2 \equiv \frac{\langle \boldsymbol{\epsilon}^2 \rangle}{\sigma_{\rm E}^2}. \tag{3.18}$$

The purpose of this parameterisation is to separate naturally the effects of nonlinearity and stochasticity, allowing them to be quantified via the relations

$$\textbf{nonlinearity} \equiv \tilde{b}/\hat{b} \tag{3.19}$$

$$\textbf{stochasticity} \equiv \sigma_b/\hat{b}. \tag{3.20}$$

There are two relations that are often quoted in the literature as measures of the bias parameter and stochasticity: the ratio of variances

$$b_{\rm var} \equiv \frac{\sigma_{\rm L}}{\sigma_{\rm E}} = \sqrt{(\tilde{b}^2 + \sigma_b^2)} \tag{3.21}$$

and the linear correlation coefficient

$$r_{\rm lin} \equiv \frac{\langle \delta_{\rm E} \delta_{\rm L} \rangle}{\sigma_{\rm E} \sigma_{\rm L}} = \frac{\hat{b}}{b_{\rm var}}. \tag{3.22}$$

As indicated, both $b_{\rm var}$ and $r_{\rm lin}$ can be written in terms of the basic parameters given above and both mix nonlinear and stochastic effects.

### 3.2.2 Redshift-space distortions

Note that this chapter works throughout with redshift-space overdensities. Redshift-space distortions are dependent on galaxy type due to the different clustering properties of early and late type galaxies. On nonlinear scales the dominant effect is the finger-of-god stretching, but on the scales of interest to this paper the linear $\beta$-effect is expected to apply (Kaiser 1987). Averaging over all angles and including stochasticity between the galaxy and matter fields, we can write the redshift-space power spectrum, $P_{\rm s}$, as

$$\frac{P_{\rm s}}{P_{\rm r}} = (1 + \frac{2}{3} r\beta + \frac{1}{5}\beta^2) \tag{3.23}$$

where $\beta = \Omega_m^{0.6}/b$, $b$ is the mass-galaxy bias, $r$ is the linear mass-galaxy correlation coefficient and $P_{\rm r}$ is the real-space power spectrum (Pen 1998; Dekel & Lahav 1999). The $\beta$-effect was measured for galaxies of different spectral class in the 2dFGRS by Madgwick et al. (2003b), obtaining $\beta_{\rm L} = 0.49 \pm 0.13$ and

$\beta_{\rm E} = 0.48 \pm 0.14$. From these results and assuming $r = 1$ we obtain

$$\frac{P_{\rm s,E}}{P_{\rm s,L}} = 0.99 \frac{P_{\rm r,E}}{P_{\rm r,L}}. \tag{3.24}$$

Although this suggests the effect is not large and currently insignificant within the errors, it is clear that in the case of zero stochasticity, redshift-space distortions will work to reduce the difference in the measured clustering between types. However, including a value of $r$ which is non-unity and dependent on galaxy type as suggested by simulations, has a significant effect. For example, taking $r_{\rm L} = 0.8$ and $r_{\rm E} = 1.0$ increases the relative distortion from 0.99 to 1.04, where $r_{\rm L}$ ($r_{\rm E}$) is the linear correlation coefficient between the mass and late (early) type galaxy fields.

### 3.2.3 One point probability distribution function

Given equation (3.11) we can create a model in two parts: firstly the distribution of early type galaxies per cell, and secondly the biasing relation connecting the two distributions. In this section we create a description of the one-point distribution of the early type galaxies, and in the following section we present the formalism with which this is combined with the biasing relations to create the two-point distribution function.

A standard description for the underlying probability distribution of a galaxy overdensity, $f(1 + \delta)$, is lognormal (Coles & Jones 1991). Applying this for example to the early type galaxies:

$$f(\delta_{\rm E})\, d\delta_{\rm E} = \frac{1}{\omega_{\rm E}\sqrt{2\pi}} \exp\left[\frac{-x^2}{2\omega_{\rm E}^2}\right]\, dx \tag{3.25}$$

where

$$x = \ln(1 + \delta_{\rm E}) + \frac{\omega_{\rm E}^2}{2} \tag{3.26}$$

and $\omega_{\rm E}^2$ is simply the variance of the corresponding normal distribution $f[\ln(1 + \delta)]$:

$$\omega_{\rm E}^2 = \langle [\ln(1 + \delta_{\rm E})]^2 \rangle. \tag{3.27}$$

The offset $\omega_{\rm E}^2/2$ is required to impose $\langle \delta_{\rm E} \rangle = 0$. If the lognormal distribution correctly describes the data, the variance of the overdensities, $\langle \delta_{\rm E}^2 \rangle$, is related to the variance of the corresponding Gaussian distribution by

$$\sigma_{\rm E}^2 \equiv \langle \delta_{\rm E}^2 \rangle = \exp[\omega_{\rm E}^2] - 1. \tag{3.28}$$

In Section 3.6.5 it is shown that fitting a lognormal distribution directly to the data does not yield quite the same values for $\sigma_{\rm E}$ and $\sigma_{\rm L}$ as a direct variance estimate. While this does not affect the final results for stochasticity, it causes an overestimate of the nonlinearity in the data on small scales. On the largest scales, a lognormal distribution is completely consistent with the 2dFGRS data, and it provides a transparent and simple way to describe the density field.

### 3.2.4 Biasing models

*Deterministic bias: linear and power law*

Firstly concentrating on deterministic bias, we can write the joint probability distribution function as

$$f(\delta_{\mathrm{L}}|\delta_{\mathrm{E}}) = \delta^{\mathrm{D}}(\delta_{\mathrm{L}} - b(\delta_{\mathrm{E}})\delta_{\mathrm{E}}) \tag{3.29}$$

where $\delta^{\mathrm{D}}$ is the Dirac delta function. This reduces directly to linear bias described by two linear biasing parameters ($b_{\mathcal{L}}$), by setting

$$b(\delta_{\mathrm{E}})\delta_{\mathrm{E}} = b_{0,\mathcal{L}} + b_{1,\mathcal{L}}\delta_{\mathrm{E}} \tag{3.30}$$

where the constraint $\langle\delta_{\mathrm{L}}\rangle = 0$ fixes $b_{0,\mathcal{L}} = 0$. A simple variation could be power law bias described by the power law biasing parameters $b_{\mathcal{L}}$:

$$b(\delta_{\mathrm{E}})\delta_{\mathrm{E}} = b_{0,\mathcal{P}}(1 + \delta_{\mathrm{E}})^{b_{1,\mathcal{P}}} - 1 \tag{3.31}$$

which avoids the negative density predictions of linear bias, and reduces to the linear biasing relation near $\delta = 0$. Rearranging equation (3.31), using the properties of lognormal distributions and the fact that $b_{1,\mathcal{P}} = \omega_{\mathrm{L}}/\omega_{\mathrm{E}}$, we find for power law bias that

$$b_{0,\mathcal{P}} = \exp[0.5\omega_{\mathrm{E}}^2(b_{1,\mathcal{P}} - b_{1,\mathcal{P}}^2)]. \tag{3.32}$$

For convenience we define

$$b_{\mathrm{lin}} = b_{1,\mathcal{L}} \tag{3.33}$$

and

$$b_{\mathrm{pow}} = b_{1,\mathcal{P}} \tag{3.34}$$

throughout the rest of this chapter.

*Stochastic bias: bivariate lognormal*

Returning to equation (3.13), a broader function than the Dirac delta function of equation (3.29) can be introduced. An interesting class of model is when both $\delta$ fields form a bivariate Gaussian distribution (i.e. an ellipsoid in [$\delta_{\mathrm{E}}$,$\delta_{\mathrm{L}}$] space), but this again becomes unphysical for $\delta < -1$ (DL99). It is however simple to cure this defect by assuming instead a *bivariate lognormal* distribution, which describes an ellipsoid in [$\ln(1+\delta_{\mathrm{E}})$,$\ln(1+\delta_{\mathrm{L}})$] space. In the same way as the bivariate Gaussian distribution tends to linear bias as the scatter becomes zero, the bivariate lognormal distribution tends to power-law bias. The joint probability distribution is given by

$$f(g_{\mathrm{E}}, g_{\mathrm{L}}) = \frac{|V|^{-1/2}}{2\pi}\exp\left[-\frac{(\tilde{g_{\mathrm{E}}}^2 + \tilde{g_{\mathrm{L}}}^2 - 2r_{\mathrm{LN}}\,\tilde{g_{\mathrm{E}}}\,\tilde{g_{\mathrm{L}}})}{2(1-r_{\mathrm{LN}}^2)}\right], \tag{3.35}$$

where $g_i = \ln(1+\delta_i) - \langle \ln(1+\delta_i) \rangle$ and $\tilde{g_i} = g_i/\omega_i$, with $i$ corresponding to early or late type. $\omega_i$ is related to the variance of the underlying Gaussian field $\ln(1+\delta_i)$ as for the one–point lognormal distribution:

$$\sigma_i^2 \equiv \langle \delta_i^2 \rangle = \exp[\omega_i^2] - 1. \tag{3.36}$$

The correlation coefficient is

$$r_{\rm LN} = \frac{\langle g_{\rm E} g_{\rm L} \rangle}{\omega_{\rm E}\omega_{\rm L}} \equiv \frac{\omega_{\rm EL}^2}{\omega_{\rm E}\omega_{\rm L}} \tag{3.37}$$

and $V$ is the covariance matrix

$$V = \begin{pmatrix} \omega_{\rm E}^2 & \omega_{\rm EL}^2 \\ \omega_{\rm EL}^2 & \omega_{\rm L}^2 \end{pmatrix}. \tag{3.38}$$

Taking $f[\ln(1+\delta_{\rm E})]$ to be a Gaussian of width $\omega_{\rm E}$ and mean $-\omega_{\rm E}^2/2$ (i.e. $f(\delta_{\rm E})$ is distributed as a lognormal, equation [3.25]), the conditional probability distribution is

$$\begin{aligned} f(g_{\rm L}|g_{\rm E}) &= \frac{f(g_{\rm E}, g_{\rm L})}{f(g_{\rm E})} \\ &= \frac{\omega_{\rm E}}{(2\pi|V|)^{1/2}} \exp\left[-\frac{(\tilde{g_{\rm L}} - r_{\rm LN}\,\tilde{g_{\rm E}})^2}{2(1-r_{\rm LN}^2)}\right], \end{aligned} \tag{3.39}$$

i.e. the distribution of $\tilde{g_{\rm L}}|\tilde{g_{\rm E}}$ is a Gaussian with mean $r_{\rm LN}\,\tilde{g_{\rm E}}$ and variance $1 - r_{\rm LN}^2$.

As $r_{\rm LN} \rightarrow 1$, equation (3.39) reduces to a Dirac delta function, and this bivariate lognormal model reduces to the power law bias model of equation (3.31). It is important to note that $r_{\rm LN}$ is not equal to the linear correlation coefficient $r_{\rm lin}$ of equation (3.22), which can differ from unity even if $r_{\rm LN} = 1$. In this sense, the lognormal parameters offer a cleaner separation of stochastic and nonlinear effects. If stochasticity is present within the data, this model may provide an improvement over the deterministic biasing models. As observational data improve, it may become possible to constrain the relative biasing function to a greater extent; the current data are insufficient for such an analysis.

### *Bias statistics for the Bivariate Lognormal Distribution*

Analytic solutions exist to the mean biasing parameters and biasing scatter function given in Section 3.2.1 for the bivariate lognormal model. In this section these relations are calculated for the general case of two overdensity perturbation fields.

The general biasing relation between the overdensities $\delta_1$ and $\delta_2$ of two subgroups of galaxies, or two types of matter, is fully described by equation (3.13)

$$b(\delta_1)\delta_1 \equiv \langle \delta_2|\delta_1 \rangle = \int f(\delta_2|\delta_1)\delta_2 d\delta_2. \tag{3.40}$$

The conditional probability distribution for the bivariate lognormal model is given by equation (3.39)

$$f(g_2|g_1) = \frac{\omega_1}{(2\pi|V|)^{1/2}} \exp\left[-\frac{(\tilde{g}_2 - r_{\rm LN}\,\tilde{g}_1)^2}{2(1-r_{\rm LN}^2)}\right] \tag{3.41}$$

where $g_i = \ln(1+\delta_i) - \langle\ln(1+\delta_i)\rangle$, $\langle\ln(1+\delta_i)\rangle = -\omega_i^2/2$ and $\tilde{g}_i = g_i/\omega_i$. $\tilde{g}_2|\tilde{g}_1$ follows a univariate Gaussian distribution with mean $r_{\rm LN}\,\tilde{g}_1$ and variance $1 - r_{\rm LN}^2$. The covariance matrix $V$ and correlation coefficient $r_{\rm LN}$ are both defined in log space, and are given explicitly in equations (3.37) and (3.38). The variance of the distribution in linear space $\sigma_i^2$ is related to the variance of the Gaussian field by

$$\sigma_i^2 \equiv \langle\delta_i^2\rangle = \exp[\omega_i^2] - 1. \tag{3.42}$$

On substituting equation (3.41) into (3.40) and integrating we find

$$b(\delta_1)\delta_1 = \exp\left[\omega_2\tilde{g}_1 - \frac{(\omega_2 r_{\rm LN})^2}{2}\right] - 1. \tag{3.43}$$

From this basic parameter we can calculate the mean biasing and its nonlinearity [equations (3.14) and (3.15)]

$$\hat{b} \equiv \frac{\langle b(\delta_1)\delta_1^2\rangle}{\sigma_1^2} = \frac{\exp[r_{\rm LN}\omega_2\omega_1] - 1}{\exp[\omega_1^2] - 1} \tag{3.44}$$

$$\tilde{b}^2 \equiv \frac{\langle b^2(\delta_1)\delta_1^2\rangle}{\sigma_1^2} = \frac{\exp[r_{\rm LN}^2\omega_2^2] - 1}{\exp[\omega_1^2] - 1} \tag{3.45}$$

We also know the ratio of variances, equation (3.21)

$$b_{\rm var}^2 \equiv \frac{\sigma_2^2}{\sigma_1^2} = \frac{\exp[\omega_2^2] - 1}{\exp[\omega_1^2] - 1}. \tag{3.46}$$

Although it is possible to derive the scatter for this model from equation (3.18), it can be shown that (DL99)

$$b_{\rm var}^2 = \tilde{b}^2 + \sigma_b^2. \tag{3.47}$$

Using this fact we obtain

$$\sigma_b^2 = \frac{\exp[\omega_2^2] - \exp[r_{\rm LN}^2\omega_2^2]}{\exp[\omega_1^2] - 1}. \tag{3.48}$$

Whilst the bivariate lognormal model contains constant scatter dependent only on $r_{\rm LN}$ in the log frame, transformation to the linear frame causes the scatter to become dependent on the widths of the univariate distributions and vary with $\delta_1$.

### 3.2.5 Including observational shot noise

It is not possible to measure the underlying probability distribution $f(\delta_{\rm E}, \delta_{\rm L})$ directly due to contamination of the observational data with noise, the dominant form of which is expected to be Poisson or ‘shot’ noise. This can be included in the models of the previous section by convolution with a Poisson distribution (Coles & Jones 1991; Blanton 2000). In this way the measured probability of finding $N_{\rm E}$ early type galaxies and $N_{\rm L}$ late type galaxies within a cell, $P(N_{\rm E}, N_{\rm L})$, can be compared with the models above. Accounting for shot noise in this way results in the models being less sensitive to outliers than equations (3.29) and (3.39).

Using equation (3.11) to combine the one point PDF (3.25) with the conditional PDF (3.29 or 3.39), provides a model for the actual joint probability distribution function $f(\delta_{\rm E}, \delta_{\rm L})$. Convolution with a Poisson distribution then gives

$$\begin{aligned} P(N_{\rm E}, N_{\rm L}) = & \int_{-1}^{\infty}\int_{-1}^{\infty} \frac{\bar{N}_{\rm E}^{N_{\rm E}}(1+\delta_{\rm E})^{N_{\rm E}}}{N_{\rm E}!} e^{-\bar{N}_{\rm E}(1+\delta_{\rm E})} f(\delta_{\rm E}) \\ & \times \quad \frac{\bar{N}_{\rm L}^{N_{\rm L}}(1+\delta_{\rm L})^{N_{\rm L}}}{N_{\rm L}!} e^{-\bar{N}_{\rm L}(1+\delta_{\rm L})} f(\delta_{\rm L}|\delta_{\rm E}) d\delta_{\rm E} d\delta_{\rm L}, \end{aligned} \tag{3.49}$$

where $\bar{N}_{\rm E}$ ($\bar{N}_{\rm L}$) is the expected number of early (late) type galaxies in a given cell, allowing for completeness.

## 3.3 The Data: the 2dFGRS

A description of the 2dFGRS is given in Chapter 2. Here a few more details of particular relevance to this chapter are provided.

For any structure analysis it is important to be aware of several problems that cause varying completeness over a survey region. Common to all similar surveys, some regions of the sky in the 2dFGRS must be masked due to bright stars causing internal holes. Furthermore, due to the adaptive tiling algorithm employed to ensure an optimal observing strategy, the sampling fraction falls to as little as 50% near the survey boundaries and internal holes due to lack of tile overlap. To account for the selection effects within the survey the publicly available redshift completeness masks are used[2]. These are sufficient for galaxies classified by colour, but the spectral type analysis introduces extra selection effects. A difference in completeness over a region of sky could occur, for example, when the spectra of a survey plate are of too poor quality to perform the spectral type analysis, yet redshifts can be obtained. Masks are created for the cells using the publicly available software which accounts for these effects.

Subsequent reanalysis of the photometry of the APM galaxy catalogue has shown the survey depth to vary slightly with position on the sky (Colless et al. 2001). To allow for this, a limiting corrected magnitude for the survey of $B_{\rm J} = 19.2$ is used throughout.

### 3.3.1 Galaxy properties in the 2dFGRS

The 2dFGRS catalogue provides two methods of classification for comparison. Firstly the well studied galaxy spectral type $\eta$, and secondly the photometric colours of the galaxies have recently been measured.

[2] http://magnum.anu.edu.au/∼TDFgg/Public/Release/Masks/

### *Spectral type, $\eta$*

The spectral type of the galaxies has been derived using PCA by Madgwick et al. (2002), (see also Folkes et al. 1999). The spectral type of the 2dFGRS galaxies is characterised by the value $\eta$, a linear combination of the first two principal components, used in order to minimise the effect of distortions and imperfections in the 2dFGRS spectra. Similar to most PCA classifications of galaxies, $\eta$ classifies galaxies according to the average emission and absorption line strength in the spectrum. $\eta$ provides a continuous classification scheme, but for our purposes it is necessary to split the galaxies into two classes, and we adopt a split at $\eta = -1.4$ as suggested in Madgwick (2003). Galaxies with $\eta < -1.4$ are shown to be predominantly passive galaxies and those with $\eta > -1.4$ predominantly star-forming (Madgwick et al. 2003a). The former are hence termed 'early type' (Type 1) and the latter 'late type' (split into Types 2-4 in Madgwick et al. (2002)). The 2dFGRS catalogue contains 74 548 early type and 118 424 late type galaxies defined in this way.

One concern with using optical fibre spectra for this type of analysis are 'aperture effects', resulting from the fixed aperture of the fibres being smaller than the size of galaxies. This could result in, for example, only the bulge components of close spirals being observed. Such effects have been studied in detail by Madgwick et al. (2002), and no systematic bias found. One possible explanation for this is the poor seeing present at the Anglo-Australian Telescope, typically $1.5''$–$1.8''$, which will cause the fibre to average over a large fraction of the total galaxy light in most cases and dilute aperture effects. An overabundance of late type galaxies was detected at redshifts beyond 0.11, which could be attributed to either aperture effects, evolution, or the easier identification of objects with strong emission lines in low signal-to-noise ratio spectra; however, this will not affect our volume limited galaxy sample with a maximum redshift of $z_{\rm max} = 0.114$ (see following section). Aperture effects are discussed further in Section 3.6.2.

### *Broad-band colours*

More recently it has been possible to obtain broad-band colours for the 2dFGRS galaxies using the same $B_{\rm J}$ UKST plates as the survey input catalogue (Hambly et al. 2001), but now scanned with the SuperCosmos machine to yield smaller errors of about 0.09 mag per band. Similar scans have also been made of the UKST $R_F$ plates. External CCD sources, mainly the SDSS-EDR, were used to calibrate the SuperCosmos magnitudes and, while the plates were scanned independently, the $(B_{\rm J} - R_F)$ colour distribution was constrained to be uniform between plates. The extinction corrections are from the dust maps of Schlegel, Finkbeiner & Davis (1998) and wavelength dependent extinction ratios are from Cardelli, Clayton & Mathis (1989). Cross et al. (2004) provide an up to date summary of the methods used to obtain the 2dFGRS magnitudes and colours.

We define rest frame colour

$$(B - R)_0 \equiv B_{\rm J} - R_{\rm F} - K(B_{\rm J}) + K(R_{\rm F}) \tag{3.50}$$

where the colour-dependent K corrections are

$$\begin{aligned}K(B_{\rm J}) =&(-1.63 + 4.53C)\,y\\&+ (-4.03 - 2.01C)\,y^2\\&- \frac{z}{[1 + (10\,z)^4]}\end{aligned} \tag{3.51}$$

$$\begin{aligned}K(R_{\rm F}) =&(-0.08 + 1.45C)\,y\\&+ (-2.88 - 0.48C)\,y^2\end{aligned} \tag{3.52}$$

with $y = z/(1+z)$ and $C = B_{\rm J} - R_{\rm F}$. A division at $(B - R)_0 = 1.07$ achieves a similar separation between 'passive' and 'actively star forming' galaxies to the spectral classification split at $\eta = -1.4$, giving a total of 77 120 red galaxies and 144 292 blue galaxies.

The distributions of $\eta$ type and colour for the 2dFGRS galaxies are shown in Fig. 3.1, and the joint distribution is shown in Fig. 3.2. The correlation between the two properties is clear, together with the distinct bimodality, yet it is obvious that the relationship is not exactly one-to-one. Table 2.1 gives the respective numbers of each galaxy type in the 2dFGRS catalogue for comparison.

### 3.3.2 The volume limited galaxy sample

It is well known that the luminosity of a galaxy is correlated with galaxy type. Therefore in a flux limited sample the fraction of early/late types varies with redshift, potentially complicating any statistical analysis comparing their distributions. The size of the 2dFGRS presents the option of studying volume limited galaxy samples rather than the more usual flux limited datasets of previous galaxy redshift surveys. By imposing a luminosity and redshift cut, volume limited samples contain a representative sample of most galaxies over a large redshift range. Although some faint galaxies at low redshift are lost from the analysis, the sample selection effects are greatly simplified.

In order to create a volume limited sample, we first need to estimate the absolute magnitude, $M_{B_{\rm J}}$ of each galaxy in the sample:

$$M_{B_{\rm J}} = m_{B_{\rm J}} - DM - K(z) \tag{3.53}$$

where m is the apparent magnitude, corrected for Galactic extinction. $DM$ is the cosmology dependent distance modulus:

$$DM = 5\log_{10}\left[\frac{(1+z)c}{H_0}\int_0^z [\Omega_\Lambda + \Omega_m(1+z)^3]^{-1/2}dz\right] + 25 \tag{3.54}$$

$K(z)$ is the K-correction, which accounts for the differing region of the galaxy's spectral energy distribution seen in the observed wavelength frame due to the galaxy's redshift. For the 2dFGRS, the K-corrections as a function of $\eta$-type have been calculated by Madgwick et al. (2002). For those galaxies without an $\eta$-type, the average K-correction was used. To give an idea of the magnitude of correction involved for galaxies in the 2dFGRS: for a galaxy at the mean redshift of the survey ($z = 0.1$) the K-correction is 0.113 mags for the most star forming galaxies ($\eta$-Type 4), 0.303 mags for passive galaxies ($\eta$-Type 1) and 0.217 mags using the average K-correction.

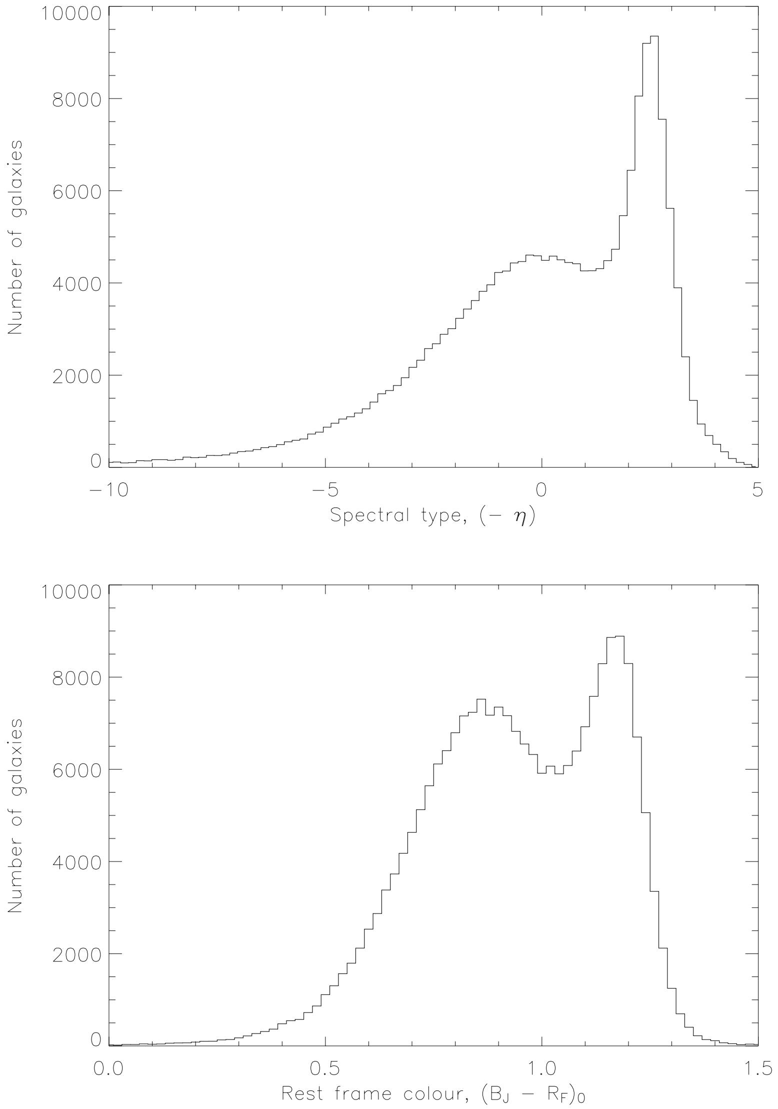

**Figure 3.1:** The distributions of spectral type and rest frame colour for all 2dFGRS galaxies. The distinction between passive and actively star-forming galaxies is clear in both distributions. Cuts at $\eta = -1.4$ and $(B - R)_0 = 1.07$ produce the four subgroups used in this chapter.

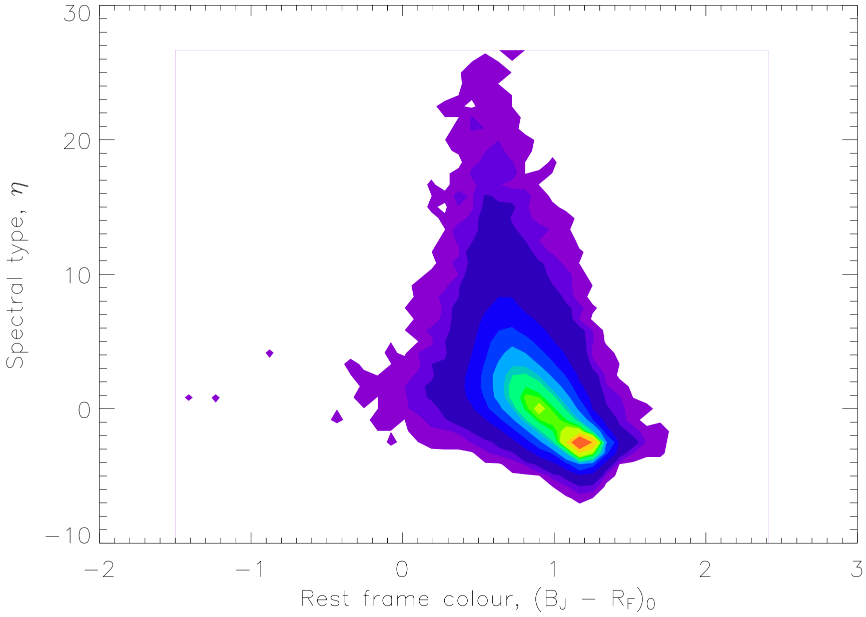

**Figure 3.2:** The joint distribution of the colour and spectral types for galaxies in the 2dFGRS. From outside inwards the contours enclose 99.9,99.5,99,95 and 90% of objects then decrease in steps of 10% till the innermost contour which encloses 10% of objects.

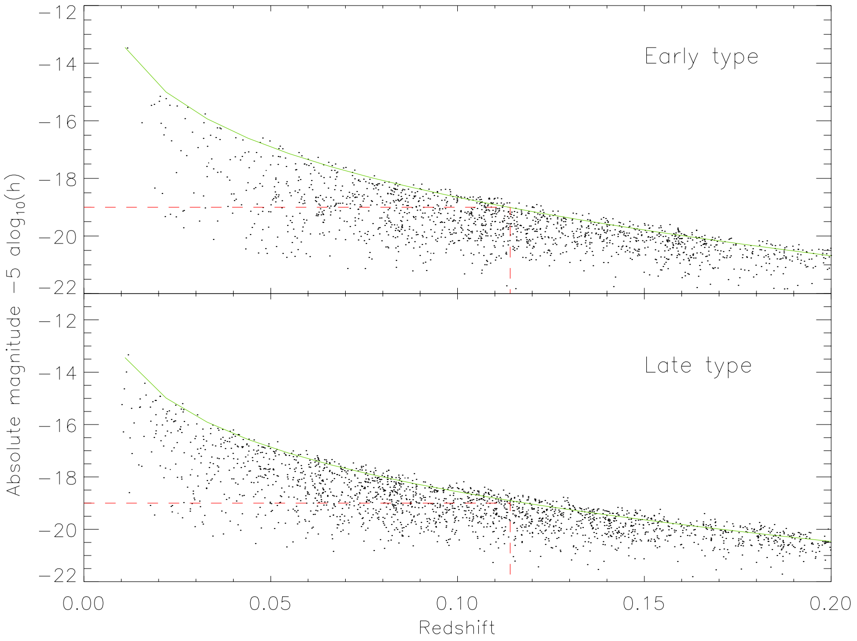

**Figure 3.3:** Creating volume limited samples: For illustrations ~2000 early- (top) and late-type (bottom) galaxies are plotted, selected randomly from the 2dFGRS catalogue. The absolute magnitudes for each galaxy are calculated according to their $\eta$-type K-corrections and redshifts. Overplotted is the survey apparent magnitude limit (19.2, green line) which is converted into an absolute magnitude using the K-correction for $-1.4 < \eta < 1.1$ galaxies. The sample limits used in this chapter are indicated in red (dashed line). These are defined by the early-type galaxies which are effected to a greater extent by their K-correction.

Next, an absolute magnitude limit must be decided upon. Here we trade off probing the fainter galaxy population against the volume of the Universe we can sample. The brighter the limit we choose, the further out we can see the faintest galaxies in the sample and the larger the volume we probe, at the expense of only retaining luminous galaxies. Due to the on average greater luminosity of early-type galaxies the limit must not be too bright as this work sets out to compare the two populations. Fig. 3.3 shows the absolute magnitude vs. redshift for 2dFGRS galaxies, split by their $\eta$-type into early and late subgroups. An absolute magnitude limit of $M_{B_{\rm J}} - 5\log_{10}(h) \leq -19.0$ gives a representative sample of the local population, maximising the number of cells, versus the number of galaxies in each cell. Solving equation (3.53) for a limiting survey magnitude of $B_{\rm J} = 19.2$ and taking the K-correction for early-type galaxies gives a maximum redshift for of $z_{\rm max} = 0.114$.

The publicly released data of 2003 June contains a total of 221 414 unique galaxies with reliable redshifts, 192 979 of which have spectral classification. The volume limited sample analysed here contains 48 066 galaxies, 46 912 of these have a spectral classification and 48 040 have measured colours.

## 3.4 Method: Counts-in-Cells

*Creating the cells*

The counts-in-cells analysis employed involves splitting the survey region into a lattice of roughly cubical cells and counting the number of galaxies in each cell. The cell dimensions are defined such that all have equal comoving volume $V \equiv L^3$, with limits to right ascension and declination that form a square on the sky. The cosmological co-moving volume element is given by:

$$dV = (R_0\, r\, d\psi)^2 R_0\, dr = A\, R_0^3\, r^2\, dr \tag{3.55}$$

where $A = d\psi \times d\psi = d\delta\, d\alpha\, \cos\delta$ is the area of the cell on the sky, $\alpha$ is right ascension and $\delta$ declination. The distance-redshift relation converts the comoving radius, $r$, into redshift for a given cosmology:

$$R_0 dr = \frac{c}{H_0} \left[(1 - \Omega_m - \Omega_\Lambda)(1+z)^2 + \Omega_\Lambda + \Omega_m (1+z)^3\right]^{-1/2} dz \tag{3.56}$$

where the radiation density has been neglected for the observational regime. Combining the above two equations with the requirement for square cells on the sky defines the right ascension, declination and redshift boundaries (via the distance-redshift relation) of each cell:

$$\delta_2 - \delta_1 = \frac{L}{R_0\, r_1} \tag{3.57}$$

$$\alpha_2 - \alpha_1 = \frac{L^2}{(R_0\, r_1)^2\, (\sin\delta_2 - \sin\delta_1)} \tag{3.58}$$

$$(R_0\, r_2)^3 - (R_0\, r_1)^3 = 3\, L\, (R_0\, r_1)^2 \tag{3.59}$$

where the subscripts $1$ and $2$ denote values at either edge of the cell.

This angular selection of the cells simplifies the treatment of the survey mask, but it means that the cells are not perfect cubes. Over the redshift range involved, this effect is small. The cells are required

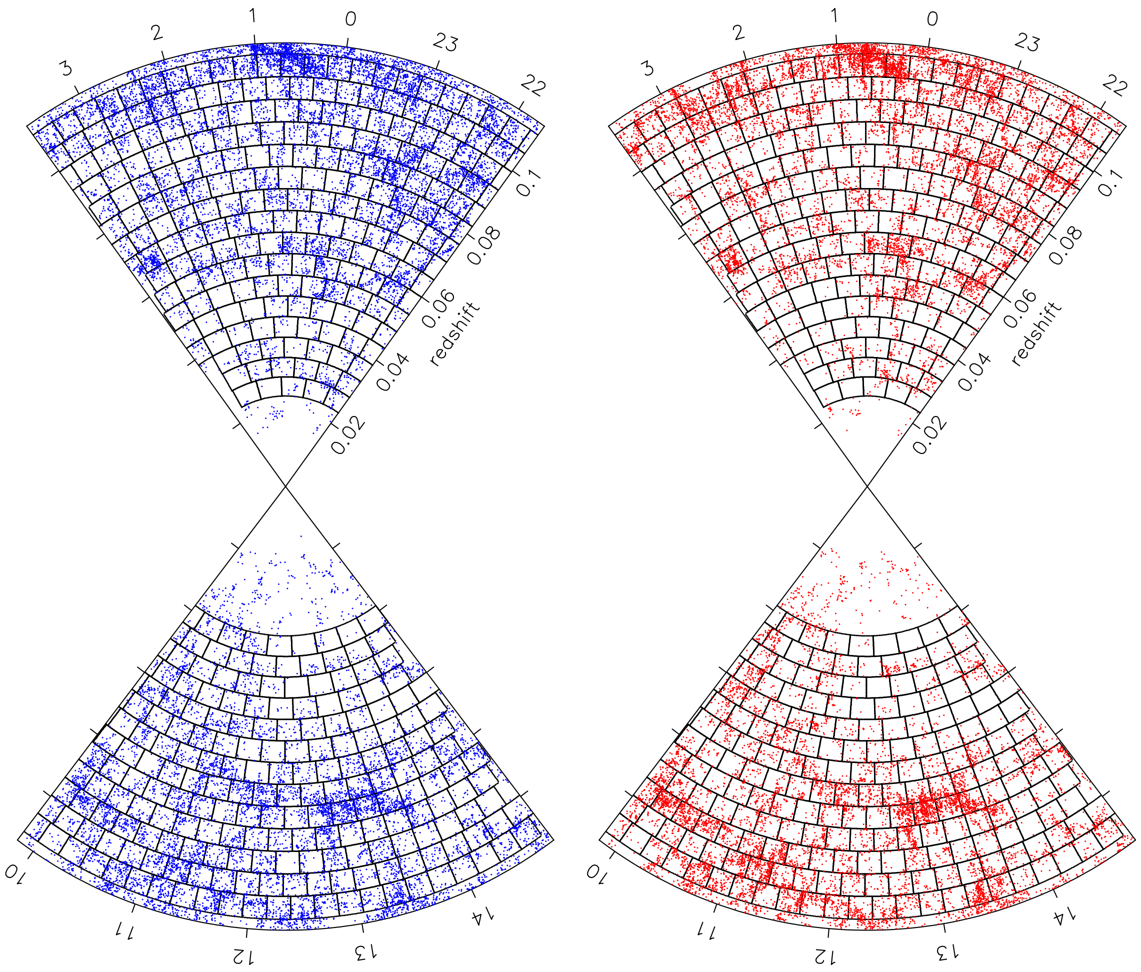

**Figure 3.4:** Wedge plots of the 2dFGRS volume limited survey region with $M_{B_{\rm J}} - 5\log_{10}(h) \leq -19.0$. Dots represent late type galaxies on the left and early type galaxies on the right (classified by spectral type). Redshift increases from the centre, and right ascension is shown on the horizontal axis, declination is projected onto the plane. Typical cell boundaries of length $L = 25\,{\rm Mpc}$ are overplotted.

to fit strictly within the 2dFGRS area, causing some parts of the survey to be unused. Although this restriction in principle removes any boundary effects, it means that cells of different sizes sample slightly different areas of the universe. The cells are defined with $h = 0.7$, and in what follows all cell lengths are quoted in $\mathrm{Mpc}$, instead of adopting the standard $h^{-1}\,\mathrm{Mpc}$ scaling. By changing the size of the cells any scale dependence in the quantities can be measured, and cells of size $10\,\mathrm{Mpc} \leq L \leq 45\,\mathrm{Mpc}$ were considered, giving a total of between 11 423 and 72 cells in the volume limited survey area after removing low completeness cells. For comparison to results in literature, these cell sizes are equivalent in volume to using top hat smoothing spheres with radii $6.1\,h^{-1}\,\mathrm{Mpc} \leq r \leq 27.9\,h^{-1}\,\mathrm{Mpc}$. Fig. 3.4 shows an example of how $25\,\mathrm{Mpc}$ cells cover the 2dFGRS volume to $z = 0.11$.

### *Calculating overdensities*

Due to internal holes in the survey and the adaptive tiling algorithm employed, the sampling fraction in the 2dFGRS varies over the sky as already discussed. Using the publicly available code (see Section 3.3), random 2dFGRS catalogues are created, which include these selection effects by making use of the calculated survey masks. When calculating the overdensity of each cell, the number of galaxies expected in that cell is weighted by the fraction of random points found in the same cell in this random catalogue. The spectral type analysis introduces extra selection effects, which are quantified by a special mask.

An overdensity $\delta$ is calculated for each cell by dividing the observed cell counts $N$ by the expected number for a given cell allowing for completeness, $\bar{N}$:

$$\delta_i = \frac{N_i}{\bar{N}} - 1 \tag{3.60}$$

This procedure is carried out for both early- and late-type galaxies within each cell. The overall density variance is defined by equation (3.10).

It is necessary to set a completeness limit to remove excessively under-sampled cells, such as those affected by holes in the survey due to stars, or cells at the less observed edges of the survey volume. Although the limits are somewhat arbitrary, it is important they are set correctly as incomplete cells could affect our measurements of scatter in the biasing relation. Fig. 3.5 shows the completeness distributions for $L = 25\,\mathrm{Mpc}$ cells. It can be seen that for both $\eta$ and colour classification the distribution has a sharp peak of almost complete cells, with a long tail to low completeness and a sharp cut off at high completeness. The figure also highlights the importance of including the effects of $\eta$ classification on the 2dFGRS mask as the completeness peak and cut off is noticeably lower for $\eta$ classification than for all galaxies. The lower completeness is reflected in the reduced number of cells available for analysis.

A completeness limit is set for each cell at 70% (or 60% for the larger cells), to include all cells within the high completeness peak. In order to check the effects of completeness on the final results, the models were also fit to only those cells with completenesses higher than 80% (70% for the larger cells), and the results found to be consistent within the errors.

A general concern with a counts-in-cells analysis of observational data is the varying survey selection function over a cell's extent. For example, a cell containing a cluster of galaxies at its most distant edge, and weighted by the average selection function over its volume, would give a different 'count' to a cell containing a cluster near its inner boundary. Furthermore, with a joint counts-in-cells analysis

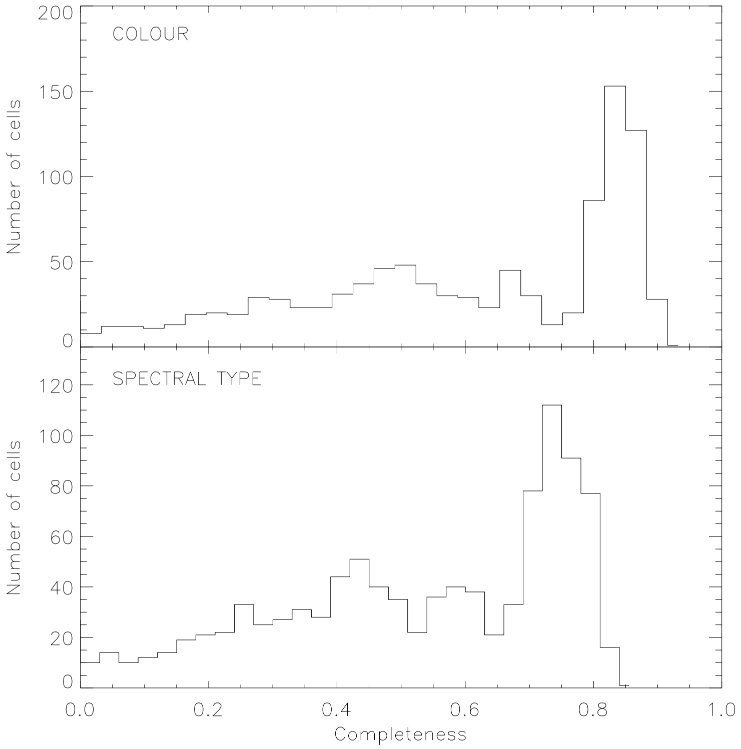

**Figure 3.5:** Histograms showing the completeness distribution of $25\,\mathrm{Mpc}$ cells using the standard 2dFGRS mask (top), and including the effects of $\eta$ classification (bottom).

any relationship between luminosity and galaxy type or colour will result in redshift dependent relative counts. With careful use of type-dependent selection functions such effects can be allowed for (Conway et al. 2004), but the size of the 2dFGRS offers a great advantage in providing volume limited samples large enough to produce reliable measurements and are unaffected by such complications.

## 3.5 Parameter fitting

### 3.5.1 A maximum likelihood approach

Once a model has been chosen and the cells created, a maximum likelihood method is used to fit the free parameters of the model to the data. Maximum likelihood is a powerful statistical tool for comparison of different datasets or models, making use of each data point rather than requiring the data to be binned such as with $\chi^2$ calculations.

Denoting the number of early (late) type galaxies within cell $i$ as $N_{\mathrm{E},i}$ ($N_{\mathrm{L},i}$), the likelihood of finding a cell containing $N_{\mathrm{E}}$ early type and $N_{\mathrm{L}}$ late type galaxies given a model with free parameters $\boldsymbol{\alpha}$, is defined as

$$L_i(N_{\mathrm{E},i}, N_{\mathrm{L},i}; \boldsymbol{\alpha}) = P(N_{\mathrm{E},i}, N_{\mathrm{L},i}|\boldsymbol{\alpha}) \tag{3.61}$$

and the total likelihood for all cells is then

$$L = \prod L_i. \tag{3.62}$$

The likelihood can be maximised with respect to the free parameters $\boldsymbol{\alpha}$ to find the best fitting values$\hat{\boldsymbol{\alpha}}$ for the model given the dataset. In practice it is easier to minimise the function

$$\mathcal{L} \equiv -\sum_i \ln L_i. \tag{3.63}$$

The models in Section 3.2 contain two or three free parameters: $\sigma_{\mathrm{E}}$ and/or $\sigma_{\mathrm{L}}$ from the one point PDFs, and $b$ or $r_{\mathrm{LN}}$ from the conditional probability function. A downhill simplex method (Press et al. 1992) was used to find the parameters which gave the minimum value of $\mathcal{L}$, fitting all parameters simultaneously.

### 3.5.2 Error estimation

As it is not possible to derive analytic solutions to the sampling distribution of our maximum likelihood estimators $\hat{\boldsymbol{\alpha}}$, the standard error on our parameters must be estimated directly from the likelihood function using Bayes' theorem and assuming a uniform prior on $\boldsymbol{\alpha}$:

$$P(\boldsymbol{\alpha}|\boldsymbol{x}) \propto L(\boldsymbol{x};\boldsymbol{\alpha}) \tag{3.64}$$

where $\boldsymbol{\alpha}$ again denotes the model parameters, $\boldsymbol{x}$ the data, and $P$ the probability.

For a single free parameter, the upper and lower limits on each parameter, $\alpha$, are found from

$$P(\alpha_- \leq \alpha < \alpha_+|\boldsymbol{x}) = \frac{\int_{\alpha_-}^{\alpha_+} L(\boldsymbol{x};\alpha)d\alpha}{\int_{-\infty}^{\infty} L(\boldsymbol{x};\alpha)d\alpha} \tag{3.65}$$

If it can be assumed that the likelihood function is reasonably approximated by a Gaussian, $1\sigma$ errors on the parameter can be estimated by fitting a Gaussian to the likelihood function. For multi-parameter models it is necessary to quantify any possible degeneracy between errors. If the multi-dimensional likelihood function can be approximated by a multivariate Gaussian distribution, individual errors and correlations between the parameters can be found.

A second method of error estimation involves creating many mock datasets from the fitted model probability distributions themselves. These datasets are made through Monte Carlo techniques: randomly selecting the number of early and late type galaxies in each mock cell according to the underlying probability distribution and the average number of galaxies expected. For all models presented in Section 3.2 the mock datasets can be created simply and quickly by making use of the standard numerical routines for normal distributions of random numbers (e.g. Press et al. 1992), translated into log-space where required. To simulate the varying completeness of the cells, the mean and variance of the true cell completeness is taken and the distribution approximated by a Gaussian. The mock cells are then weighted randomly according to numbers drawn from this distribution. Finally a Poisson sampling is applied to the mock cells, again making use of the standard routines for creating random numbers following a Poisson distribution and taking the mean of the distribution to be the number of galaxies in the mock cell multiplied by the cell completeness. The mock datasets are designed to closely reproduce the true data in sample size.

On applying the above likelihood techniques to each of these mock datasets, the best fit and true parameters of the underlying distribution can be compared to estimate the errors. The advantage of this method is that no assumptions need to be made about the shape of the likelihood function. The disadvantage is that we are assuming that the model is a correct representation of the data, as the errors strictly apply only to the model not the data. By increasing the size of the mock datasets, this method can also be used to check for any bias inherent in the fitting method. This process was carried out for each model in this paper, finding the parameter estimations to be unbiased.

In all cases, it has been assumed that the density fluctuations in each cell can be treated as independent. This is clearly not true in detail, since the existence of modes with wavelength$\gtrsim L$ will cause a correlation between nearby cells. This was considered by Broadhurst, Taylor & Peacock (1995), who showed that the correlation coefficient was low even for adjacent cells: $r \simeq 0.2$. As we shall see, it is $(1-r^2)^{1/2}$ that matters for joint distributions, and so the failure of independence is negligible in practice.

### 3.5.3 Model comparison

Once we have found the best fit parameters for each of our three models, we would like to know the goodness-of-fit of the models and the significance of any differences between the fits. We adopt two different methods.

*Likelihood ratios*

To test the significance of one model against another model the likelihood ratio test is used. In its simplest form we define the maximum likelihood ratio for hypothesis $H_0$ versus $H_1$

$$\lambda = \frac{L(\boldsymbol{x}|H_0)}{L(\boldsymbol{x}|H_1)} \tag{3.66}$$

where $\boldsymbol{x}$ is the data, and $L$ represents the *maximum* likelihood value. However, this test is strictly only applicable to models with the same number of free parameters. This is intuitively reasonable as a better fit should be always be achievable when the number of free parameters with which to fit the data is increased even when the model is incorrect. This will be especially important for assessing the evidence for stochasticity, where we will compare a model of perfect correlation with one where $r_{\rm LN} \neq 1$ is allowed, effectively introducing an extra parameter.

The key question is how large a boost in likelihood is expected from the introduction of an extra parameter. This was considered by Liddle (2004) who advocates the “bayesian information criterion”, defined as

$$B = -2\ln L + p\ln N, \tag{3.67}$$

where $p$ is the number of parameters and $N$ is the number of data points. This measure of information effectively says that going from a satisfactory model with $p$ parameters to one that over-fits with $p+1$ parameters would be expected to increase $\ln L$ by $0.5\ln N$. Therefore, in order to achieve evidence in favour of the increase to $p+1$ at the usual $1\sigma$ threshold (equivalently 5%), we require

$$\Delta \ln L = -\ln 0.05 + 0.5 \ln N, \tag{3.68}$$

which is between 5 and 8 for the number of cells considered here. An unequivocal detection of stochasticity thus apparently requires a likelihood ratio between $r_{\rm LN} \neq 1$ and the best $r_{\rm LN} = 1$ model in excess of $\lambda \simeq \exp(5)$ to $\exp(8)$.

Monte Carlo simulations may be used to check the validity of this analytic method. This is computationally expensive, so only an upper limit may be set on the significance of an observed likelihood ratio. We create 40 mock datasets following a power law bias model convolved with a Poisson distribution, defining the mean cell counts, number of cells, one point PDF fit parameter $\sigma_{\rm E}$ and model parameter $b_{\rm pow}$ to emulate a range of datasets. To these we fit both power law and bivariate lognormal models with the usual maximum likelihood fitting procedure. This allows us to assess the largest likelihood ratio that should arise by chance if the true model is in fact perfect power law bias. The results suggest a substantially smaller critical value is required than equation (3.68), closer to $\Delta \ln L = 1$ to reject the model at the $95\%$ confidence limit. It therefore appears that the assumptions used to derive the *bayesian information criterion* do not apply to this problem.

### *Kolmogorov-Smirnov test*

Although the likelihood ratio test can eliminate one model in favour of another, it can not tell us how well the preferred model fits the data. A Kolmogorov-Smirnov (KS) test can be used to test for a difference between an observed and modelled cumulative probability distribution. This test provides the probability that the data are drawn from the model probability distribution, with a low probability representing a poor fit. A resulting probability above about 0.1 is generally accepted as a reasonable fit, as the KS test is unable to rule out the model being the true underlying distribution at greater than 90% confidence. Strictly the test becomes invalid once the data has been used to fix any free parameters of the model, as in this method (Lupton 1993). However, as long as the number of data points is much greater than the number of free parameters any effects should be small.

To transform a bivariate distribution to a one dimensional (1D) variable on which we can perform the standard KS test, an integrated probability distributions is created from both the model and the data, integrating within constant model probability contours centered on the position of maximum probability. This gives cumulative probability distributions for model and data from which the KS probability (that the data follow the same underlying distribution as the model) can be derived.

The KS test has been generalised to bivariate analyses by Peacock (1983) and Fasano & Franceschini (1987). However, this two dimensional KS test was found to lack power compared to the previous method for the present application.

## 3.6 Results

Table 3.1 summarises the details of each of the samples, including average count per cell, and completeness limit. The following sections look in detail at the univariate and bivariate model fits to each dataset.

**Table 3.1:** Completeness limits, total number of cells and average cell counts for each dataset after corrections for completeness have been applied. Note that numbers do not scale exactly as $L^3$ due to edge effects.

| | Cell size (Mpc) | compl. | no. cells | $\langle N_{\rm E}\rangle$ | $\langle N_{\rm L}\rangle$ |
|---|---|---|---|---|---|
| Colour | 10 | 0.7 | 11423 | 1.5 | 1.8 |
| | 15 | 0.7 | 3019 | 5.1 | 6.3 |
| | 20 | 0.7 | 1104 | 12.1 | 14.6 |
| | 25 | 0.7 | 484 | 25.6 | 30.1 |
| | 30 | 0.7 | 234 | 41.3 | 48.6 |
| | 35 | 0.6 | 169 | 57.4 | 70.7 |
| | 40 | 0.6 | 115 | 88.3 | 105.4 |
| | 45 | 0.6 | 72 | 125.9 | 149.2 |
| $\eta$ | 10 | 0.7 | 9668 | 1.9 | 1.9 |
| | 15 | 0.7 | 2567 | 6.5 | 6.3 |
| | 20 | 0.7 | 930 | 15.3 | 14.7 |
| | 25 | 0.7 | 404 | 32.1 | 30.2 |
| | 30 | 0.7 | 187 | 54.2 | 49.6 |
| | 35 | 0.6 | 115 | 71.8 | 71.4 |
| | 40 | 0.6 | 74 | 106.4 | 108.7 |

In Section 3.6.2 the scale dependence of nonlinearity and stochasticity in the 2dFGRS is investigated. In Sections 3.6.3 and 3.6.4 the origin of the stochasticity is discussed and some consistency checks are performed on the results.

### 3.6.1 One point probability distributions

Fig. 3.6 shows the bivariate distributions of cell counts, together with the one point distribution functions for a range of cell sizes. Before considering the bivariate distributions further, we look in detail at the individual lognormal fits to these one point distributions. A lognormal distribution convolved with Poisson noise is fit to the early and late number counts individually, using the method described in the previous sections. The best fitting lognormal models are shown overplotted in the Figure. It can be seen that on large scales the lognormal model alone fits the data well, but on small scales the deviation due to discreteness is substantial. For this reason it is important to account for shot noise in the fitting procedure.

In order to assess goodness-of-fit of the Poisson sampled lognormal model quantitatively, we compare a mock dataset derived from the best-fitting model with the true data using a KS-test. Firstly, many Monte Carlo cells are created in a similar manner to that described in Section 3.5.2 with best fitting parameters and expected number counts to match each dataset. Completeness effects are allowed for by randomly assigning each cell a completeness value from a Gaussian distribution of width and mean equal to those of the dataset. Fig. 3.7 shows the distributions of early type galaxies and Monte Carlo cells with matching model parameters. Overplotted are the best fitting lognormal model (dashed line) and a lognormal curve defined by the variance derived non-parametrically directly from the second moment of the cell counts (dotted line, see Section 3.6.5).

The Monte Carlo data can now be compared with the true data through a KS test (i.e. compare the red and blue histograms in Fig. 3.7). For large cells ($\geq 25\,{\rm Mpc}$) KS probabilities are found in excess

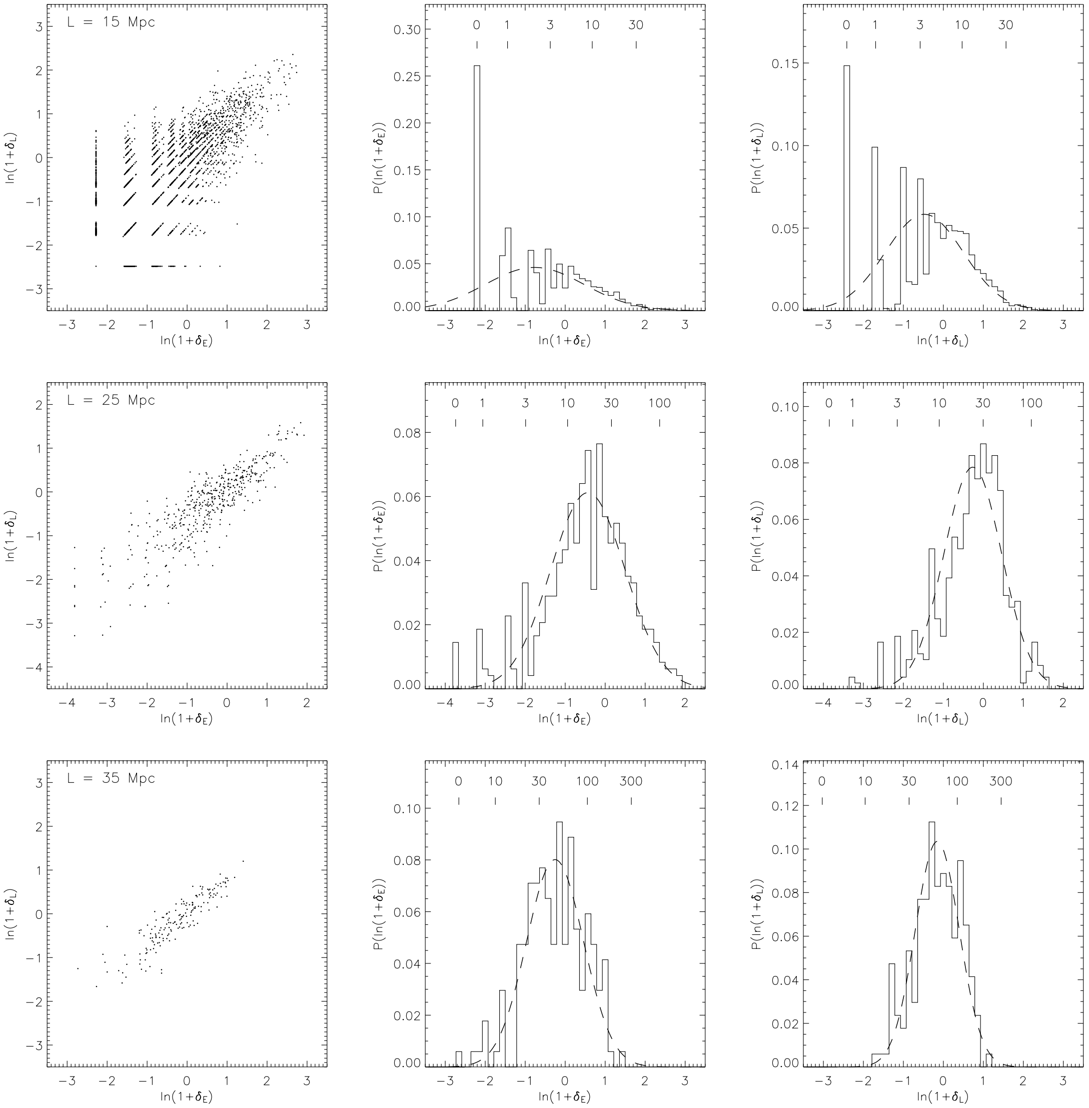

**Figure 3.6:** On the left, the bivariate counts-in-cells distributions with early and late type galaxies classified by colour. The axis give the log of the ratio of cell counts to expected counts (see equation 3.60) for early- (E) and late-type (L) galaxies. The points mark density values of individual cells, and from top to bottom $L = 15$, 25 and $35\,\mathrm{Mpc}$ cells are shown. 1D projections of the distributions are shown for early types (centre) and late types (right), to which a Poisson sampled lognormal model has been fitted. All distributions are normalised to a total probability of one under the histogram/curve. The best–fitting lognormal curves are overplotted (dashed line). Due to the logarithmic axes, a bin for cells containing zero galaxies has been artificially positioned on the horizontal axis. Note the discreteness of the galaxy counts: the actual number of galaxies contained in the cells is indicated by the numbers over the 1D distributions. Further, note the survey completeness effects on smaller counts per cell, causing the spread of points around the mean value. Correcting zero counts for completeness is non-trivial and not included in this analysis, hence there is no spread of these points.

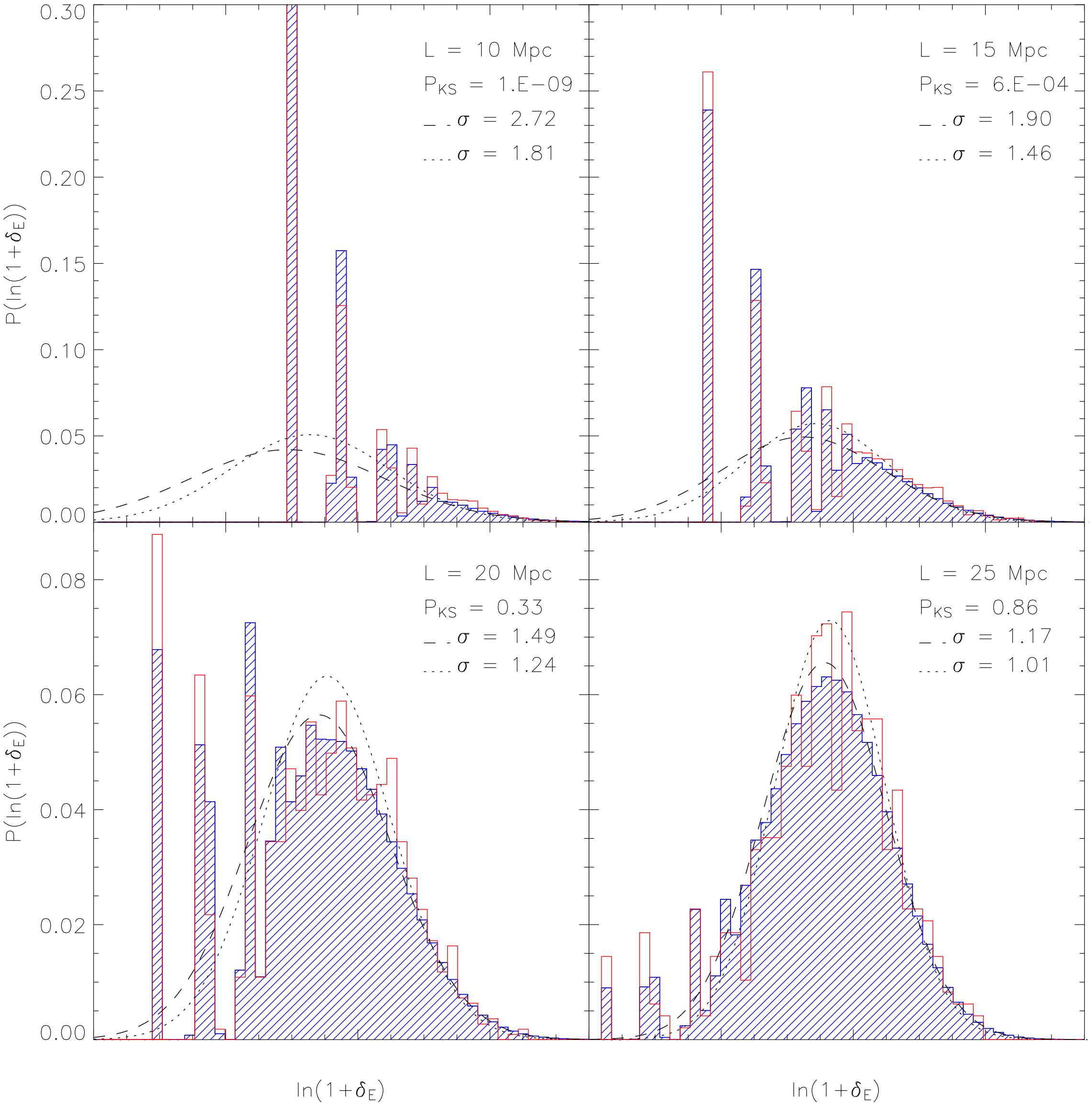

**Figure 3.7:** The univariate distributions of early type galaxies for $L = 10, 15, 20$ and $25\,\mathrm{Mpc}$ cells (empty, red histogram), together with the distribution of Monte Carlo cells (hatched, blue histogram) with parameters equal to those obtained from fitting a Poisson sampled lognormal distribution to the data cells. Completeness effects are modelled as a Gaussian with variance equal to that found in the dataset. The dashed curve shows this best-fitting lognormal distribution; the dotted curve shows the lognormal curve with variance equal to that measured directly from the data. Both variances are given in the upper right, together with KS probabilities that the Monte Carlo cells are drawn from the same distribution as the data cells. Due to the logarithmic axes, a bin for cells containing zero galaxies has been artificially placed on the horizontal axis. For $L = 10\,\mathrm{Mpc}$ about 60% of the cells contain zero early type galaxies, and the vertical axis has been truncated to allow a better view of the remaining bins. Note the discreteness of the data and simulations, as discussed in the caption to Fig. 3.6.

of 0.8, but as cell size decreases the KS probabilities decrease. On the smallest scales of $10\,\mathrm{Mpc}$, KS probabilities are obtained of $\sim 10^{-9}$. This poor model fit on small scales causes the lognormal model to overestimate variances in comparison with direct methods, and therefore explains the difference between the dotted and dashed curves. The Figure shows there to be an excess of data cells with moderate overdensities compared to the best-fitting lognormal model, particularly on the smallest scales. On small scales the majority of cells contain zero or one galaxy and in attempting to fit to these cells the lognormal model fails to fit the distribution at larger overdensities. This is equivalent to saying that the data contains more extreme underdense cells than if the data were to follow a lognormal distribution with parameters based on the data at moderate overdensities.

The overplotted curves reveal the same story. The dotted curve derived from the direct variance estimate lies below the dashed curve of the best-fit lognormal in the underdense regions. This results in an underprediction of the number of cells containing zero or one galaxy when combining the direct variance with the lognormal model (as discussed in more detail by Conway et al. 2004). One could imagine a weighting scheme to overcome this problem by down-weighting zero and one count cells in the likelihood analysis, however the key problem remains that the model is not a good description of the data on small scales and an alternative should be found.

### *Failures of the Poisson sampled lognormal distribution*

The discrepancy between the observed and predicted distributions of cell counts shows that at least one of the two assumptions made above about the galaxy field is incorrect: either the lognormal PDF or the Poisson sampling hypothesis.

The lognormal distribution is simply a convenient functional form which has been shown to fit galaxy distributions from previous surveys (e.g. Hamilton 1985; Kofman et al. 1994) and N-body matter distributions successfully (Kofman et al. 1994; Kayo et al. 2001, and references therein). Deviations from this simple model are evident in detailed numerical simulations (e.g. Bernardeau & Kofman 1995), and at some level at least such deviations would be expected in the data. Various alternative distributions have been suggested in the literature such as the skewed lognormal, negative binomial or Edgeworth expansions (see also Sheth et al. 1994; Valageas & Munshi 2004), however, until now there has been no requirement to make use of these more complicated models for galaxy surveys.

However, the fact that the model fails in detail on scales at which shot noise dominates the distribution in underdense regions suggests that the *Poisson clustering hypothesis* is at least partly to blame. By attempting to fit the model to these underdense cells, the variance is increased and the moderately overdense regions are no longer well fitted. On small scales the majority of cells contain zero or one galaxy, hence the preference of the model to fit these cells and not those containing more galaxies.

Without delving deeper into finding the best description of the galaxy probability distribution function the failures of the Poisson sampled lognormal model place some restrictions on the present analysis. In particular there are two points to make. The first is that cells of side about 10 Mpc are the smallest that can sensibly be discussed with this approach; reducing the cell size would lead to distributions that are dominated by discreteness effects. More importantly, it should be stressed that analytic models of this sort are not really physical. Whether the data is best-fit by a lognormal distribution or some slight variation tells us nothing about galaxy formation without some theoretical or numerical model with which to compare. As a theoretical framework for galaxy biasing does not exist, in the end what matters is whether the 2dFGRS data match the predictions of a proper numerical calculation of galaxy formation. Such a comparison is carried out at the end of the chapter, and the results of a Poisson-sampled lognormal fit are a convenient statistic to use for this purpose. Provided true data and mock data are treated identically, small imprecisions in the function used for the fit are irrelevant.

### 3.6.2 Joint distributions and biasing models

Each of the three biasing models in Section 3.2 is fitted to the datasets described in Table 3.1. Best fit parameters for the models are estimated simultaneously through the maximum likelihood method of Section 3.5.1. Errors are determined through multivariate Gaussian fits to the likelihood surfaces, which were found to agree well with Monte Carlo error estimates. Tables 3.2 and 3.3 shows the best fitting parameter values for the two deterministic models, together with log-likelihood differences between the model and the bivariate lognormal model. The values of $b_{\rm pow}$ and $b_{\rm lin}$ clearly show how early type galaxies are more clustered than late type galaxies, as is well known. Table 3.4 gives the best fitting parameter values and errors for the stochastic bias model.

Fig. 3.8 shows the joint probability distribution of the data for $L = 20\,{\rm Mpc}$ cells, together with Monte Carlo realisations of the best fitting linear, power law and bivariate lognormal models for comparison. The Monte Carlo realisations include completeness effects by randomly selecting a completeness value for each cell from a Gaussian with mean and width equal to that of the true distribution of cell completeness. This scale is chosen to illustrate all the properties of the data and the figures clearly show the shot noise in underdense regions, together with the effects of survey incompleteness. By eye we can see the differences between the linear and power law bias models (curved vs. straight locus), and the effect of stochasticity in the bivariate lognormal model (broadening of the distribution beyond the Poisson scatter). The best fitting linear model has a mean bias closer to unity at high density, and cannot fit the nonlinearity seen in the data. The power law model corrects for this, but the scatter about the mean is insufficient to match the data. The stochasticity introduced by the bivariate lognormal model is evident and is matched well by the data. The likelihood ratios shown in Tables 3.2 and 3.3 quantify the differences and show that on all scales the bivariate lognormal model fits significantly better than the deterministic biasing models.

The analysis is now repeated, splitting galaxies by spectral type $\eta$, rather than by colour. The colour split allows a larger sample of cells to be included in the analysis, but Fig. 3.2 shows that a division by spectral type does not always select the same galaxies as a colour split, so it is interesting to see how the results compare. The second section of Tables 3.1, 3.2, 3.3 and 3.4 give details of the datasets and results of the model fits to cells with galaxies classified by $\eta$. The joint distributions for $20\,{\rm Mpc}$ cells are shown in Fig. 3.9.

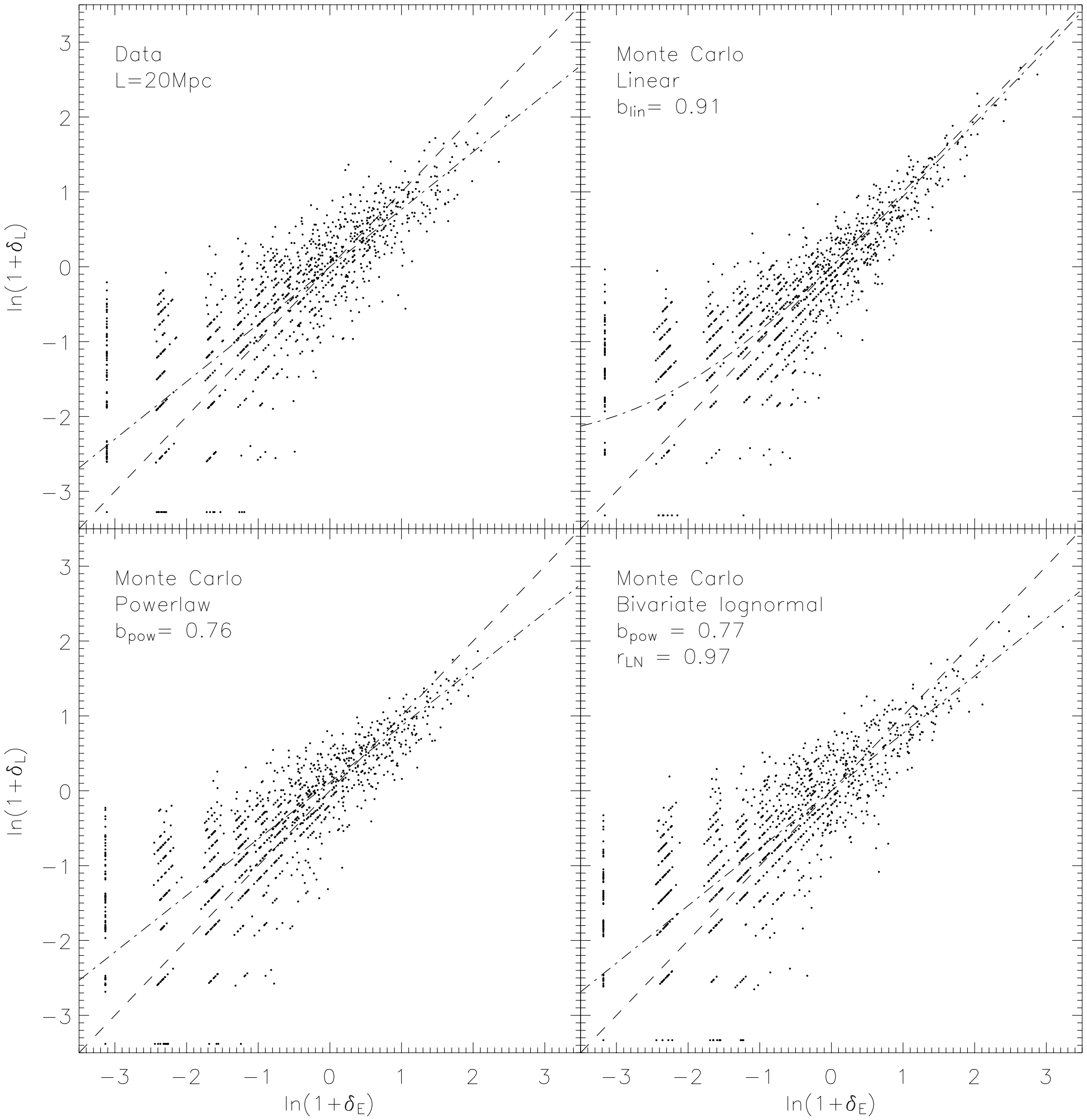

**Figure 3.8:** On the top left, the bivariate counts-in-cells distributions for $20\,\mathrm{Mpc}$ length cells, with early and late type galaxies classified by colour. The points mark density values of individual cells. The other three panels show Monte Carlo realisations of the best fitting linear, power law and bivariate lognormal models. The realisations are created to match the data as far as possible, with equal cell numbers and average number counts. Cell completeness is included by assuming the distribution of cell completeness to be a Gaussian of mean and width equal to that of the data cells. In each panel the dashed line shows the $b = 1.0$ case, and the dash-dot line shows the mean biasing of each model (for the top left plot, the dash-dot line shows the mean biasing of the best fitting bivariate lognormal model). Poisson sampling of the galaxies is assumed in all cases. Note that for all but $b = 1.0$, linear bias appears as a curve on the log-log plots. Due to the logarithmic axes, cells containing zero early or late type galaxies have been artificially positioned.

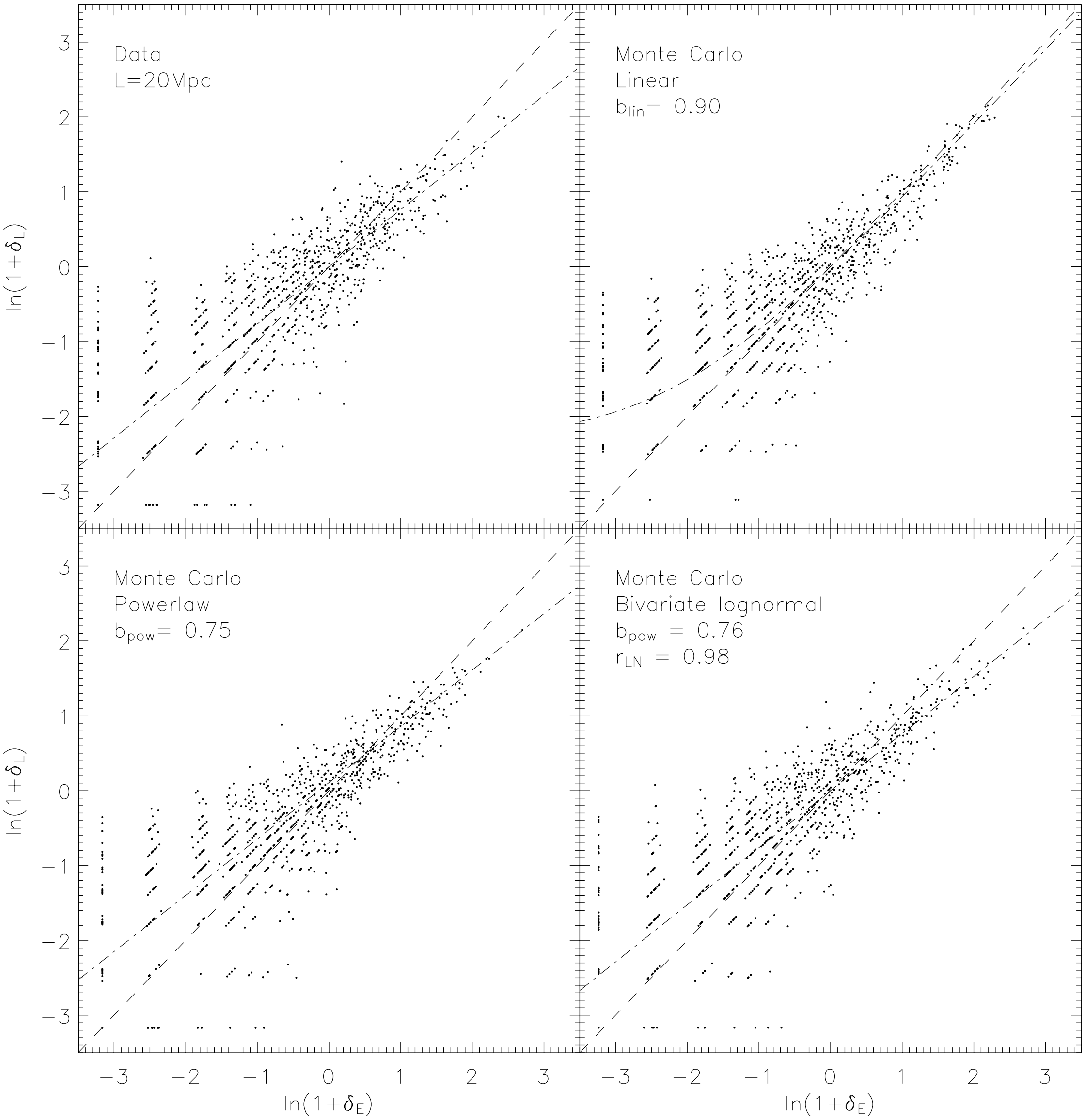

**Figure 3.9:** Same as Fig. 3.8, with galaxies classified by $\eta$ type.

**Table 3.2:** The best fitting deterministic biasing models parameters to each dataset split by colour. The level of nonlinearity given the model is given by $\tilde{b}/\hat{b}$, which is unity by definition for the linear model. The penultimate column shows the log-likelihood differences between the best fit linear or power law models and bivariate lognormal model. A positive value indicates the bivariate lognormal model is a better fit to the data. The final column shows how many cells must be removed to reduce the power law likelihood ratio to $\sim \exp(1)$ (Section 3.6.3).

| Cell size | Model | $\omega_{\rm E}$ | $b_{\rm lin}$ or $b_{\rm pow}$ | $\hat{b}$ | $b_{\rm var}$ | $r_{\rm lin}$ | $\tilde{b}/\hat{b}$ | $\mathcal{L}-\mathcal{L}^{\rm LN}$ | Outliers |
|---|---|---|---|---|---|---|---|---|---|
| 10 | linear | 1.41 | 0.93 | 0.93 | 0.93 | 1.00 | 1.000 | 291.1 | |
| | power law | 1.51 | 0.78 | 0.56 | 0.58 | 0.96 | 1.044 | 75.6 | 159 |
| 15 | linear | 1.23 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 185.0 | |
| | power law | 1.26 | 0.77 | 0.61 | 0.63 | 0.97 | 1.030 | 55.5 | 86 |
| 20 | linear | 1.04 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 133.6 | |
| | power law | 1.10 | 0.76 | 0.63 | 0.65 | 0.98 | 1.024 | 37.9 | 40 |
| 25 | linear | 0.92 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 87.1 | |
| | power law | 0.94 | 0.76 | 0.67 | 0.68 | 0.98 | 1.016 | 41.3 | 34 |
| 30 | linear | 0.80 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 39.3 | |
| | power law | 0.77 | 0.77 | 0.72 | 0.72 | 0.99 | 1.009 | 22.4 | 18 |
| 35 | linear | 0.73 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 16.0 | |
| | power law | 0.70 | 0.76 | 0.71 | 0.72 | 0.99 | 1.008 | 6.6 | 6 |
| 40 | linear | 0.67 | 0.95 | 0.95 | 0.95 | 1.00 | 1.000 | 18.1 | |
| | power law | 0.68 | 0.81 | 0.77 | 0.78 | 1.00 | 1.005 | 12.4 | 8 |
| 45 | linear | 0.59 | 0.97 | 0.97 | 0.97 | 1.00 | 1.000 | 12.9 | |
| | power law | 0.59 | 0.86 | 0.83 | 0.83 | 1.00 | 1.002 | 10.4 | 1 |

**Table 3.3:** Same as Table 3.2 except with the dataset split by $\eta$-type.

| Cell size | Model | $\omega_{\rm E}$ | $b_{\rm lin}$ or $b_{\rm pow}$ | $\hat{b}$ | $b_{\rm var}$ | $r_{\rm lin}$ | $\tilde{b}/\hat{b}$ | $\mathcal{L}-\mathcal{L}^{\rm LN}$ | Outliers |
|---|---|---|---|---|---|---|---|---|---|
| 10 | linear | 1.43 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 238.1 | |
| | power law | 1.51 | 0.77 | 0.55 | 0.58 | 0.96 | 1.046 | 49.8 | 110 |
| 15 | linear | 1.22 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 131.1 | |
| | power law | 1.24 | 0.77 | 0.64 | 0.65 | 0.97 | 1.026 | 35.6 | 58 |
| 20 | linear | 1.03 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 91.9 | |
| | power law | 1.07 | 0.75 | 0.67 | 0.68 | 0.98 | 1.019 | 17.9 | 25 |
| 25 | linear | 0.93 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 70.3 | |
| | power law | 0.96 | 0.74 | 0.65 | 0.67 | 0.98 | 1.018 | 23.7 | 21 |
| 30 | linear | 0.81 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 28.4 | |
| | power law | 0.81 | 0.73 | 0.67 | 0.68 | 0.99 | 1.013 | 13.3 | 9 |
| 35 | linear | 0.74 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 13.6 | |
| | power law | 0.75 | 0.69 | 0.65 | 0.66 | 0.99 | 1.013 | 4.3 | 2 |
| 40 | linear | 0.64 | 0.95 | 0.95 | 0.95 | 1.00 | 1.000 | 4.8 | |
| | power law | 0.65 | 0.83 | 0.82 | 0.82 | 1.00 | 1.003 | 1.2 | 0 |

**Table 3.4:** The best-fitting bivariate lognormal model parameters to each dataset. Errors are shown, derived from Gaussian fits to the parameter likelihood surface. $\Delta(r_{\rm LN})$ is derived from propagation of $\Delta[(1-r_{\rm LN}^2)^{1/2}]$. The remaining columns give the average biasing parameters. Section 3.2.4 gives the analytic solutions for each parameter in the case of the bivariate lognormal model. The final two parameters measure the nonlinearity and stochasticity of the model (equations 3.19, 3.20).

| | Cell size | $\omega_{\rm E}$ | $\Delta(\omega_{\rm E})$ | $\omega_{\rm L}$ | $\Delta(\omega_{\rm L})$ | $r_{\rm LN}$ | $\Delta(r_{\rm LN})$ | $\sigma_{\rm E}$ | $\sigma_{\rm L}$ | $r_{\rm lin}$ | $\hat{b}$ | $b_{\rm var}$ | $\tilde{b}/\hat{b}$ | $\sigma_b/\hat{b}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Colour | 10 | 1.52 | 0.01 | 1.20 | 0.01 | 0.958 | 0.004 | 3.01 | 1.80 | 0.88 | 0.52 | 0.55 | 1.054 | 0.44 |
| | 15 | 1.26 | 0.02 | 0.99 | 0.02 | 0.966 | 0.004 | 1.99 | 1.29 | 0.92 | 0.60 | 0.62 | 1.033 | 0.35 |
| | 20 | 1.10 | 0.02 | 0.85 | 0.02 | 0.969 | 0.005 | 1.54 | 1.02 | 0.93 | 0.62 | 0.64 | 1.026 | 0.31 |
| | 25 | 0.95 | 0.02 | 0.73 | 0.02 | 0.959 | 0.007 | 1.21 | 0.84 | 0.93 | 0.64 | 0.66 | 1.020 | 0.34 |
| | 30 | 0.78 | 0.03 | 0.61 | 0.03 | 0.962 | 0.009 | 0.92 | 0.68 | 0.94 | 0.69 | 0.70 | 1.011 | 0.31 |
| | 35 | 0.71 | 0.04 | 0.54 | 0.03 | 0.976 | 0.009 | 0.81 | 0.58 | 0.96 | 0.70 | 0.70 | 1.009 | 0.24 |
| | 40 | 0.68 | 0.05 | 0.55 | 0.04 | 0.971 | 0.009 | 0.76 | 0.60 | 0.96 | 0.75 | 0.75 | 1.006 | 0.27 |
| | 45 | 0.60 | 0.04 | 0.51 | 0.04 | 0.970 | 0.010 | 0.66 | 0.55 | 0.96 | 0.80 | 0.80 | 1.003 | 0.27 |
| $\eta$ | 10 | 1.51 | 0.02 | 1.18 | 0.01 | 0.963 | 0.005 | 2.95 | 1.75 | 0.89 | 0.53 | 0.56 | 1.052 | 0.40 |
| | 15 | 1.23 | 0.02 | 0.98 | 0.02 | 0.966 | 0.005 | 1.89 | 1.27 | 0.92 | 0.62 | 0.64 | 1.029 | 0.34 |
| | 20 | 1.07 | 0.02 | 0.82 | 0.02 | 0.976 | 0.005 | 1.47 | 0.98 | 0.94 | 0.63 | 0.64 | 1.025 | 0.27 |
| | 25 | 0.92 | 0.02 | 0.70 | 0.02 | 0.966 | 0.007 | 1.16 | 0.79 | 0.94 | 0.64 | 0.65 | 1.019 | 0.31 |
| | 30 | 0.81 | 0.05 | 0.60 | 0.04 | 0.965 | 0.010 | 0.97 | 0.66 | 0.94 | 0.64 | 0.65 | 1.016 | 0.30 |
| | 35 | 0.73 | 0.04 | 0.51 | 0.03 | 0.980 | 0.009 | 0.84 | 0.55 | 0.96 | 0.63 | 0.64 | 1.015 | 0.22 |
| | 40 | 0.66 | 0.04 | 0.54 | 0.04 | 0.988 | 0.008 | 0.73 | 0.58 | 0.98 | 0.78 | 0.79 | 1.004 | 0.17 |

Comparing the results for $20\,\mathrm{Mpc}$ cells, the results for colour and $\eta$ are generally similar. The likelihood ratios again favour the bivariate lognormal model over our two deterministic biasing models, and suggest a slightly smaller difference between the stochastic and deterministic models than found in the colour dataset. This is verified by the smaller stochasticity found in the best fitting bivariate lognormal model at all scales. Unlike for the colour datasets, power law bias is only marginally inconsistent with the data on the largest scales studied here. This difference between colour and $\eta$ type results may reflect a physical difference in the relative biasing relations, but firm conclusions are not yet possible as it is difficult to derive suitable properties from galaxy simulations with which to compare.

### *The goodness-of-fit statistics*

Tables 3.2 and 3.3 shows the log-likelihood differences $(-\ln\lambda)$ of the parameter fits, taking the bivariate lognormal model as our null hypothesis. In all cases the linear model provides a worse fit to the data than the power law model, and the power law a worse fit than the bivariate lognormal model. This latter statement may however simply be due to the addition of an extra free parameter to the model. The significance of the difference between the power law and bivariate lognormal model must be established: either the bayesian information criteria (equation 3.68) or results from fitting the bivariate model to Monte Carlo simulations of power law distributions may be used (see discussion in Section 3.5.3). For $L = 20\,\mathrm{Mpc}$ cells the likelihood ratios of exp(38) and exp(18), for colour and $\eta$-type classification respectively, are highly significant even for the more stringent bayesian information criteria which requires a likelihood ratio in excess of $\exp(6.5)$.

Although the likelihood ratios favour the bivariate lognormal distribution over the two deterministic models, they do not tell us how well the best-fitting distribution matches the data. For this the KS statistic described in Section 3.5.3 is used. On scales $L \geq 15\,\mathrm{Mpc}$ this KS statistic accepts the model with a probability greater than 0.5. On the smallest scales studied here the probability is found to decrease, in line with the trend for the univariate lognormal distribution (Section 3.6.1).

### *Stochasticity and nonlinearity*

In order to quantify the nonlinearity and stochasticity of the joint distribution of early and late type galaxies it is assumed that the bivariate lognormal model is an accurate representation of the data. In this analysis the log-density correlation coefficient $r_{\mathrm{LN}}$ provides a complete measure of the stochasticity; to aid comparison with other work the mean biasing, its nonlinearity and the average biasing scatter of equations (3.14) to (3.18) are quoted. For clarity we concentrate briefly on the results for $20\,\mathrm{Mpc}$ cells, shown in the third line of Table 3.4. These indicate that whilst the nonlinearity [equation (3.19)] is only 1.03, the stochasticity [equation (3.20)] is 0.31. This high stochasticity is reflected in the deviation of $r_{\mathrm{LN}}$ from unity, and the low linear correlation coefficient of $r_{\mathrm{lin}} = 0.93$. It is important to note that these statistics account for Poisson noise, as the models were convolved with a Poisson distribution before being fitted to the data.

Tables 3.2 and 3.3 shows for comparison some biasing statistics for the two deterministic models. It can be seen that a similar nonlinearity is measured by the power law model, whilst the linear correlation coefficient remains close to one, reflecting the inability of the model to measure stochasticity. The best fitting linear bias model has a mean biasing parameter closer to one, importantly indicating that by assuming this model other studies may be underestimating the magnitude of relative biasing, and therefore potentially galaxy-mass biasing.

It may be considered surprising that the correlation parameter $r_{\mathrm{LN}}$ can be measured so precisely that a value of 0.97 can be so clearly distinguished from a value of 1.00. The reason for this can be seen by examining the expression for the bivariate lognormal distribution [equation (3.39)], in which the scatter in $\delta_{\mathrm{L}}$ at fixed $\delta_{\mathrm{E}}$ is proportional to $S \equiv \sqrt{1 - r_{\mathrm{LN}}^2}$. This is a more meaningful quantity than the correlation coefficient, but it lacks a standard name. In this context, the obvious term for $S$ would be 'stochasticity', but this is already taken and the temptation to expand the terminology further is resisted. The stretched nature of this measure of correlation is quite extreme (as noted independently by Seljak & Warren (2004)): $S = 0.5$ corresponds to $r_{\mathrm{LN}} = 0.87$. Therefore, $r_{\mathrm{LN}} = 0.87$ is effectively half-way to no correlation at all. This is why even a correlation as high as $r_{\mathrm{LN}} = 0.97$ is noticeably imperfect in terms of density-density plots.

### *Scale dependence*

We now look at how the results depend on scale. This is interesting because it may potentially distinguish whether the efficiency of galaxy formation in a particular region of space is affected by local or non-local factors. Examples of local factors could be density, geometry, or velocity dispersion of the dark matter. Non-local factors could involve, for example, effects of ionising radiation from the first stars or quasars on galaxy formation efficiency, causing coherent variation over larger scales than possible from local factors.

Fig. 3.10 shows the bivariate distributions for 15, 25 and $35\,\mathrm{Mpc}$ cells with galaxies split by colour. The likelihood ratio tests (Tables 3.2 and 3.3) show that the bivariate lognormal bias model provides a significantly better fit to the data than both deterministic models on all scales for colour selection and all but the largest scales when classifying by $\eta$.

The nonlinearity and stochasticity as a function of scale are plotted in Fig. 3.11, with errors derived by propagation from those shown in Table 3.4. The commonly quoted parameters $b_{\mathrm{var}}$ and $r_{\mathrm{lin}}$, which combine both nonlinearity and stochasticity, are plotted in Fig. 3.12 to facilitate comparison with results in the literature. On small scales ($\leq 20\,\mathrm{Mpc}$) the average biasing statistics suffer systematic errors from overestimates of the variance by the Poisson sampled lognormal model fit, as discussed in detail in Section 3.6.5. To indicate the magnitude of these effects, open diamonds show results for the colour dataset replacing the variance estimated during the lognormal fit with direct variance estimates during calculation of the average biasing statistics, where the difference is greater than $1\sigma$. It can be seen that although both nonlinearity and $b_{\mathrm{var}}$ show noticeable change with scale, this can be mostly explained by the poor fit of the model. There is little effect on stochasticity, and both mocks and data are affected in the same way, making comparison practical. The nonlinearity reaches $< 1\%$ by around $35\,\mathrm{Mpc}$ with results for colour and $\eta$ classification barely distinguishable. A little care is needed in interpreting this result, however, negligible 'nonlinearity' does not mean that linear bias is a good fit. As much as anything, this is a statement that the amplitude of fluctuations declines for large $L$, so most cells have $|\delta| \ll 1$ and any deviations from linear biasing will appear smaller in magnitude than on smaller scales simply due to the decrease in magnitude of $|\delta|$.

The stochasticity also declines, although on large scales the errors prevent distinction between a flat

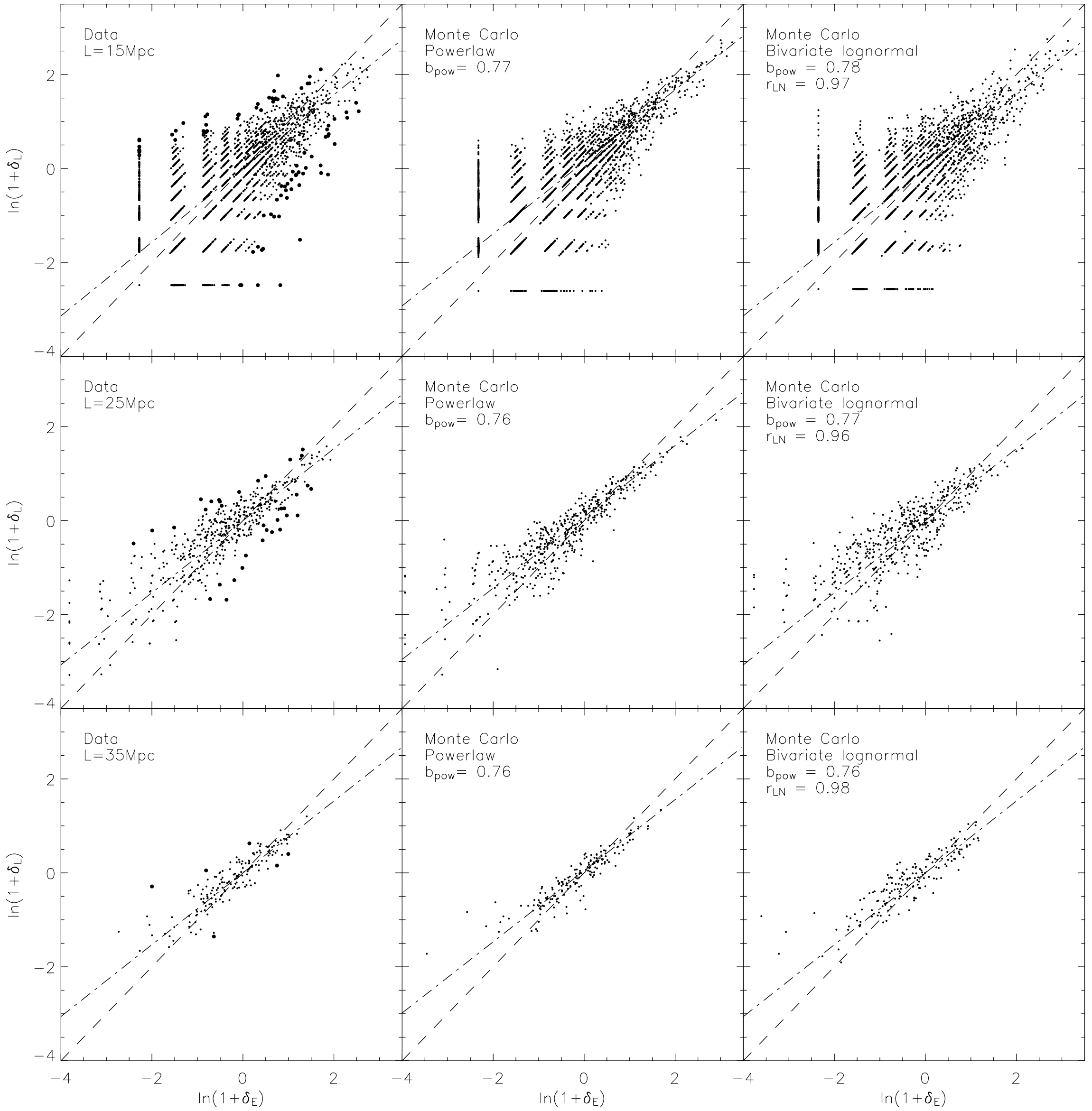

**Figure 3.10:** The bivariate counts-in-cells distributions for 15 (top), 25 and $35\,\mathrm{Mpc}$ length cells. The left hand panel shows the data with early and late galaxies classified by colour. Larger points indicate the cells identified as outliers from $r_{\mathrm{LN}} = 1$ (see Section 3.6.3). The central column shows a Monte Carlo simulation of the best fitting power law model and the right hand column the best fitting bivariate lognormal model. The dashed line indicates a mean biasing of $b = 1.0$, the dot-dash line shows the best fit mean bias.

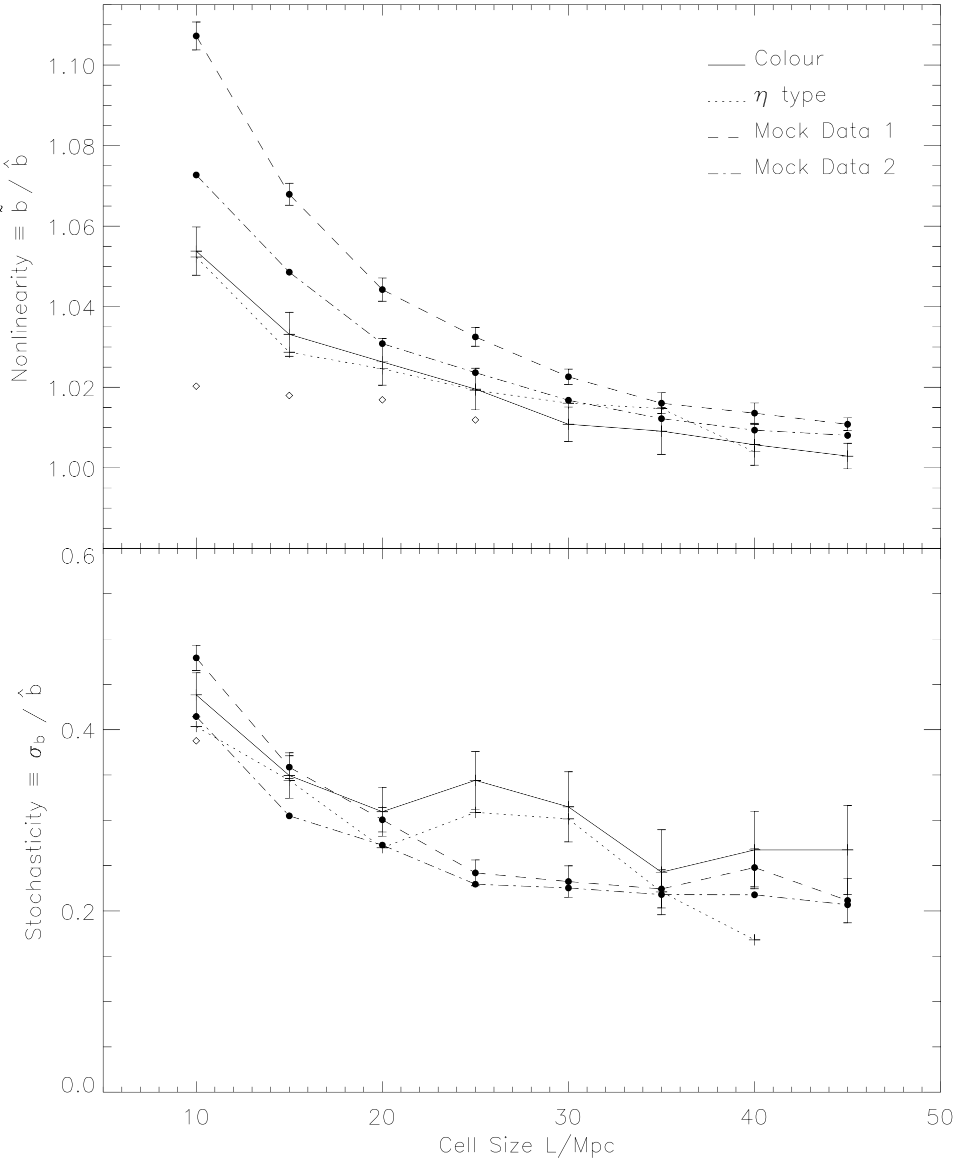

**Figure 3.11:** The scale dependence of nonlinearity and stochasticity in the 2dFGRS. The solid line shows results for galaxies classified by colour and dotted line for galaxies classified by $\eta$ type. The circles show the two semianalytic datasets. Mock 1 is described in the text as the "superwind" model, and Mock 2 as the "low-baryon" model. The open diamonds indicate values measured for the colour dataset using direct variance estimates, where they differ by more than $1\sigma$ from the results derived from model fitting (see Section 3.6.5). For clarity, errors are omitted for the $\eta$ and second mock datasets.

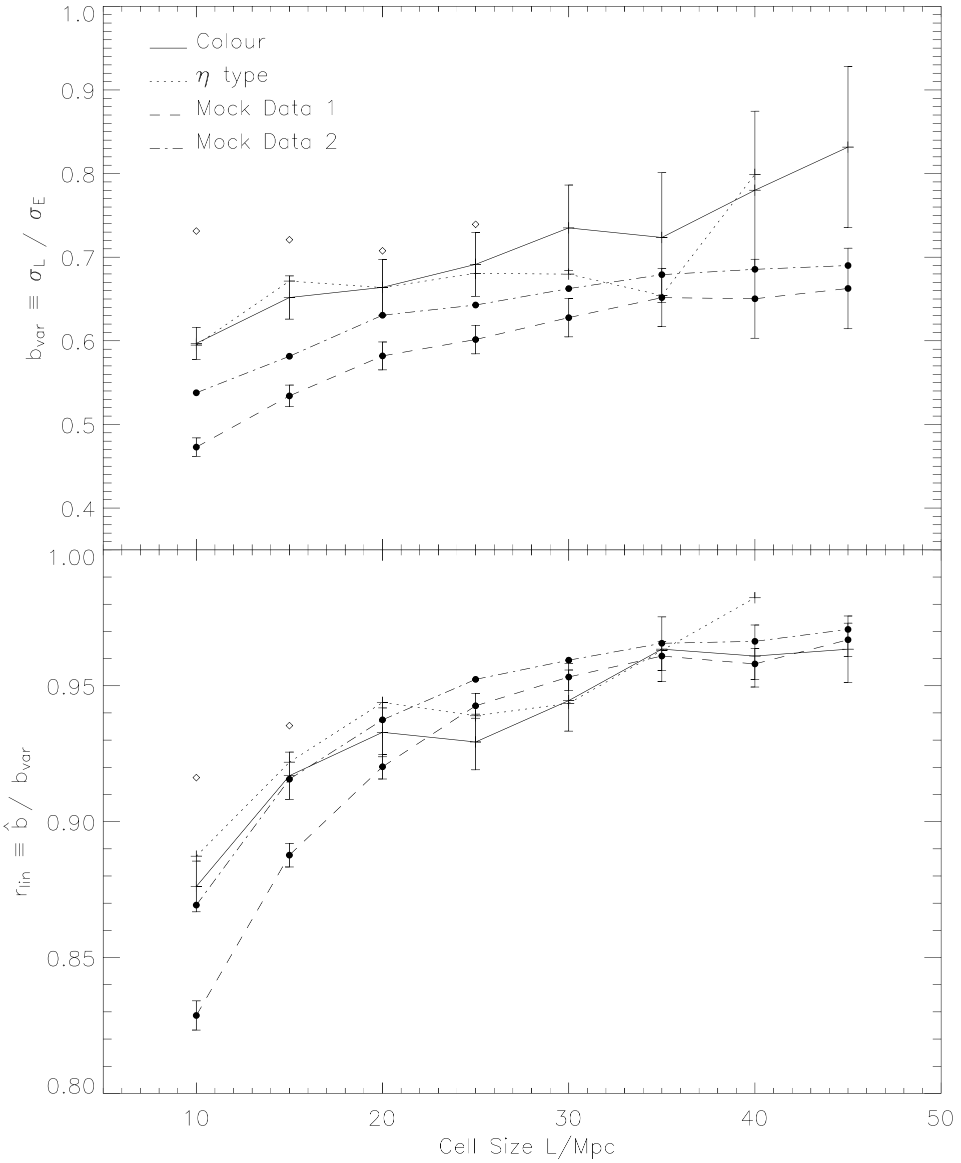

**Figure 3.12:** The scale dependence of the ratio of variances, $b_{\rm var}$, and the linear correlation coefficient, $r_{\rm lin}$. Symbols as in Fig. 3.11.

**Table 3.5:** Colour and $\eta$ samples with cell length $L = 20\,\mathrm{Mpc}$ are each split into two redshift groups at $z = 0.09$. This Table shows results for the bivariate lognormal model fit to each subsample.

| | | $\omega_{\mathrm{E}}$ | $\Delta(\omega_{\mathrm{E}})$ | $\omega_{\mathrm{L}}$ | $\Delta(\omega_{\mathrm{L}})$ | $r_{\mathrm{LN}}$ | $\Delta(r_{\mathrm{LN}})$ |
|---|---|---|---|---|---|---|---|
| Colour | low $z$ | 1.11 | 0.03 | 0.85 | 0.03 | 0.956 | 0.010 |
| | high $z$ | 1.11 | 0.03 | 0.85 | 0.03 | 0.974 | 0.006 |
| $\eta$ | low $z$ | 1.10 | 0.03 | 0.85 | 0.03 | 0.985 | 0.007 |
| | high $z$ | 1.07 | 0.03 | 0.82 | 0.03 | 0.976 | 0.007 |

or declining function with scale. There is a tendency for the stochasticity of the $\eta$ datasets to lie a little below that of the colour datasets, but this is not significant within the errors. The dashed and dash-dot lines show the results for two semianalytic mock universes which will be discussed in detail in Section 3.7.2. One can immediately note the encouraging general agreement: stochasticity is clearly expected at about the detected level. The small increase in stochasticity of the data around $25\,\mathrm{Mpc}$ could potentially be of interest if shown to be real. However, within the errors and with the knowledge that adjacent points are correlated, it is unlikely that the data show a significant deviation from the mock datasets.

### *Division by luminosity and redshift*

Splitting galaxies by their luminosity allows an investigation into whether the results of the previous sections could be due to the luminosity difference of the galaxy types. By dividing galaxies at $M - 5\log_{10}(h) = -19.5$, two similar sized groups are formed with class 1 being more luminous than class 2. The models are fitted as before by replacing E (L) with class 1 (2). In contrast to the outcome when galaxies are divided by type, the likelihood ratios between the best fitting power law and bivariate lognormal models are small on all scales, ranging from 0 to 3. It is found that $r_{\mathrm{LN}}$ is roughly constant with a value of $\sim 0.99$ for the best fitting bivariate lognormal models. If stochasticity is caused by some variable other than the local density during galaxy formation, then perhaps luminosity is less dependent on this variable than galaxy type and colour. Other explanations could be that our volume limited sample is too shallow to find the expected bimodality in luminosity, and the position of our boundary between bright and faint galaxies is arbitrary.

It is also of interest to see if the results are independent of redshift. The survey is divided at $z = 0.09$ and the models fitted to both high and low redshift galaxies using a cell length of $L = 20\,\mathrm{Mpc}$. Table 3.5 shows the best fitting bivariate lognormal parameters for galaxies split by colour and $\eta$ for both redshift groups. Due to the fibre apertures of the 2dF instrument some redshift dependence may be expected for galaxies classified by $\eta$ (see Section 3.3.1), yet precise predictions are difficult. Certainly no difference is seen within the errors between these two redshift groups, and the difference for colour classification can not be attributed to such effects as the colours are derived directly from the survey plates. It is possible that the changing errors on the colour at high redshift contribute to the decrease in stochasticity, although evolution can not be ruled out. There is certainly room for further investigation with forthcoming larger redshift surveys.

*Comparison with other 2dFGRS results*

This work has been carried out in conjunction with that of Conway et al. (2004) who investigate the variance and deviation from linear bias in the 2dFGRS NGP and SGP regions using flux-limited samples, including a counts-in-cells analysis. They find similar discrepancies between the Poisson sampled lognormal model and the data, investigating the causes and magnitude of the problem in detail. After accounting for this effect in both analyses the results agree within the $1\sigma$ errors where comparable: Conway et al. find $1/b_{\rm var} = 1.25 \pm 0.05$, and nonlinearity $(\tilde{b}/\hat{b})$ of a few per cent on the smallest scales measured. The results in this chapter for $b_{\rm pow}$ agree, but are consistently higher for $b_{\rm lin}$. This is due to different fitting procedures; Conway et al. give greater weight to overdense regions.

Madgwick et al. (2003b) measure the square root of the ratio of the correlation functions of early and late type galaxies to be around 1.2 on their largest scales of $8 < r < 20\,h^{-1}\,{\rm Mpc}$. Their bias parameter corresponds to $1/\sqrt{\hat{b}}$ in the notation of Section 3.2.1 (DL99). This gives a value for $\hat{b}$ a little higher than the results of Table 3.4, but entirely consistent when lognormal variance estimates are replaced by direct measures as in Section 3.6.2.

### 3.6.3 Origin of the stochasticity signal

Before the detection of stochastic bias is accepted, and we proceed to confront the result with theoretical models, a degree of skepticism is in order. The results thus far have indicated that some regions of space have a number ratio of early and late type galaxies that differs from the typical value by too much to be consistent with Poisson scatter. Such an outcome seems potentially vulnerable to systematics in the analysis as any source of error in classifying galaxies could introduce an extra scatter, spuriously generating the impression of stochasticity. However, it is not clear which way this effect would go. Suppose the survey finds galaxies with perfect efficiency, but then assigns them a random class. Any true initial stochasticity is erased by the classification ‘errors’ and we measure $r_{\rm LN} = 1$. In order to generate apparent stochasticity where none is present we would need something more subtle. Possibilities could include a perfect efficiency in detecting early type galaxies, but a fluctuating efficiency in finding late types; a spatially varying boundary between early and late types; or large variations in the survey selection function on scales smaller than the cell length. To assess the possible contribution of this latter effect to the measured stochasticity, small scale incompleteness masks were applied to semi-analytic datasets (see Section 3.7.2)[3]. The large scale stochasticity of these models was affected by less than $1\sigma$.

Whether or not a spurious generation of stochasticity seems plausible, it is worth looking more closely at the data to see how the signal arises. In order to do this, we focus on the outliers from the relation $\ln(1+\delta_{\rm L}) \propto \ln(1+\delta_{\rm E})$, but a careful definition of an outlier is required. We want to ask how much the numbers $(N_{\rm E}, N_{\rm L})$ differ from their expectation values when clustering is included, but the latter are unknown. Therefore, we take the best-fitting power law model with $r_{\rm LN} = 1$ and integrate over the distribution of densities to calculate the probability for obtaining this outcome, $P(N_{\rm E}, N_{\rm L})$, accounting for Poisson noise. The most outlying points are those with the lowest values of $P$, and these are removed in succession until the remaining cells are consistent with an $r_{\rm LN} = 1$ model. The numbers of outliers in this sense are listed in Tables 3.2 and 3.3, and Fig. 3.10 shows their positions on density-density plots.

[3] The test was carried out by John Peacock.

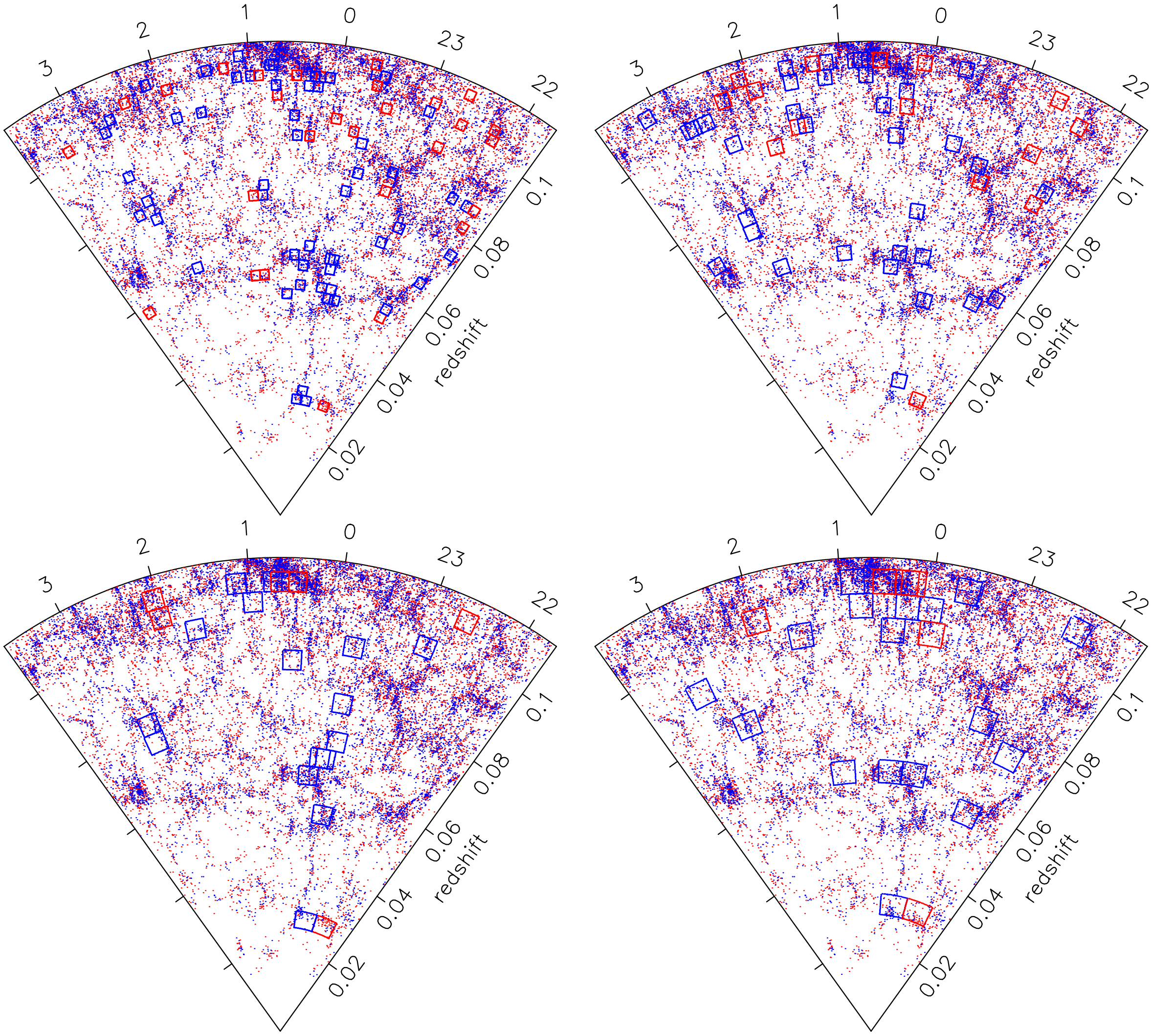

**Figure 3.13:** Wedge plots of the 2dFGRS volume limited survey SGP region as for Fig.3.4. Both early and late type galaxies are shown. Overplotted are cells identified as causing the stochasticity signal, from top left: 10,15,20,25 $\mathrm{Mpc}$ cells.

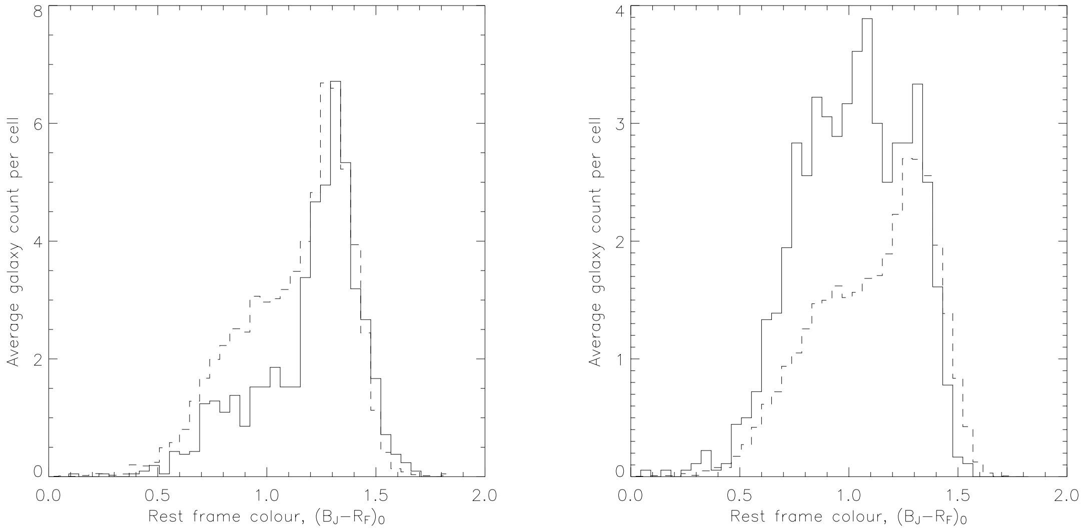

**Figure 3.14:** The colour distribution of cells identified as causing the stochasticity signal ($L = 20\,\mathrm{Mpc}$, filled line). The left (right) plot shows all those cells with an excess of red (blue) galaxies. The comparison plot (dashed line) is calculated from those cells with similar $\delta_{\mathrm{E}}$ to the outlier cells, in order to account for nonlinearities.

Having identified the cells that provide the evidence for stochasticity, their properties can be examined in more detail. Fig. 3.13 shows the spatial distribution of the outlying cells within the 2dFGRS for a range of cell sizes, from which it can be seen that they are often associated with overdense regions. This should not be taken as indicating that stochasticity is confined to such regions: given that the degree of stochasticity is small, the cells that contain the most galaxies will provide the best signal-to-noise ratio for the effect. The colour distribution of galaxies in the outlying cells is shown in Fig. 3.14, compared to the distribution of 'normal' cells. To allow for nonlinearity in the density-density relation only those cells with similar values of $\delta_{\mathrm{E}}$ are considered for the comparison distribution. The distributions cover sensible ranges of colours, and the peaks corresponding to early and late types appear to be in the correct places. What causes these cells to be outliers is that the ratio of the two populations differs greatly from what is typical, and it is hard to see how this result can be in error. The completeness values in these cells are typically 0.8, and yet we see variations in the early:late ratio by more than a factor 2. Moreover, similar variations are seen whether colour or spectral type is used for classification. The conclusion is therefore drawn that these variations are a real property of the galaxy distribution.

### 3.6.4 Consistency checks

The analysis is repeated for cells with galaxies classified randomly, recovering a best–fitting bivariate lognormal model with $r_{\mathrm{LN}} = 1$ exactly. By fitting the bivariate lognormal model to Monte Carlo simulated power law mocks (Section 3.5.3), any bias inherent in the fitting procedures can be checked for. The best fit models have mean $r_{\mathrm{LN}} \gtrsim 0.998$, not significantly different from the $r_{\mathrm{LN}} = 1$ of power law deterministic bias.

### 3.6.5 Direct variance estimates

As mentioned previously in this chapter, it is possible to determine the variance $\sigma^2(L)$ directly without assuming any model. Optimal power spectrum estimates perhaps provide the most accurate determinations of variance (Tegmark et al. 2004; Pen et al. 2003), however for the purposes of this work it suffices to use a simpler method presented by Efstathiou et al. (1990). Their estimator calculates $\Delta N = N - \langle N \rangle$ for each cell and subtracts the Poisson variance from $(\Delta N)^2$ to form an estimate of $\sigma^2$ for each cell. This is then averaged over all cells. The estimator only applies in the case of a uniform survey, where $\langle N \rangle$ is the same for all cells. For the general case of an incomplete survey, Efstathiou et al. derive a slightly different estimator, assuming a Gaussian density field. In fact, this is a poor assumption even on the largest scales considered here. Monte Carlo realisations of lognormal fields were created and used to show that their estimator for $\sigma$ is biased low by around 1-2%, and has an uncertainty often several times larger than that expected for the Gaussian model. Table 3.6 gives the direct variance estimates for the data cells, with errors from both Efstathiou et al. (1990) and Monte Carlo simulations.

Even accounting for the bias, these direct variance estimates remain generally 10–20% lower than those estimated by fitting a Poisson sampled lognormal curve. For early type galaxies in $L = 10\,\mathrm{Mpc}$ cells the discrepancy is nearly 40%. Imposing different weighting schemes during fitting can lower the lognormal variance to meet the direct variance results (Conway et al. 2004, see Section 3.6.1), but this only corrects the symptoms and not the underlying problem that the Poisson sampled lognormal model does not appear to be a good fit to the data.

This failure of the lognormal model to recover the true variance of the data may be due to the assumption of the *Poisson Clustering Hypothesis* which we know to be incorrect in detail. On the smallest scales the cells used in this chapter are largely shot noise dominated, and it is on these scales that the discrepancy is greatest (see also Section 3.6.1). It remains important to emphasise that the variance estimates quoted throughout this chapter are model dependent, and not to be taken as the true variance of the galaxies in the survey, which can be estimated more accurately through model independent methods.

An unfortunate side effect of this difficulty in obtaining accurate estimates for variance, is that the average biasing statistics of Section 3.2.1 are dependent on $\sigma$ (see also Section 3.6.2). Tests show this to have little effect on stochasticity ($\sigma_b/\hat{b}$), but on scales $\leq 20\,\mathrm{Mpc}$ nonlinearity is overestimated. By replacing our measured variance with results obtained from bias corrected direct estimation we find nonlinearity to decrease to around 2% for $L = 10\,\mathrm{Mpc}$, decreasing with scale gradually to match our measured values by $L = 30\,\mathrm{Mpc}$. Stochasticity is decreased by about 2$\sigma$ at $L = 10\,\mathrm{Mpc}$ to around 0.39, but the effect is insignificant on all other scales. These results may be compared with the measurements of variance and deterministic bias in the 2dFGRS using flux-limited samples over a slightly larger volume (Conway et al. 2004).

**Table 3.6:** Different variance estimates and errors for early and late datasets defined by colour. (a) From the bivariate lognormal model fit with errors derived from a multidimensional Gaussian fit to the likelihood surface; (b) Efstathiou et al. (1990) direct variance estimator and errors; (c) direct variance estimator after using Monte Carlo simulations of lognormal fields to correct for bias due to non-Gaussianity, with rms errors from the simulations.

| cell size | ${\sigma_{\rm E}}^{\rm a}$ | $\Delta(\sigma_{\rm E})^{\rm a}$ | ${\sigma_{\rm E}}^{\rm b}$ | $\Delta(\sigma_{\rm E})^{\rm b}$ | ${\sigma_{\rm E}}^{\rm c}$ | $\Delta(\sigma_{\rm E})^{\rm c}$ | ${\sigma_{\rm L}}^{\rm a}$ | $\Delta(\sigma_{\rm L})^{\rm a}$ | ${\sigma_{\rm L}}^{\rm b}$ | $\Delta(\sigma_{\rm L})^{\rm b}$ | ${\sigma_{\rm L}}^{\rm c}$ | $\Delta(\sigma_{\rm L})^{\rm c}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 3.014 | 0.074 | 1.806 | 0.016 | 1.814 | 0.142 | 1.799 | 0.037 | 1.321 | 0.013 | 1.323 | 0.059 |
| 15 | 1.987 | 0.060 | 1.462 | 0.021 | 1.472 | 0.129 | 1.295 | 0.033 | 1.054 | 0.016 | 1.056 | 0.056 |
| 20 | 1.542 | 0.057 | 1.242 | 0.028 | 1.253 | 0.126 | 1.024 | 0.035 | 0.879 | 0.021 | 0.882 | 0.055 |
| 25 | 1.214 | 0.048 | 1.008 | 0.034 | 1.019 | 0.107 | 0.839 | 0.032 | 0.746 | 0.026 | 0.750 | 0.054 |
| 30 | 0.922 | 0.045 | 0.877 | 0.042 | 0.890 | 0.102 | 0.677 | 0.033 | 0.659 | 0.032 | 0.663 | 0.061 |
| 35 | 0.807 | 0.064 | 0.728 | 0.041 | 0.736 | 0.087 | 0.584 | 0.042 | 0.547 | 0.032 | 0.551 | 0.050 |
| 40 | 0.764 | 0.067 | 0.716 | 0.049 | 0.727 | 0.095 | 0.596 | 0.049 | 0.557 | 0.038 | 0.563 | 0.061 |
| 45 | 0.657 | 0.054 | 0.547 | 0.047 | 0.557 | 0.071 | 0.546 | 0.044 | 0.467 | 0.041 | 0.473 | 0.058 |

This is a suitable point to discuss a subtlety of cell counts that has been neglected so far. Variation in the survey mask is represented by $\langle N \rangle$ varying between cells. We have treated this as a simple variation in sampling efficiency that is uniform over the cell. However, this cannot be precisely correct: where sampling of a cell is low because it encounters one of the larger drills in the input catalogue (unsurveyed regions due to bright stars), it would be more correct to assume a completely sampled cell of smaller volume. This alternative extreme was explored [4] by assuming that $\sigma \propto \langle N \rangle^{-0.3}$, as expected for a $\xi(r) \propto r^{-1.8}$ spectrum. Since $\langle N \rangle \propto$ completeness, and the typical cell completeness is about 0.8, the measured values of $\sigma$ are increased by about 7%, approximately a $1\sigma$ shift. This has no effect on the detection of stochasticity, and because the 'lost volume' assumption will not apply in all cases, the overall effect is expected to be negligible.

## 3.7 Comparison with simulations

In order to interpret the measurements of stochastic bias, we need to make a comparison with theory. In practice, this means considering the results of numerical simulations that are sufficiently detailed to predict the spatial distributions of the different classes of galaxy. There are currently two main methods of simulating the large scale structure of the visible universe: semianalytic or hydrodynamic. Each are considered in turn.

### 3.7.1 Previous work

Somerville et al. (2001a) used semianalytic models to measure the relative bias between early and late type galaxies (as defined by bulge to total luminosity) and red and blue galaxies on scales of $r = 8\,h^{-1}\,\mathrm{Mpc}$. They set a limiting magnitude of $M_B - 5\log\,h \leq -18.4$, and split galaxies by colour at $B - R = 0.8$, making their samples reasonably comparable to cells of $L = 20\,\mathrm{Mpc}$ in this analysis. The value obtained of $b_{\mathrm{var}} = 0.66$ (Table 3.4) falls in between their values of $0.77$ for late/early types and $0.55$ for blue/red galaxies. They find $r_{\mathrm{lin}} = 0.87$ for both subgroups, slightly lower than results found here for both colour and spectral type splits. Unfortunately the results are not split into stochasticity and nonlinearity, making it difficult to make further comparisons. It is however interesting that a lower amplitude of relative bias is found between the two colour groups in the 2dFGRS than is seen in these models.

The hydrodynamic simulations of Yoshikawa et al. (2001) classify galaxies by their formation redshift, and are smoothed with top hat spheres of radius $8\,h^{-1}\,\mathrm{Mpc}$. By using this classification scheme, hydrodynamic models approximate early type galaxies as those that form at high redshifts via initial starbursts, whereas late type galaxies have a lower formation redshift and undergo slower star formation. They find that old galaxies are positively biased with respect to matter with a linear correlation coefficient of less than 1, whereas young galaxies are slightly antibiased with a correlation coefficient closer to 1. They measure the relative bias between galaxy types by $b_{\xi}^{\mathrm{rel}} \equiv (\xi_{\mathrm{young}}/\xi_{\mathrm{old}})^{1/2}$ where $\xi_{\mathrm{young}}$ ($\xi_{\mathrm{old}}$) is the two-point correlation function of the young (old) galaxies. This is equivalent to $b_{\mathrm{lin}}$. They obtain values of between 0.5 and 0.66 for scales of $1\,h^{-1}\,\mathrm{Mpc} < r < 20\,h^{-1}\,\mathrm{Mpc}$, lower than the equivalent

[4] Calculations carried out by John Peacock.

values for the linear biasing model with $L \leq 25\,\mathrm{Mpc}$ cells (Tables 3.2 and 3.3). Once again results for stochasticity and nonlinearity are not quoted for the relative bias.

### 3.7.2 Preliminary mock comparison

None of this past work really allows a direct comparison with the results of this chapter, so two new theoretical 'datasets' were generated from the results of large semianalytic calculations carried out using the 'Cosmology machine' supercomputer at Durham. The background model is that deduced from the simplest WMAP+2dFGRS analysis of Spergel et al. (2003): flat, $\Omega_m = 0.27$, $\Omega_b = 0.045$, $h = 0.72$, $n = 0.97$, $\sigma_8 = 0.8$, applying the semianalytic apparatus of Cole et al. (2000) to a simulation with $N = 500^3$ particles in a box of side $250\,h^{-1}\,\mathrm{Mpc}$. As shown by e.g. Benson et al. (2003), a problem faced by such modelling is a tendency to over-produce massive galaxies, as a result of excessive cooling arising from the higher baryon density now required by CMB+LSS. This problem is particularly severe for disk (late type) galaxies. The first mock adopted the 'superwind' approach of Benson et al. (2003) in an attempt to alleviate this problem, but the cure is not total. The second mock attempted to reduce cooling by retaining the low baryon density of Cole et al. (2000). Although this conflicts with CMB data, it provides a useful comparison. An empirical approach was taken in which a monotonic shift in luminosity was applied to force the models to have the observed luminosity function as in Madgwick et al. (2002)[5]. The model colour distribution was bimodal to a realistic degree, so this shift was applied separately to generate model distributions of early and late type galaxies in which the global luminosity functions were correct. The resultant mock cell counts were analysed identically to the real data.

In some respects, these simulations match the real data very well. For the low-baryon model, the amplitude of the cell variances for early-type galaxies agree to within 3% on small scales and 10% on large scales. The superwind model variances agree to within 10% on all scales. The relative bias of the low-baryon model agrees to within 10 and 15% with observation, and the superwind model to within 10 and 20%. Significant stochasticity and nonlinearity is also required, which can be measured accurately as a function of scale because we are able to use more mock cells than are available in the real data. The mocks are affected in a similar manner to the data by the discrepancy between direct estimates of variance and those from lognormal model fits. On small scales this significantly increases the estimated nonlinearity; as the effect is equivalent between mocks and data. However, a direct comparison remains instructive. Fig. 3.11 shows the resulting stochasticity and nonlinearity as a function of scale, compared to that of the 2dFGRS data. The impression is that the mock results show a greater nonlinearity than the real data on small scales, while stochasticity is well matched within the errors.

Given the known imperfect nature of the semianalytic simulations (e.g. the failure to match luminosity functions exactly), the correct attitude is probably to be encouraged by the degree of agreement with the data. It is certainly plausible that the existing calculations contain all the relevant physical contributions to the observed bias, but perhaps not yet in quite the right proportions. As usual with such numerical comparisons, this raises the question of whether the issue of stochasticity can be understood in a more direct fashion. In the end, the effects we are seeing must be reducible to the way in which the early:late type galaxy ratio varies between and within virialized systems of different mass, so that in effect we are dealing with a more general version of the morphological segregation that is familiar

[5] Carried out by John Peacock

from the study of rich clusters (Narayanan, Berlind & Weinberg 2000). Such a follow-up investigation is beyond the scope of this chapter and discussed in Section 7 along with other further work suggested by the results presented here.

## 3.8 Summary and Conclusions

Fits have been presented of three relative biasing schemes to joint counts-in-cells distributions of 2dF-GRS galaxies, separated by both colour and spectral type $\eta$. Each scheme is convolved with a Poisson distribution to account for statistical 'shot noise'. The first two models present two alternative types of deterministic biasing: linear and power law bias. Linear bias is an important concept in cosmology and many results are linked to it, but it is not physically plausible as it allows negative densities. Power law bias presents a simple cure for this problem, but still has little physical motivation. With the advent of large semianalytic and hydrodynamic simulations, interest has grown in 'stochastic' bias models. Bias could be determined by parameters other than the local overdensity of the dark matter, and considerable scatter could occur in the relation. Galaxy distributions have previously been measured to be well approximated as lognormal, therefore a bivariate lognormal distribution seems a natural model for relative bias between galaxy types. This model incorporates stochasticity and nonlinearity in a well defined manner, which is mathematically simple and consistent with observation.

To account for the discrete nature of galaxies, the *Poisson clustering hypothesis* is assumed, and all models are convolved with a Poisson distribution. On small scales where the cell counts become shot noise dominated, this hypothesis may be failing, causing overestimates of variance compared to direct estimation methods. The main symptom of the discrepancy is a number of completely empty cells that exceeds the Poisson sampled lognormal prediction. An alternative explanation is that the lognormal model can no longer be used as a good description of the galaxy distribution as galaxy surveys increase in size and statistical power. The problem is found not to affect the final results for stochasticity, and the same effect is seen in the simulations, but it emphasises the need for a greater understanding of Poisson statistics in relation to galaxy clustering.

A significant deviation from $r_{\mathrm{LN}} = 1$ has been detected in the 2dFGRS and confirmed this detection of stochasticity through likelihood ratio tests, Kolmogorov-Smirnoff probability testing, and Monte Carlo simulations. Stochasticity is measured at a level of $\sigma_b/\hat{b} = 0.44 \pm 0.02$ or $r_{\mathrm{LN}} = 0.958 \pm 0.004$ on the smallest scales ($10\,\mathrm{Mpc}$), declining with increasing cell size. The nonlinearity of the biasing relation is less than $5\%$ on all scales. The small measured values of stochasticity and nonlinearity support the use of galaxy redshift surveys for studies of the large scale distribution of matter in the universe, and the measurement of cosmological parameters. However, as precision in cosmology increases and new techniques are developed, the effects of stochastic bias on parameter estimation should be understood. For example, studies of cosmology through weak gravitational lensing requires knowledge of nonlinear and stochastic bias (Seljak & Warren 2004). Our results for $\eta_{\mathrm{lin}}$ on 10 Mpc scales are consistent within the (large) errors with galaxy-mass correlations measured by weak lensing surveys (Hoekstra et al. 2002) on the largest scales probed.

A comparison with semianalytic simulations shows a similar variation of nonlinearity and stochasticity with scale. The amplitude of stochasticity appears a little lower than in the true data, particularly on large scales, and the nonlinearity is greater on small scales. Nevertheless, given the known imperfections of the current generation of semianalytic calculations, the general agreement is certainly encouraging. It is hoped that this work will stimulate the investigation of more detailed biasing models. Through the linking of new simulations to observations, a more thorough understanding of the processes of galaxy formation and evolution should be within our reach.

# 4

# PCA sky-residual subtraction of SDSS spectra

*The development of optical fibre spectroscopy has presented many new challenges to astronomical data reduction. Despite considerable improvements in recent years, accurate sky subtraction continues to pose a significant problem for both multi-object and integral-field instruments. The SDSS spectra provide a considerable improvement over 2dFGRS spectra, but problems in data reduction still impose restrictions on some potential scienctific applications. In this chapter a new method is developed to automatically remove residual OH sky emission lines from SDSS spectra using a principal component analysis in the observed frame. The new procedure allows the extensive wavelength range of the SDSS to be used to its full as exemplified in proceeding chapters.*

## 4.1 Introduction

The SDSS spectra present a dramatic improvement in quality over data from previous surveys: extended wavelength coverage, $3800-9200$Å; intermediate resolution, $\lambda/\Delta\lambda \simeq 1800$; high-quality relative and absolute flux-calibration; typical signal-to-noise ratios (S/N) of $10-30$; availability of accurate noise and mask arrays. Coupled with the overall homogeneity of the dataset this makes them suitable for an almost limitless number of quantitative investigations. Upon visual inspection, however, significant systematic sky-subtraction residuals longward of $6700$Å are evident in many spectra. For fainter objects in particular, the sky-subtraction residuals are the dominant source of uncertainty over a wavelength interval of nearly 2000 Å. The spectroscopic region involved includes features of significant astrophysical interest; examples include the Calcium triplet ($8500, 8545, 8665$Å), a powerful diagnostic of stellar populations accesible for low-redshift galaxies, and the H$\beta$ $\lambda4863$ + [OIII] $\lambda\lambda4960,\ 5008$ emission region in quasars and active galactic nuclei (AGN) in the redshift interval $0.4 < z < 0.8$.

Two examples where sky residuals in SDSS spectra cause a significant reduction in the scientific value of the catalogue are in the work of Bolton et al. (2004) on identification of spectroscopic gravitational lenses in the Luminous Red Galaxy (LRG) sample, and in that of Nestor et al. (2005) in the creation of metal absorption line catalogues in the quasar sample. The success rate of locating secondary objects in spectra, i.e. where a second object lies within the field of view of a fibre, falls dramatically as their spectroscopic features enter the red end ($> 6700$Å) of the spectra. Estimation of the statistical significance of detection of features in this region requires great care.

The problem of sky-subtraction is not a new one to astronomy; there is a long history of developing methods to better remove night sky photons from our images and spectra. In the following subsections some background is given firstly into the general issues relating to sky-subtraction and then to the specific problems in the SDSS dataset.

### 4.1.1 Sky light

So what makes the sky shine? There are four main contributors (Chamberlain 1961; Wyse & Gilmore 1992):

**Airglow:** non-thermal atmospheric emission, most notably [OI] $\lambda\lambda5577,\ 6300,\ 6364$ and [NaI] $\lambda5893$, the many OH emission bands redwards of $6500$Å, and $O_2$ A-band absorption at $7600$Å.

**Astronomical sources:** unresolved stars, clusters and galaxies, and nebular light.

**Aurora:** caused by interaction of atoms and molecules in the atmosphere with charged particles from the sun, contributing a complex emission spectrum - particularly [OI] $\lambda5577$ and $N_2^+$ $\lambda3914$Å. Fluctuations over less than a degree average out over a few seconds but can be orders of magnitude in intensity.

**Zodiacal light:** sunlight scattered off interplanatary dust, varying smoothly with time and position over scales of $\sim 1$ degree.

The many factors involved mean that in practice sky spectra must be recorded at the same time as the object+sky spectra, and within a few degrees of the target on the sky. Fig. 4.1 shows an example of the spectrum of the night sky in the red part of the optical wavelength range.

It is the non-thermal airglow and OH emission bands in which this chapter is principally interested, and the method presented here is designed to remove. The following subsections give a brief background to the methods employed in spectroscopic sky subtraction concentrating particularly on the removal of these sharp emission features.

### 4.1.2 Sky subtraction methods

The standard technique for sky-subtraction in fibre spectroscopy is to allocate around $10 - 20\%$ of the fibres to sky observations, spread uniformly across the telescope, detector and spectrograph fields (Wyse

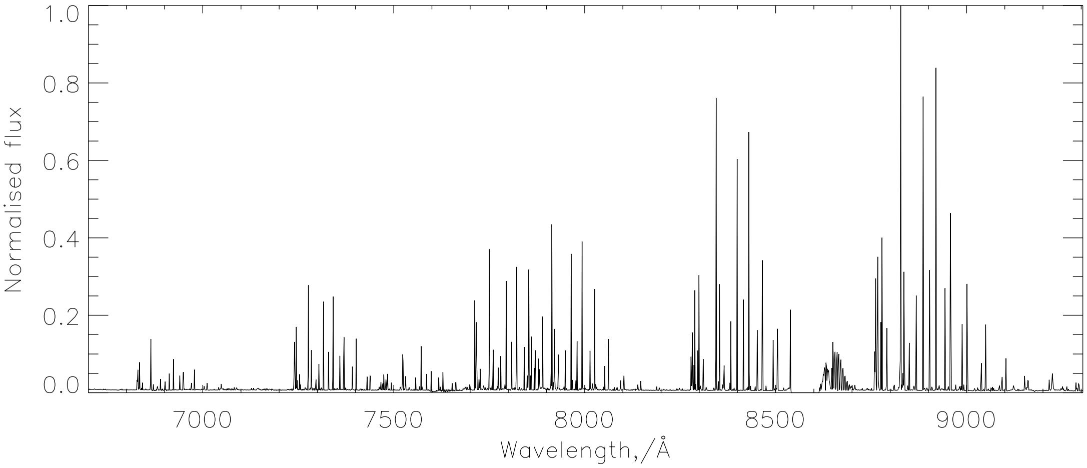

**Figure 4.1:** Night sky emission in the red end ($> 6700$ Å) of the optical wavelength range. The many emission lines are mostly due to OH transitions in the atmosphere.

& Gilmore 1992; Watson et al. 1998). Relative transmissions of all fibres are determined through sky or synthetic flat fields and all fibres are assumed to have identical spectral transmission. The spectra are then wavelength calibrated and averaged to give a master sky spectrum for the plate. This is scaled by the relative transmissions of each fibre and subtracted from each object+sky spectrum. The primary disadvantage of this method is that the sky is not observed locally to each object, and the light takes a different path through the optical system.

In contrast, in long or multi-slit spectroscopy the sky is observed locally in the telescope field, detector and spectrograph as the sky light passes through the same slit as the object+sky light. With regions of sky above and below the object, interpolation can further improve subtraction and systematic errors are small in comparison to Poisson noise. The result is that for faint object spectroscopy (fainter than about 20th magnitude in B-band) fibre systems are considered far inferior to longslit or even multislit (slitlet) systems which can now allow observations of up to $\sim 1000$ objects at once (e.g. GMOS, Hook et al. 2004).

### 4.1.3 Undersampling of OH lines

The problem of accurate subtraction of the strong, narrow OH sky features in which this chapter is interested, is however common to both long-slit and fibre observations. In general these lines are not sampled sufficently in the wavelength direction to fully determine their profiles which, together with small wavelength calibration errors, makes it near impossible to accurately subtract them from the object spectrum. This results in large oscillations in the final spectrum, in practice often making the spectrum unuseable beyond around $7000$Å (Parry & Carrasco 1990).

By combining the sky spectra into the "master spectrum" before rebinning them onto a common wavelength scale, some oversampling can be achieved (Lissandrini et al. 1994), as is done in the SDSS spectra. Kelson (2003) provides a recent summary of the difficulties associated with accurate sky-subtraction in long-slit observations, presenting a highly effective procedure for removing the signature of even rapidly varying and poorly sampled emission line features. This method makes use of the full extent of the skyline across the detector field, together with knowledge of the camera distortions and curvature of the spectral feature.

The problem remains that in optical fibre spectroscopic datasets the region beyond 7000Å contains noise significantly in excess of the Poisson limit and is unuseable for many scientific applications. The problem can be overcome during observations by employing e.g. a beam switching technique, however this drastically reduces the efficiency of the observations. The technique of Kelson (2003) may be adabtable to optical fibre datasets, but only with access to the raw data and full knowledge of the optical distortions of the instrumentation. Whilst large surveys are providing overwhelming amounts of data, seemingly simple data reduction problems are placing restrictions on the information we can retrieve from them.

### 4.1.4 Sky in the SDSS spectra

The sky-subtraction procedures incorporated in the SDSS spectroscopic pipeline are very effective, drawing on many years of experience with similar instruments. Brief details can be found in the Early Data Release paper (Stoughton et al. 2002, Sections 4.8.5 and 4.10.1) and information on calibration improvements for the Second Data Release (DR2) are given in Abazajian et al. (2004a). Basically for each spectroscopic plate 32 fibres are assigned to regions of blank sky and their spectra are combined into a "master-sky" spectrum which is then scaled and subtracted from each individual science object. This technique has proved extremely successful in accurately removing the correct amount of sky signal and allowing the spectra of even faint objects to be detected. The fundamental problem, however, lies in the ability to remove the particularly sharp OH emission features from spectra in which the line profiles are barely resolved i.e. the full width half maximum (FWHM) of these lines is less than the wavelength interval covered by an individual CCD pixel. Combined with sub-pixel changes in the pixel-to-wavelength calibration between spectra, sharp residuals often remain after subtraction of the master spectrum. The residuals arise from the subtraction of two essentially identical tooth-comb signatures (the master-sky spectrum from the science spectrum) that have been very slightly misaligned relative to one another, leading to well-defined patterns. Fig. 4.2 shows some examples of the type of patterns caused by the imperfections in the sky-subtraction procedure.

For the red end of the spectra to be used in some analyses, a method must be developed to deal with these residuals. Simply masking the pixels known to be affected would provide one solution, at the expense of loosing a significant fraction of the wavelength range beyond ∼7200Å. However the systematic deviations, correlated over extended wavelength ranges, suggests that a technique capable of quantifying the form and amplitude of the correlated deviations could allow the removal of a substantial fraction of the sky-subtraction noise. As described in Chapter 2, PCA is a well-established statistical method that identifies and quantifies correlations in datasets, and has a number of desirable properties for such an application.

The idea of employing PCA for sky-subtraction has been suggested by Kurtz & Mink (2000), who present a complex method to remove the sky signal without the need for concurrent sky observations. However, the scheme appears to have generated little interest, perhaps because of the ambitious goal of the technique and their focus on observations in which sky spectra are unavailable. By contrast the method developed in this chapter is targeted at a much more specific application of the PCA technique

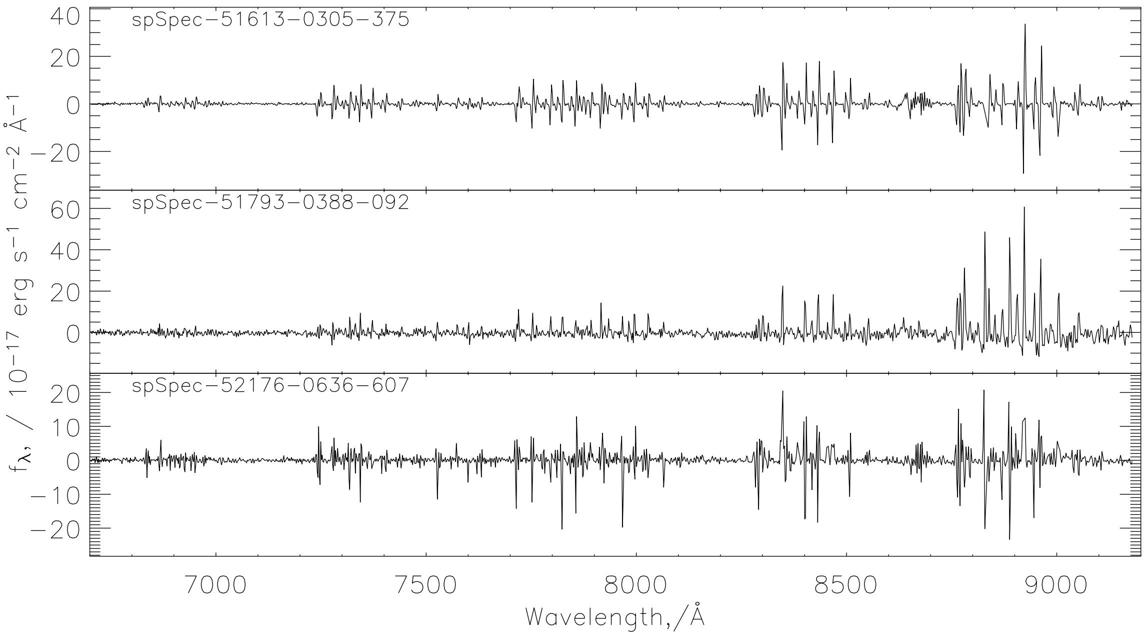

**Figure 4.2:** Examples of the red half of sky spectra taken directly from the SDSS DR3 catalogue. These spectra have been sky-subtracted along with all other spectra on their plates and should have zero flux. The SDSS sky-subtraction achieves excellent results in terms of the overall level of flux removed. Note however the well-defined patterns present due to the poor OH line subtraction, particularly the characteristic positive-negative residuals.

to the SDSS spectra after they have been sky-subtracted by the automated reduction pipeline, in particular to the removal of OH sky lines longward of $6700$Å. The aim in developing this technique is to increase the potential use of the red half of the SDSS spectra, of use to a wide variety of scientific investigations. Despite the specific task for which the method was first conceived, it is fully expected that only minor adaptations would be required before it could be applied to more general problems of fibre sky-subtraction with other instruments.

## 4.2 Preparing the sky spectra

The overall approach taken is determined by the extent of the spectroscopic information provided in a SDSS public data release. Typically four or five exposures are combined per spectroscopic plate to make up a single observation of a field and only the latter is publicly available. This precludes any possibility of improving the sky-subtraction on an exposure by exposure basis. However, the SDSS data releases do provide the final reduced spectra for all fibres on a plate. This includes not only those allocated to target objects, but also those fibres assigned to stars for spectrophotometric calibration and those positioned on blank sky regions and employed to define the underlying spectrum of the night sky. It is this latter set in which we are principally interested as they contain significant diagnostic information about the quality of sky subtraction for each spectroscopic plate.

Each SDSS spectroscopic plate contains 640 fibres. For each observation approximately 32 of these, 16 for each of the two spectrographs, are assigned to blank sky regions, selected from areas containing no detected objects in the SDSS imaging survey. The sky fibres are identified as “SKY” in the catalogue’s

“OBJTYPE” and “SPECCLASS” field. It is these 32 sky spectra that are combined to create the master-sky spectrum for each spectroscopic plate, which is then scaled and subtracted from each of the 640 spectra, producing 608 sky-subtracted object spectra and 32 sky-subtracted sky spectra, all of which are part of the standard SDSS data releases. Throughout the remainder of the chapter the sky-subtracted sky spectra will be referred to as “sky spectra”, and the sky-subtracted object spectra as “object spectra”.

The availability of large numbers of sky spectra, over 18,500 in DR2, thus provides a direct empirical measure of the noise resulting from counting statistics as well as any systematic residuals from the sky-subtraction procedure. Many scientific investigations making use of the SDSS catalogues rely on their sheer scale and this new technique is no exception. The application of empirical self-calibration techniques, based on the statistical properties of the data set itself, can be a powerful tool.

The principles behind this new sky-subtraction method are simple to understand: a PCA of the sky spectra produces a set of orthogonal components that provides a compact representation of the systematic residuals resulting from the sky-subtraction process. The components are then added in linear combinations to remove the systematic sky residuals from the spectra of target objects, such as galaxies and quasars.

The analysis focuses only on the red part of the SDSS spectra ($6700-9180$Å), as the strong emission features blueward of $6700$Å are small in number and easily masked. The upper limit of $9180$Å is set by the requirement that the entire wavelength region be included in all spectra when creating the PCA components from the sky spectra. Although techniques exist to overcome this (see Section 2.3.2), they would not greatly improve the applicability of the procedure for scientific objectives due to the small fraction of plates affected and the generally low S/N ratio at the far extent of the spectroscopic range.

### 4.2.1 A sample of sky spectra

The goal of the sky-subtraction scheme is to remove any systematic residuals from the bulk of the SDSS spectra. Pathological spectra with very unusual deviations can have a disproportionate effect on the generation of the PCA components. Therefore, a restricted subset of the full 18 764 sky spectra are identified for use in deriving the PCA components.

For the wavelength range used in the analysis ($6700 - 9180$Å) the spectra must satisfy the following criteria, where the numerical values are in the units of the SDSS spectra (10$^{-17}$ erg s$^{-1}$ cm$^{-2}$ Å$^{-1}$):

1. $-0.2 <$ mean flux $< 0.2$, ensuring the spectra have a mean close to zero

2. variance of the flux $< 0.8$, ensuring that the amplitude of pixel-to-pixel fluctuations is not unusually high

3. “spectral colours” have values $|a - b| < 0.1$, $|a - c| < 0.3$ and $|b - c| < 0.3$ ensuring that the spectra do not exhibit significant large scale gradients. $a$, $b$ and $c$ are calculated by averaging the flux in three wavelength regions largely free of OH sky emission (7000:7200Å [$a$], 8100:8250 Å [$b$] and 9100:9180 Å [$c$]).

Furthermore, only spectra with a minimum of 3800 good pixels (“NGOOD” parameter) over the entire wavelength range of the spectrum are accepted, to eliminate spectra with substantial numbers of missing pixels. Following a trial run of the PCA, 300 spectra are identified as outliers in the distribution

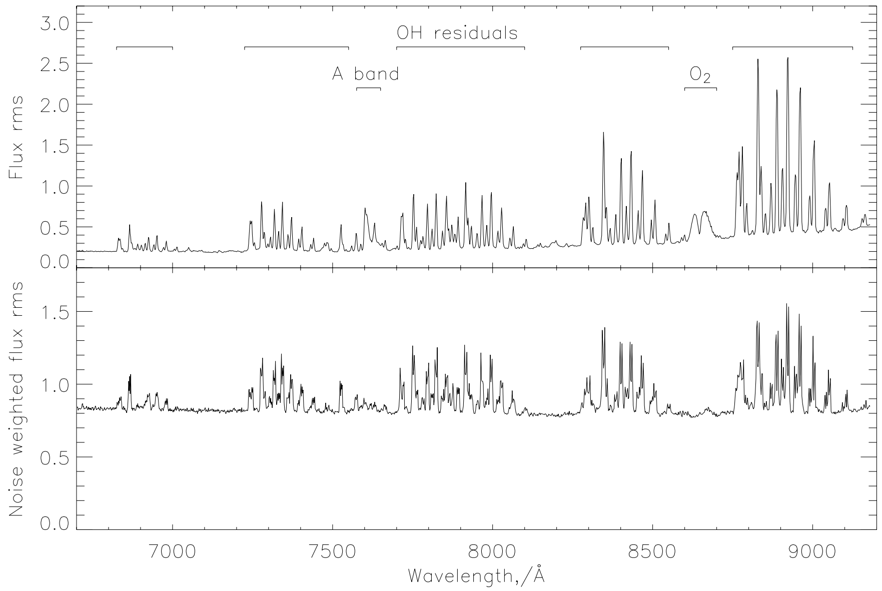

**Figure 4.3:** Top: the flux rms (67th percentile) of the 15 178 sky spectra. Bottom: same as top, with the flux normalised by the associated median SDSS spectroscopic plate noise arrays before the rms is calculated. The impact of the presence of the OH sky emission lines is evident from the increase in the rms (by nearly a factor of two at maximum), even after the empirical scaling of the noise arrays in the SDSS spectroscopic reduction pipeline.

of principal components, therefore dominating particular components, and discarded (see Section 4.3). Application of these selection criteria leaves 15 178 sky spectra. Finally, the median flux over the wavelength range $6700 - 9180$Åwas subtracted from the spectra. None of the final results depend on the details of the criteria used to define the subset of sky spectra to be used in the PCA-analysis.

### 4.2.2 Poisson error normalisation

The true error on each pixel in each spectrum is made up of components due to Poisson noise and systematic sky-subtraction errors. The method for removing sky residuals relies on determining the best-fit PCA-components via a minimum least-squares criterion over all pixels within the $6700 - 9180$Å wavelength range. A pixel which varies greatly between spectra will be weighted highly during the generation of the PCA-components and in the subsequent reconstruction of the sky-residuals present in individual spectra. For example, there will be a greater variance among spectra at wavelengths in the vicinity of strong OH emission lines, purely due to Poisson noise and independent of whether there is a contribution due to sky-subtraction errors. It follows that to achieve effective removal of the systematic errors it is important to normalise each spectrum by the Poisson noise expected at each pixel. Failure to do so results in over-subtraction for certain pixels, with "corrected" pixels apparently exhibiting fluctuations below the Poisson limit.

Each SDSS spectrum includes a companion error array based on the original Poisson errors derived from photon counts, CCD read-noise and so forth. Fig. 4.3 shows the flux standard deviation[1] (rms) of

[1] Unless otherwise stated estimates of rms amplitudes are taken to be the 67th percentile of the data. This is less sensitive to the presence of small numbers of extreme, non-Gaussian, outliers.

the 15 178 sky spectra with and without normalisation by their associated SDSS noise arrays. Throughout this chapter these arrays are referred to as the "flux rms array" and "normalised flux rms array" respectively. If the SDSS noise arrays accounted for all the variance present in the spectra, the lower plot in Fig. 4.3 would be flat. In fact the SDSS noise arrays account well both for the increased "continuum" noise at red wavelengths, the presence of the atmospheric A-band at $\sim 7600$Å and the contribution from the increased sky emission associated with the presence of the broad $O_2$ airglow emission centred on $\sim 8650$ Å. However, significant additional residual variance resulting from the incomplete subtraction of the barely resolved OH emission lines remains. Unfortunately for our application, the quoted error arrays have also been systematically increased in these regions of poor sky subtraction. This correction is carried out separately for each plate and details can be found in the data reduction source code at http://spectro.princeton.edu/idlspec2d_doc.html#SKYSUBTRACT. For many applications the scaled error arrays in the data releases, which better reflect the true error in each spectrum, are an improvement. However, using the modified SDSS noise arrays for the sky-subtraction procedure does not allow it to reach its full potential, due to the excessive down-weighting of systematic fluctuations associated with strong OH features.

### *Rescaling the SDSS noise arrays*

It is not possible to recover directly the original Poisson errors that are required; so instead, an empirical approximation must be developed. The files for several spectroscopic plates (0412, 0725 and 0745) prior to sky subtraction, exposure combination and rebinning onto a common wavelength scale, were kindly made available to us by the SDSS project (E. Switzer, P. Macdonald and D. Schlegel). From these we could investigate precisely the effect the SDSS pipeline program SKYSUBTRACT has on the data. All the sky spectra on a single plate are offset in wavelength space slightly due to instrumental effects, so each samples slightly different parts of the underlying sky spectrum. By sorting all the pixels contained in all the sky spectra by their exact wavelength, SKYSUBTRACT creates the supersampled master sky spectrum. This is then fitted with a B-spline in order to create an analytic function which can be solved for each fibre individually, and subtracted leaving the sky subtracted object spectra.

More importantly for our needs, the files provided by the SDSS team contained both the original and rescaled error arrays for the fibres. From these we could calculate the rescaling applied, compare with the master sky spectrum and learn whether the error arrays around the OH lines had been systematically increased compared to the rest of the spectrum. This is certainly the case for some exposures on some plates, although the pattern is not clear in all cases. Fig. 4.4 shows, for one particular exposure and fibre number, the input error array (objvar), master sky error array (skyvar), relative $\chi^2$ between the object flux array and supersky B spline fit, the output error (newvar) and the overall rescaling applied to the error array. It is clear in this case that the OH lines have been systematically rescaled.

The rescaling is calculated by

$$rescaling \equiv \frac{newvar}{objvar} = 1 + \frac{(\chi^2 - 1) \times skyvar}{objvar} \tag{4.1}$$

indicating that the rescaling will be large either when the B-spline fitted master sky spectrum is a poor fit to the fibre flux, or the variance of the master sky spectrum is large. The system of B-spline fitting

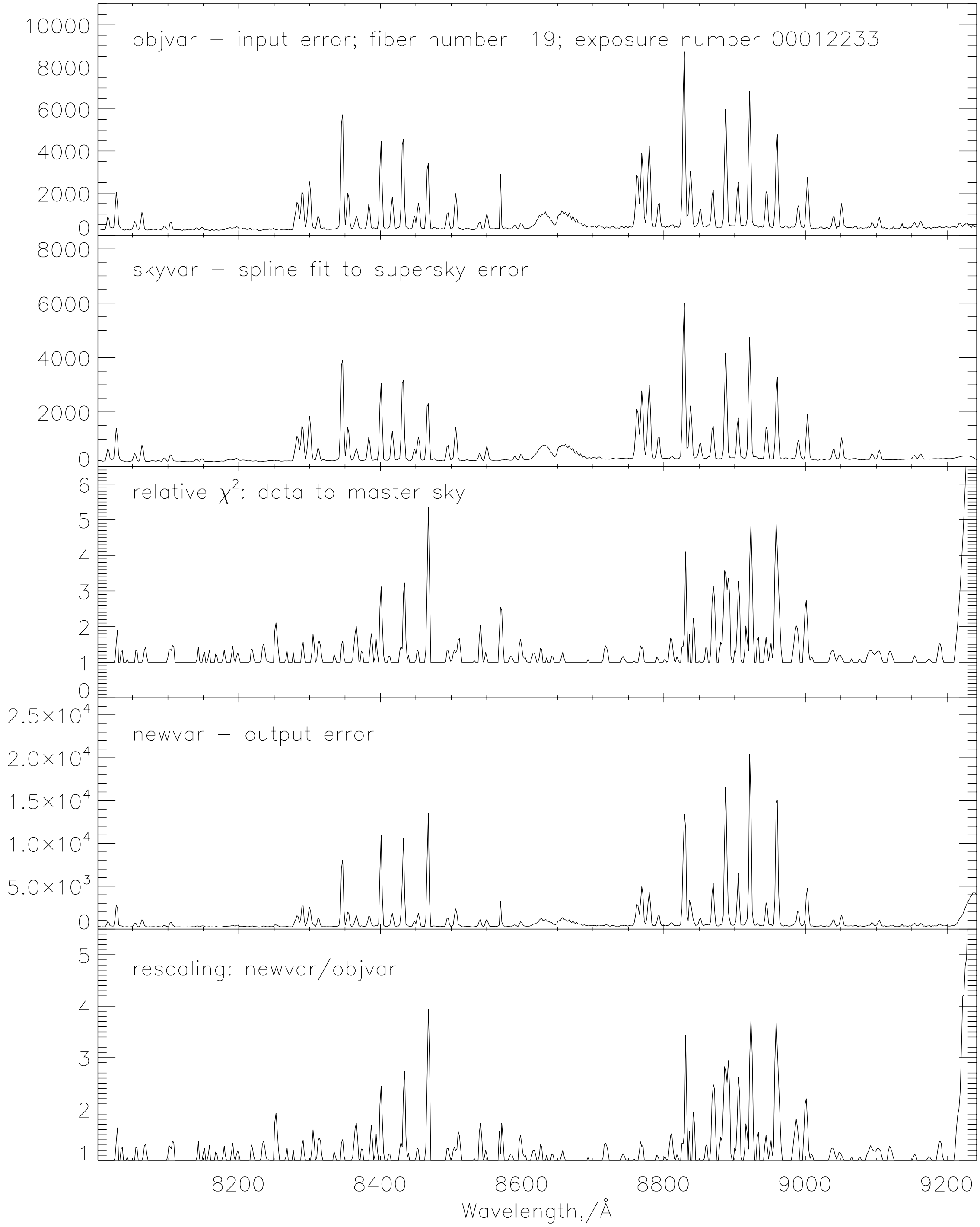

**Figure 4.4:** Analysis of raw SDSS spectroscopic plate 725, exposure ID 00012233 and fibre number 19. From top to bottom the input error array (objvar), master sky error array (skyvar), relative $\chi^2$ between the object flux array and supersky B-spline fit, the output error (newvar) and the overall rescaling applied to the error array for a small portion of the wavelength range. It is clear that the $\chi^2$ is increased at the positions of the OH lines. This causes the error arrays to be rescaled to a greater extent than in other regions.

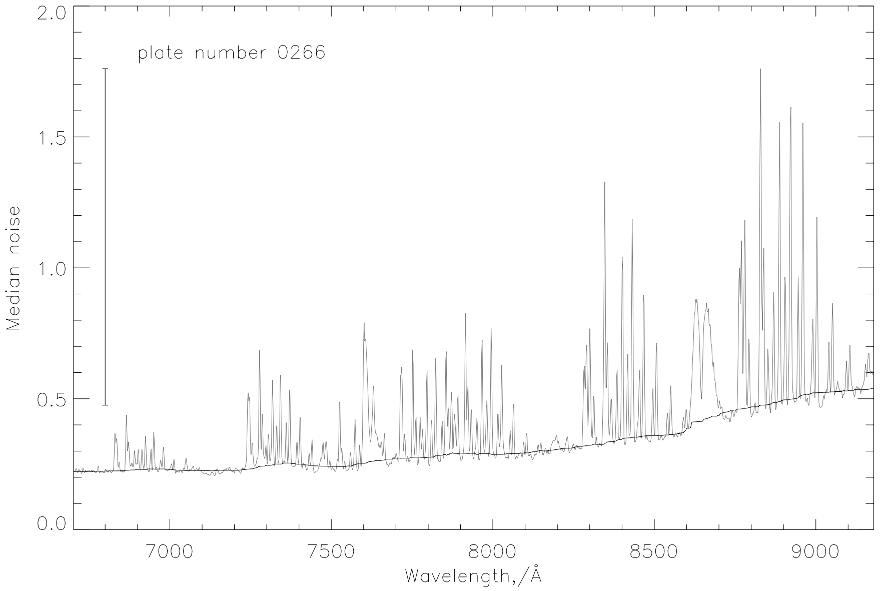

**Figure 4.5:** The median noise array for plate number 0266 (grey) and the estimated continuum overplotted (black). The bar to the left hand side indicates the amplitude of max$(n - c)$ for this plate, which occurs at the wavelength of $8829$ Å.

the master sky spectrum may be partly to blame as the B-splines may struggle with the peaks of the sharpest OH features. Alternatively the strengths of the OH lines are varying greatly between the fibres, causing a poor $\chi^2$ between master sky and flux. Note that random fluctuations of the $\chi^2$ value will pick out those pixels with large errors in the master sky spectrum - predominently the OH features, due to the dependence of the rescaling on the master sky error array (skyvar). Most likely, all these effects combine to produce the final rescaling. It is further observed that the normalised flux rms array show each OH line to be double peaked (Fig. 4.3 bottom panel), indicating that the centers of the lines appear to have been rescaled to a greater extent than their sides. This may provide clues as to the dominant factor in the rescaling, however further investigation was beyond the requirements of this project.

The same noise scaling is applied to all spectra on a plate, allowing derivation of a single rescaling for each one of the SDSS plates in the following way. For each plate we take the median at each pixel of the error arrays of the 32 sky spectra, this provides a good reference array with which to estimate the rescaling to be applied to individual error arrays. From the reference array a “continuum” noise is calculated over the full wavelength range of interest, interpolating across pixels containing OH sky emission lines. The amplitude of this continuum noise is dominated by counting statistics; it rises towards the red (Fig. 4.3; upper panel) due to an increase in the sky level and decrease in CCD sensitivity. This rise must be taken into account as no rescaling of this continuum noise is required. A simple function is then used to approximate the rescaling of the noise, $S_i$, above the continuum for each pixel $i$ on each plate:

$$S_i = 1 + \left[\frac{(n-c)_i}{\max(n-c)}\right]^\alpha \times \beta \tag{4.2}$$

where $n$ is the median noise of the plate, $c$ the continuum noise, and max$(n - c)$ the maximum noise,

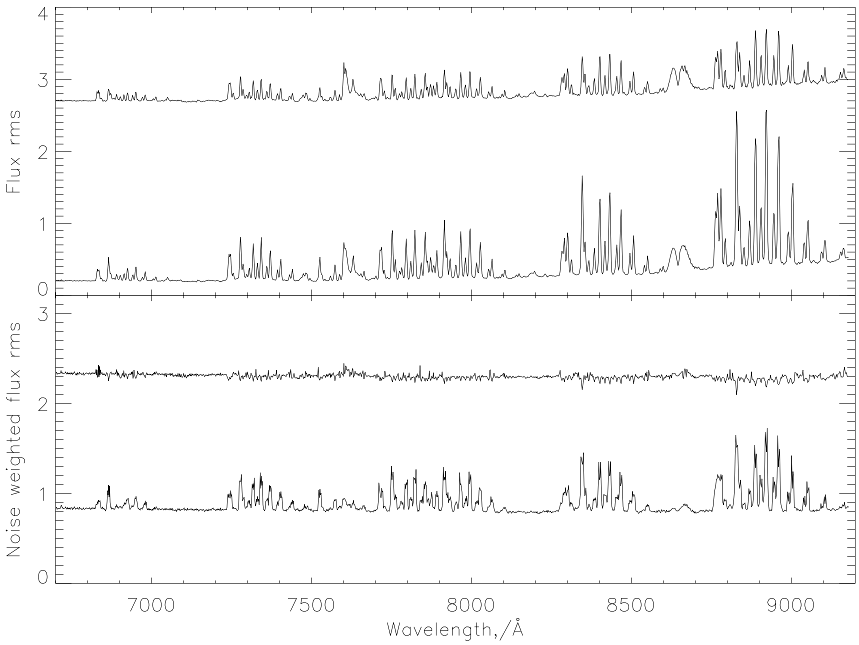

**Figure 4.6:** Top: the flux rms of the 15 178  sky spectra. The lower (upper) spectrum is before (after) the sky-residual subtraction is performed on the individual flux arrays. Bottom: same as top, with flux normalised by the associated median SDSS spectroscopic plate noise arrays, scaled according to Eq. 4.2 ($\alpha = 1, \beta = 0.3$). For clarity, the upper spectra have been displaced vertically by 2.5 units (upper panel) and 1.5 units (lower panel). The effectiveness of the sky-residual subtraction procedure is seen by the reduction by more than half of the systematic spikes in the top figure, and complete removal in the bottom figure.

above the continuum level, over all pixels. As the exact effects of the SDSS noise scaling are not understood, the form of this function is chosen to provide flexibility in both the total rescaling and dependence with wavelength. Fig. 4.5 shows an example plate median noise array, with estimated continuum and max$(n - c)$ marked. $\alpha$ and $\beta$ are determined empirically as described below and represent, respectively, the relative fraction of rescaling between pixels with differing amounts of noise, and the total rescaling of the pixel with the highest noise above continuum level.

Anticipating the results of Section 4.4, the effectiveness of the PCA sky-residual subtraction procedure can be seen in Fig. 4.6. The top panel shows the effect of the sky-residual subtraction on the flux rms array of the sky spectra: the lower spectrum shows the rms as a function of wavelength with no sky-residual subtraction performed on the flux arrays (reproducing the top spectrum from Fig. 4.3); the upper spectrum shows the flux rms array after application of the sky-residual subtraction scheme. Note how the features due to the atmospheric A-band ($\sim 7600$Å) and $O_2$-emission ($\sim 8450$ Å) are unaffected, while the height of the spikes associated with the OH emission lines is greatly reduced. The variation with wavelength of the rms, post application of the sky-residual subtraction, is explained almost entirely by counting statistics. The bottom panel of Fig. 4.6 shows the flux rms array of the sky spectra each weighted by the modified noise arrays (with $\alpha = 1, \beta = 0.3$): the lower and upper spectra show the rms before and after subtraction of the sky residuals respectively.

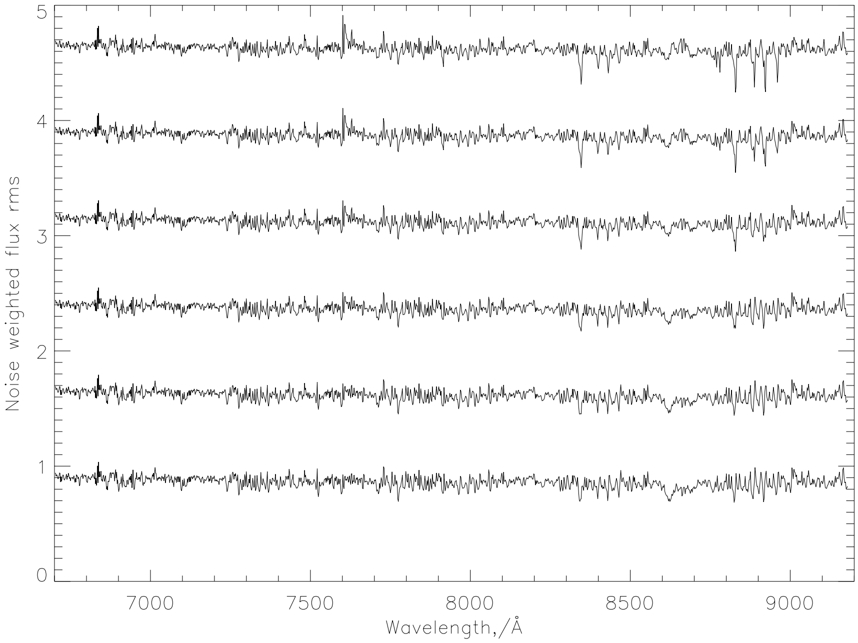

**Figure 4.7:** Flux rms of the 15 178 sky spectra after sky-residual subtraction is performed, with flux normalised by the associated median SDSS spectroscopic plate noise arrays, scaled according to Eq. 4.2. From bottom to top $\beta = 0, 0.1, 0.2, 0.3, 0.4, 0.5$ and $\alpha = 1$.

### *Estimating the rescaling constants*

Fig. 4.7 shows the effect of varying the noise rescaling ($\beta$ in Eq. (4.2)) applied to the sky spectra prior to creating the PCA eigenspectra, on the final rms of the 15 178 sky spectra after the sky-residual subtraction has been performed. From bottom to top $\beta = 0, 0.1, 0.2, 0.3, 0.4, 0.5$. For low values of $\beta$ (small amounts of rescaling of the noise arrays) the noise weighted flux rms array is approximately constant with wavelength, i.e. the sky-residual subtraction procedure is not removing signal below the Poisson limit for any pixels, although the procedure may also not be removing the full systematic contribution to the errors. As $\beta$ increases, initially the rms remains constant until a point is reached where significant dips appear in the red half of the spectra. The presence of these dips show that an unphysical situation has resulted, where, following the PCA sky-subtraction, the noise falls below the Poisson limit for those pixels. The same effect is seen if the sky residual subtraction is undertaken without weighting the flux by the noise arrays, i.e. the noise is constant, independent of wavelength The fact that the noise arrays can be rescaled by some amount before these unphysical dips appear, indicates that the SDSS noise arrays have indeed been scaled to take account of the presence of a systematic contribution to the noise at wavelengths where the OH-emission is significant.

Under-subtraction of the sky-residual signal due to over estimation of the Poisson errors prevents the sky-residual subtraction method from reaching its full potential. Over-subtraction of the sky-residual signal causes downward spikes to appear at the location of the OH emission lines (Fig. 4.7), as we begin to erroneously subtract Poisson noise. This suggests a simple method to estimate the values of $\alpha$ and $\beta$ in Eq. 4.2: the weighted flux rms, post sky-residual subtraction, is required to be as constant with wavelength as possible whilst maximising the noise rescaling. Taking a grid of values for $\alpha$ and $\beta$, for

**Table 4.1:** Standard deviation of the flux rms array of the 15 178 sky spectra post sky-residual subtraction and weighted by the rescaled SDSS noise arrays, as a function of $(\alpha, \beta)$ used during the rescaling.

| $\alpha\backslash\beta$ | 0.0 | 0.3 | 0.5 |
|---|---|---|---|
| 1.00 | 0.035 | 0.032 | 0.039 |
| 0.75 | 0.035 | 0.034 | 0.045 |
| 0.50 | 0.035 | 0.038 | 0.059 |

each combination the PCA is run on the 15 178 sky spectra with noise normalisation scaled according to Eq. (4.2), then the sky-residual subtraction is performed on each sky spectrum keeping the same noise weighting.

By calculating the standard deviation of the resulting flux rms array for each value of ($\alpha$, $\beta$) the best empirical rescaling of the SDSS noise arrays can be derived: we require the greatest value of $\beta$ that does not introduce upward or downward spikes, whilst $\alpha$ allows purely for potential wavelength variations. Table 4.1 presents these measurements for a selection of values of $(\alpha, \beta)$, and $\alpha = 1$ and $\beta = 0.3$ results in the minimum rms whilst maximising $\beta$. This implies that by systematically decreasing the amplitude of the noise arrays by 30% at positions of OH lines the effectiveness of the sky-residual subtraction can be substantially increased, without over-subtracting residual sky features beyond the limit set by counting statistics. The final results of this chapter are insensitive to small changes in $\alpha$ or $\beta$, and the measured value of $\alpha = 1$ suggests that the empirical rescaling of the noise arrays within the SDSS spectroscopic pipeline is predominantly a function of the height of the OH spikes above the continuum noise.

It should be emphasised that it is not possible to derive the true noise array, based on counting statistics, for each SDSS spectrum from the information contained in the publicly available SDSS data releases. However, the close to constant rms as a function of wavelength achieved following the empirical rescaling of the noise arrays described above (Fig. 4.6; bottom panel) is strong evidence that, while approximate, the scheme does achieve a highly effective correction, reducing the rms close to the level set by counting statistics at all wavelengths. If a version of this sky-residual subtraction scheme could be applied to the SDSS spectra using the noise arrays based on counting statistics then further improvements should result.

### 4.2.3 Identification of sky pixels

Finally, before applying the PCA to the sky spectra, those pixels that require correction must be identified, as the strong sky emission lines which lead to the presence of systematic sky-residuals occupy only a fraction of the $6700 - 9180$ Å wavelength range. The affected pixels can be found via the calculation of the rms of the 15 178 sky spectra, normalised by the corresponding SDSS noise arrays (as shown in the lower panel of Fig. 4.3). All pixels above a threshold value are defined as "sky pixels". The "non-sky pixels" are not included in the PCA sky-residual removal procedure but provide an empirical estimate of the level of Poisson noise (i.e. excluding systematic sky-residuals) present in each spectrum. Adopting a threshold level of 0.85 for the sky/non-sky boundary results in 670 sky pixels and 697 non-sky pixels. The results shown in subsequent sections are not sensitive to reasonable variations in the threshold level. Note that the bottom panel of Fig. 4.3 shows the average rms of the spectra normalised by the median noise arrays for their plates to lie below unity. The SDSS team are aware of this effect but the cause remains unclear. It does not effect our method except for the positioning of the threshold level for sky/non-sky pixels.

## 4.3 Creating the principal components

Having created a sample of sky spectra containing only those pixels with sky signal, and weighted by the empirically derived noise array for the relevant plate, the PCA (see Chapter 2) is run. The output consists of $M$ principal components which are $M$ pixels long, where, in this case $M = 670$, the number of sky pixels. Each component has an associated variance, which gives the percentage of the total variance of the dataset contained in that component. During calculation of the PCA components, sky pixels in each spectrum with an associated error of zero (i.e. no data) are set to the mean value of that pixel in all other sky spectra in order to minimise their effect.

As PCA can be sensitive to spectra with particularly unusual features in which we are not interested (as mentioned in Section 4.2.1), spectra are removed which dominate the signal in individual components after one trial run. The pruning, of 300 spectra in total, is achieved by removing those spectra with principal component amplitudes more than $5\sigma$ from the mean. The PCA is then rerun on the resulting sample of sky spectra. Figs. 4.8 and 4.9 show the first 10 principal components resulting from the analysis. Note the distinctive, correlated features present in each component.

Once the components are created, they can be used as templates to reconstruct the input sky spectra and later the sky residuals in the object spectra (Section 4.3.2). This is done by projection of each spectrum onto the components and summation of the components weighted by the projection coefficients (Eq. 2.4 and 2.5). The reconstruction may then be subtracted from the sky or object spectra leaving residual-free spectra, termed sky-residual subtracted spectra.

### 4.3.1 The number of components

As with most applications of PCA, the number of components to use in the reconstruction is not well defined. The use of too many components results in the artificial suppression of noise below the Poisson limit, with the PCA acting as an (undesirable) high-frequency filter. The use of too few components means that the removal of sky residuals is sub-optimal and, in some cases, the overall quality of the spectra can decrease.

Fig. 4.10 shows the mean ratio of the rms of the noise-weighted flux in the sky pixels to that in the non-sky pixels[2], for the 15 178 sky-residual subtracted sky spectra, as a function of the number of components used in the reconstructions. The rms of the non-sky pixels in each spectrum remains constant and the ratio decreases monotonically as more components are used in the reconstructions, with an increasing fraction of the noise in the sky pixels removed.

The reduction of the noise in the sky pixels below the noise in the non-sky pixels is clearly unphysical, and the number of components to employ in the reconstruction of each spectrum is therefore estimated by adopting the non-sky pixel rms of that spectrum as a reference. The reconstruction of a spectrum proceeds one component at a time, with the rms ratio calculated for the sky-residual subtracted spectrum. Reconstruction is stopped when the rms ratio reaches unity, i.e. when the noise weighted flux rms is the

[2] Calculation of the non-sky rms requires a robust estimator unaffected by non-Gaussian outliers, whereas the sky rms must remain sensitive to outliers. Therefore, while a 67th percentile rms is used for the former, the standard deviation of the data is used for the latter.

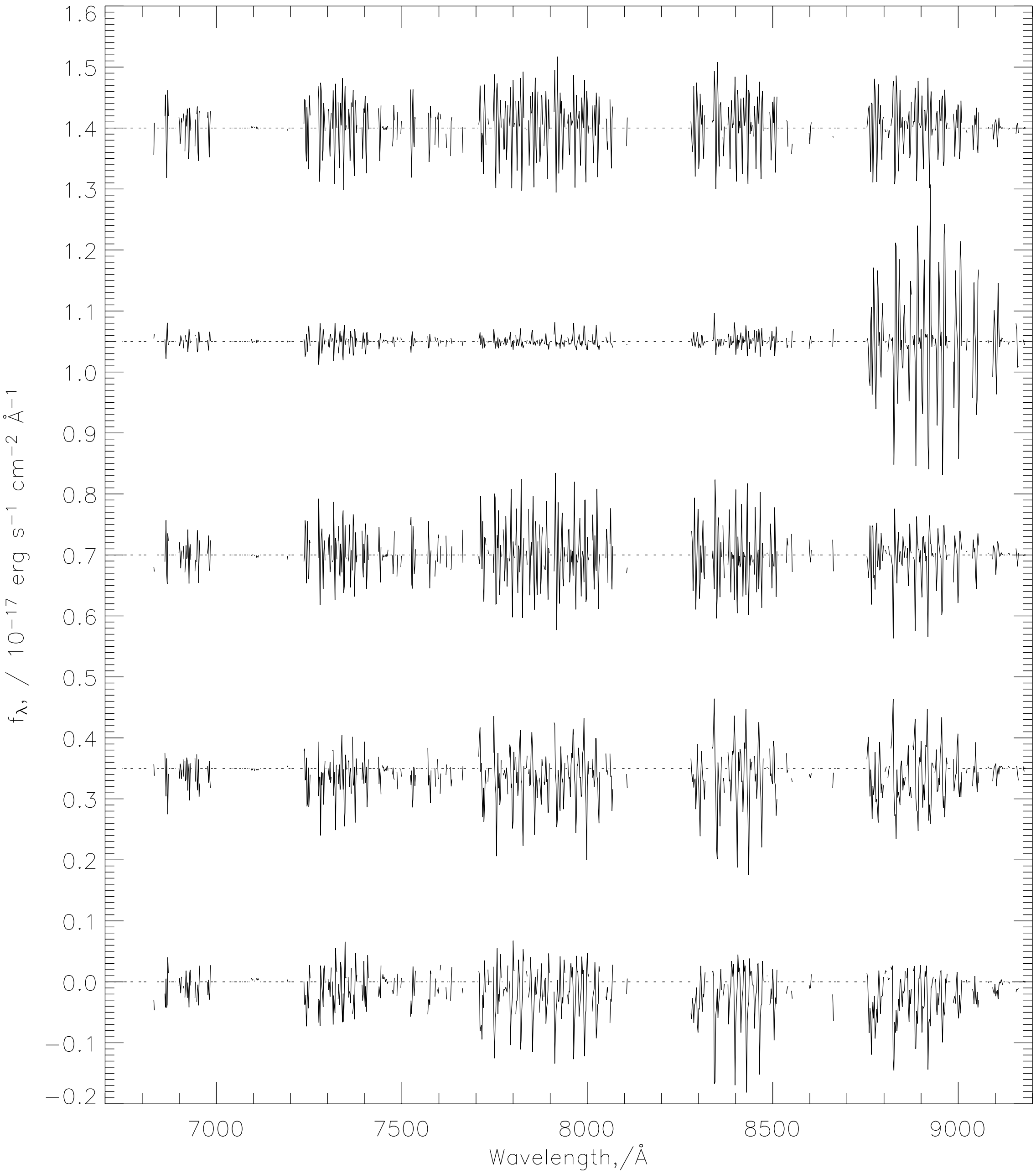

**Figure 4.8:** From top to bottom, the first five sky principal components, each offset vertically from the next by a value 0.35. A horizontal dashed line indicates zero flux in each component.

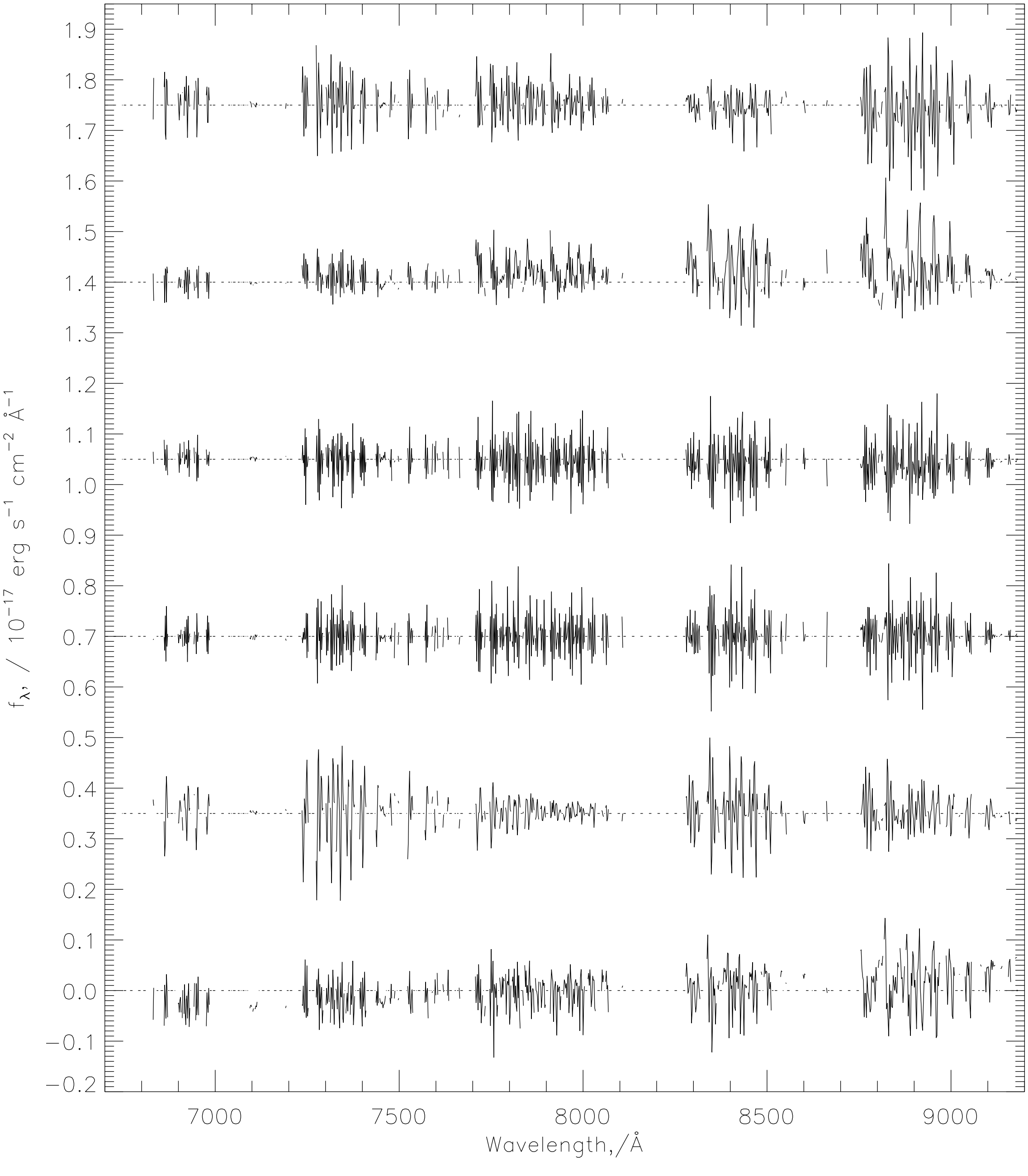

**Figure 4.9:** Same as Fig. 4.8, except the 6th-10th principal components are shown from top to bottom.

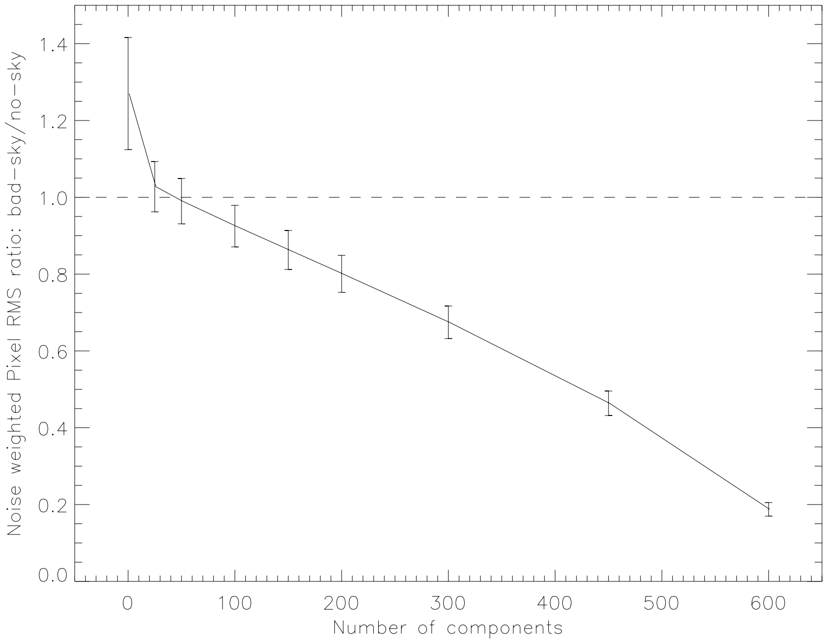

**Figure 4.10:** The mean ratio of sky to non-sky flux rms over all sky spectra in the sample as a function of the number of components used during sky residual reconstruction. Fluxes are weighted by their respective scaled plate noise arrays. The horizontal dashed line indicates where the two rms values are equal. The error bars show the standard deviation of all objects in the sample.

same for the sky and non-sky pixels. The scheme is largely self-calibrating. For example, in spectra with high S/N the systematic sky residuals typically contribute only marginally to the sky-pixel noise, the rms ratio thus starts close to unity and only a small number of components are necessary to achieve equality in the rms ratio.

The large number of pixels that contribute, combined with the robust estimator of the non-sky rms, mean the amplitude of the non-sky rms is well determined, providing an excellent indicator of when to halt the reconstructions. Fig. 4.11 shows the distribution of the number of components used for DR2 sky, galaxy and quasar spectra in the final reconstructions. For bright objects the contribution to the noise from imperfections in the sky subtraction is small and few components are required to bring the sky and non-sky rms to the same level. The presence of many spectra with high S/N produces the “spike” at low component numbers in both the galaxy and quasar sample histograms. The fainter limiting magnitude of the quasar sample make sky-residuals a problem in a larger fraction of the spectra and, on average, more components are required to remove the systematic errors present following the default SDSS-pipeline sky-subtraction. To limit the effect of very occasional poor estimation of the non-sky rms the maximum number of components is set to 150 and 200 for the galaxy and quasar samples respectively.

### 4.3.2 Reconstructing sky residuals in an object spectrum

The preceding sections have developed a procedure that achieves a substantial reduction in the amplitude of the systematic sky residuals in spectra that possess no large scale signal, i.e. the sky spectra are known, by definition, to possess zero signal at all wavelengths. We now turn to the more interesting application

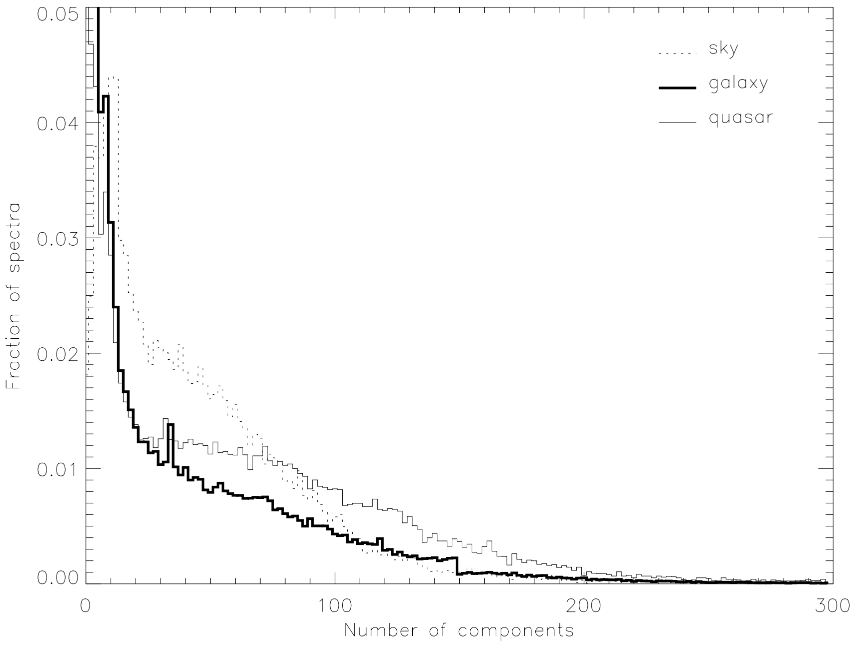

**Figure 4.11:** Histograms showing the number of components used in the final reconstruction of the sky signal in the sky (dotted line), galaxy (thin line) and quasar (thick line) DR2 samples. The y axis is truncated slightly in order to show the majority of the objects, 18% (5%) of galaxies (quasars) require zero components.

of removing the systematic sky-residuals from the SDSS science spectra of targets such as galaxies and quasars.

The problem is made tractable by the success of the sky-subtraction procedure performed as part of the standard SDSS spectroscopic pipeline. Detailed examination of the properties of the sky spectra shows that both the mean level and the large scale shape of the sky spectrum to be subtracted from each spectrum have been determined to very high accuracy. As a result, there is effectively no sky “continuum” to remove, rather the problem is confined to removing the high-frequency structure due to the presence of the strong OH sky emission lines. The key goal is to ensure that object continuum and intrinsic absorption and emission line features are not removed as a result of the procedure used to reduce the amplitude of the systematic sky-residuals.

As it is not necessary to identify any sky continuum present, the continuum of the object spectrum can be removed using a median filter before projection of the spectrum onto the principal components derived from the sky spectra. The smallest filter size (about 40 pixels) can be found from median filtering the sky spectra as it is important that the filter is unaffected by the OH residuals. As the filter size is increased above about 40 pixels, an increasing number of features intrinsic to the object fail to be removed. This generally causes an increase in the rms of the non-sky pixels used as a reference point to halt the reconstruction (see Section 4.3.1) and therefore a decrease in the number of components used and slight decrease in the effectiveness of the sky-residual subtraction procedure. However, the final effect is small and insignificant for a reasonable range of filter sizes (up to about 80 pixels). A filter size of 55 pixels has been used throughout this chapter. In certain circumstances, it may be desirable for users of this procedure to reduce the filter scale in order to follow the broad emission line profiles in quasars closer to the line centroids.

Narrow line features present the greatest challenge, as these can be mistaken for an OH sky residual by the PCA. The advantage of PCA in this task is that it looks for patterns in the spectrum, linking lines together which are correlated in the input sky spectra, and weights all bins equally. However, if an emission or absorption feature occurs exactly at the location of an OH line, some combination of principal components can sometimes be found to reconstruct the feature, without increasing the rms in the rest of the spectrum significantly. Such behaviour is particularly likely if the line feature lies in a noisy part of the spectrum. For most applications in which the sky-residual subtraction scheme might be employed, it is possible to mask known emission and absorption features, thereby circumventing the problem. Section 4.5.1 shows how such masked features are unaffected by the sky-residual subtraction procedure. The disadvantage is that real sky features which happen to fall in the masked regions do not contribute to the projection onto the principal components and the sky-residual subtraction may not be fully effective in these regions. In some potential applications, it is not known in advance where absorption and emission features may occur and such an example is presented in Section 4.5.2. Features are masked by removing the relevant pixels during projection onto the principal components. Alternatively, replacing the pixels with the local mean results in almost identical reconstruction. Bad pixels are similarly masked (where the SDSS noise array is set to zero).

Once the object spectrum has been continuum subtracted, it is divided by the derived noise spectrum of the relevant plate and projected onto the sky principal components (Eq. 2.4), leaving out those pixels identified as possibly containing emission and absorption features. The resulting principal component amplitudes are used to reconstruct the residual sky-subtraction signal (Eq. 2.5) and the reconstruction is subtracted from the object spectrum, including from previously masked pixels. Re-multiplication by the noise spectrum, followed by the re-addition of the object continuum, returns the object spectrum cleaned of OH sky emission line residuals. Fig. 4.12 illustrates the process of sky-residual subtraction on a typical SDSS galaxy spectrum.

## 4.4 Application of method to SDSS spectra

Each class of spectroscopic science targets has different spectral characteristics. The effective application of the sky residual subtraction scheme requires the removal of the large scale “continuum” signal from the target object and the identification of wavelength intervals where narrow emission or absorption features may be present. The extremely high identification success rate achieved in the SDSS means that the wavelengths of strong absorption and emission features in the spectra of stars and galaxies are known; there are only 3 002 spectra without secure identifications among the 329 382 object spectra included in DR2. Potential systematic biases in the sky residual subtraction can be prevented by masking the wavelength intervals that include such features.

The exact nature of the pre-processing applied to spectra prior to the implementation of the sky-residual subtraction depends on the scientific goal. However, in subsequent sections schemes are described that are likely to have wide application in the analysis of the three main classes of SDSS science targets: galaxies, quasars and stars.

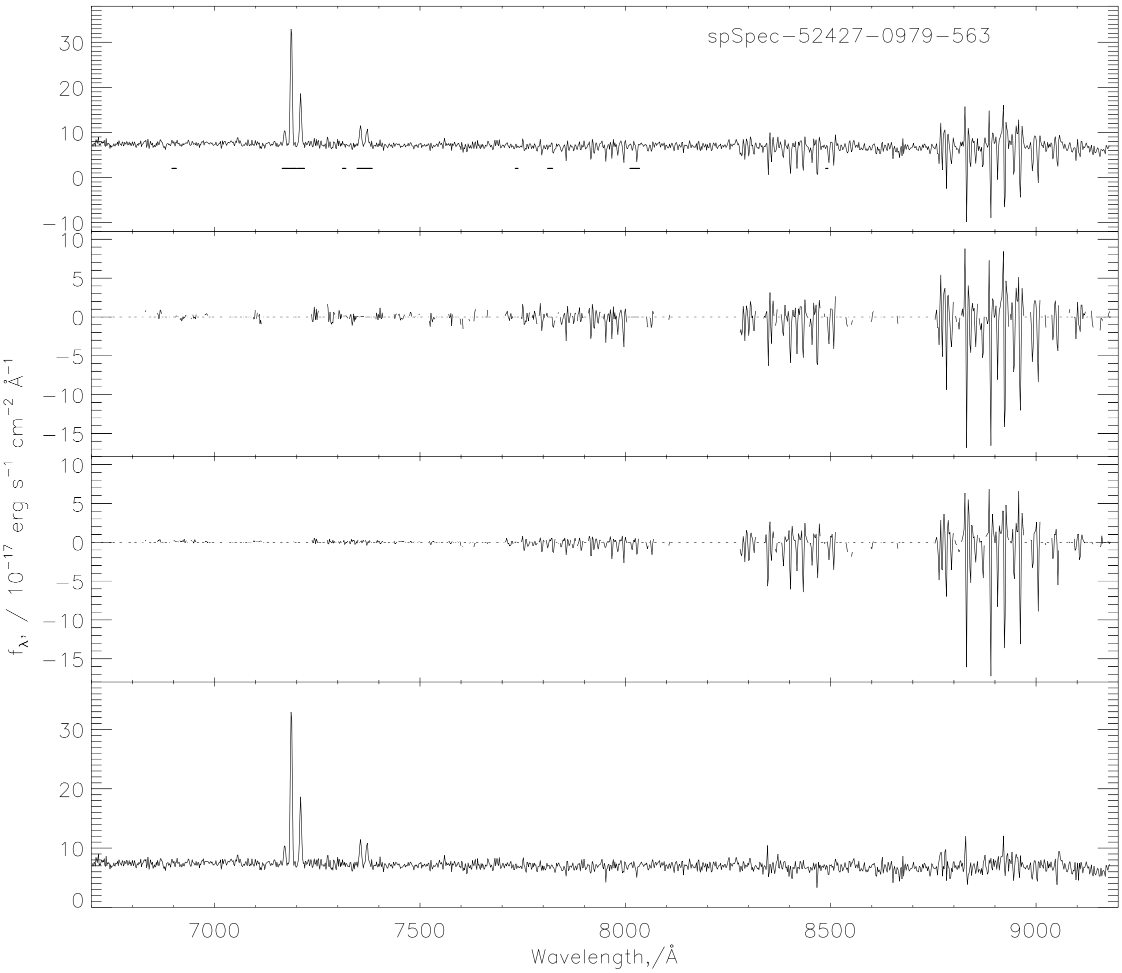

**Figure 4.12:** 1 (top): The wavelength region $6700 - 9180$ Å, of the SDSS galaxy spectrum spSpec-52427-0979-563. Pixels included in the feature mask are indicated by black horizontal bars under the spectrum; 2: The same spectrum but showing only "sky" pixels, regions between sky pixels are joined by a dotted line; 3: The reconstructed sky spectrum; 4 (bottom): The final sky-residual subtracted galaxy spectrum.

### 4.4.1 Galaxies

The DR2 catalogue contains 249 678 unique objects spectroscopically classified as galaxies. The strong systematic trends in the absorption- and emission-line properties of the galaxies as one moves from early through to late type objects necessitates a sub-classification of the population prior to the application of the sky-residual subtraction. Therefore, each galaxy is classified as either an absorption line, emission line or extreme emission line object. The spectral type parameter ("ECLASS") in the SDSS catalogue, which is derived from a preliminary PCA type analysis of the spectra by the SDSS team, provides the basis for the absorption line (ECLASS$< -0.05$) and emission line (ECLASS $\geq -0.05$) classification. The extreme emission line galaxies are identified through the presence of emission lines with very large observed frame equivalent widths (EW) ($\mathrm{EW}_{\mathrm{H}\alpha} > 200\,\mathrm{\AA}$, $z < 0.4$; or $\mathrm{EW}_{[\mathrm{O\ iii}]\lambda5008} > 200\,\mathrm{\AA}$, $z < 0.84$). These values are calculated during the standard SDSS reduction and provided with each object spectrum. An appropriate feature mask is then applied depending on the galaxy type (e.g. see Fig. 4.12). A total of 331/286/469 Å are masked in absorption/emission/extreme emission line objects respectively, over the entire rest-frame wavelength range of $3830 - 9100$Å, i.e. in the observed-frame wavelength range of $6700 - 9180\,\mathrm{\AA}$ only a small fraction of pixels are affected by the line masks. Fig. 4.13 shows examples of galaxy spectra before and after sky residual removal.

### 4.4.2 Quasars

The DR2 catalogue contains 34 674 unique objects spectroscopically classified as quasars or high- redshift quasars. A feature mask includes narrow emission lines (e.g. [O III] $\lambda\lambda4961, 5008$) and $70$Å intervals[3] (observed frame) centred on the broad emission lines (e.g. C IV $\lambda1550$). The wings of the broad emission features can, in practice, be regarded as "continuum" in the context of the sky residual subtraction as they are removed by the median filtering. Fig. 4.14 shows examples of quasar spectra before and after sky subtraction.

### 4.4.3 Stars

There are 42 027 unique objects classified as stars or late-type stars in the DR2. Due to their single redshift, application of the sky-subtraction procedure is even simpler. Depending on the final science required and the type of stars involved, a suitable feature mask and continuum estimation can be straightforwardly derived but our adopted median filter scale of 55 pixels, employed for the galaxies and quasars, produces very satisfactory results. Fig. 4.15 shows examples of stellar spectra before and after sky subtraction.

## 4.5 Tests of method on SDSS science objects

This section presents the quantitative results of applying the sky-residual subtraction to a variety of object spectra.

[3] Corresponding to approximately half the filter size in this wavelength range

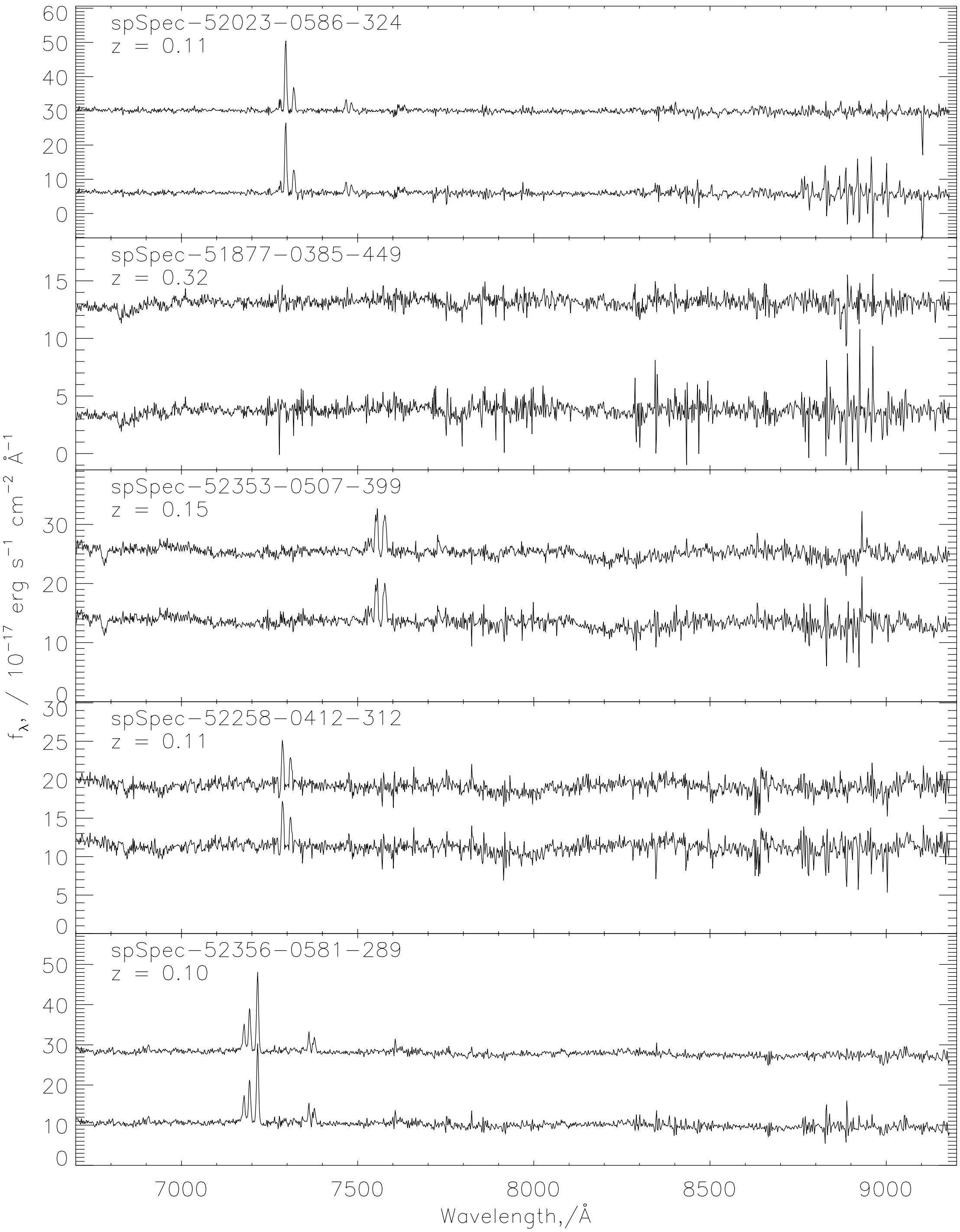

**Figure 4.13:** Examples of sky-residual subtraction applied to typical galaxy spectra effected by OH sky residuals. In each panel the lower spectrum is the raw SDSS data and the upper spectrum is after application of the sky-residual subtraction procedure. The upper sky-residual subtracted spectrum is offset for clarity.

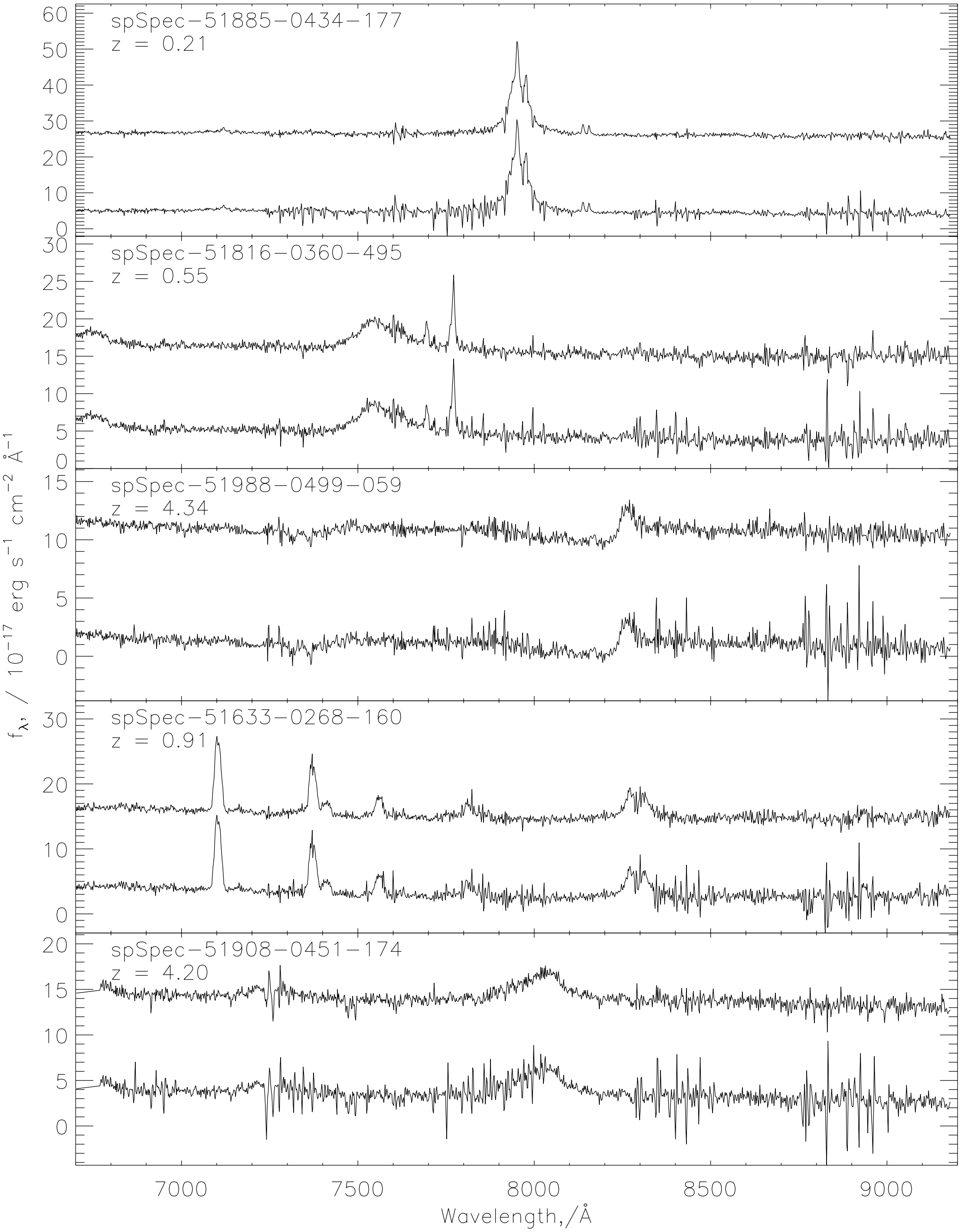

**Figure 4.14:** Same as Fig. 4.13 for quasar spectra.

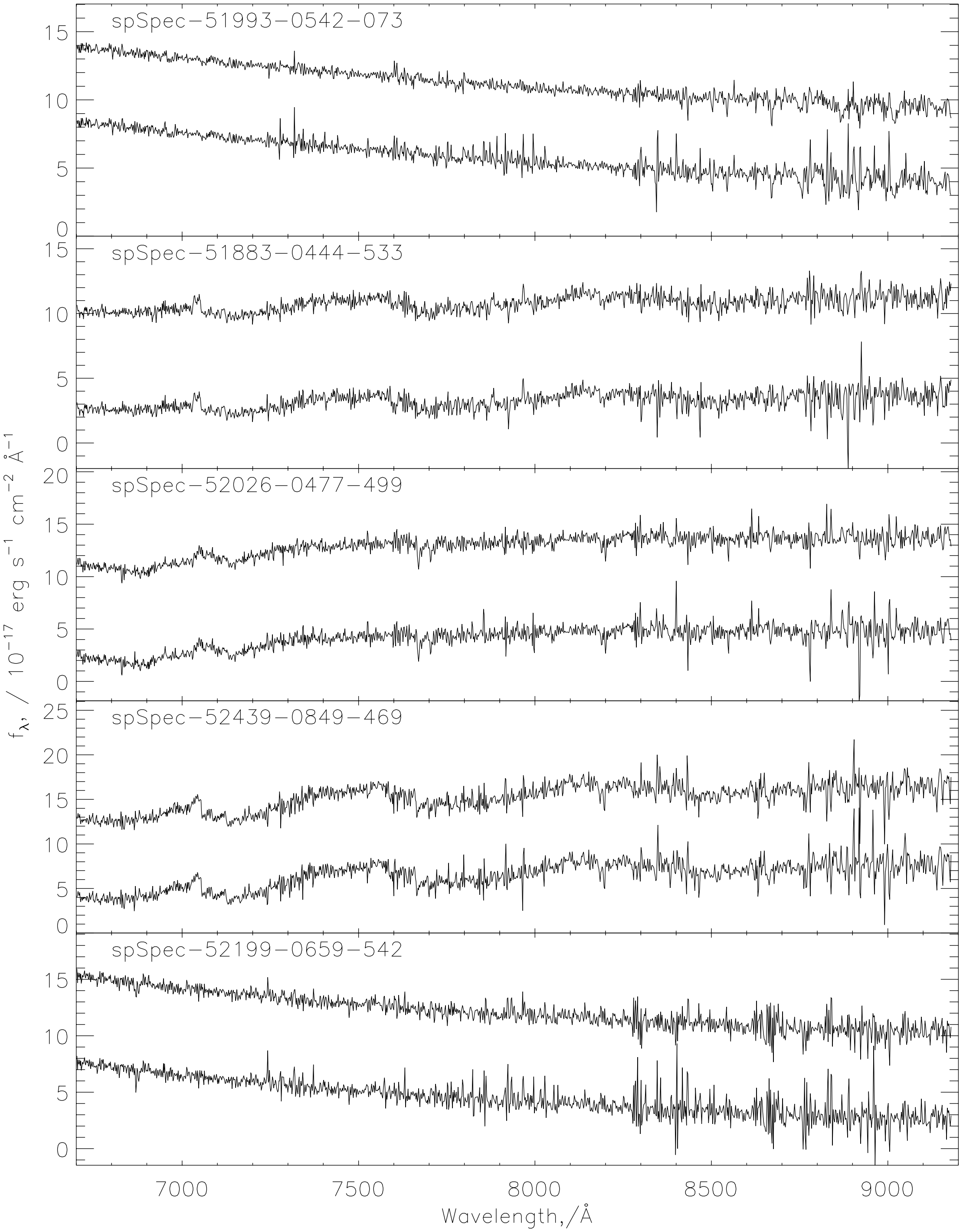

**Figure 4.15:** Same as Fig. 4.13 for stellar spectra.

### 4.5.1 Galaxy absorption features: The Ca II triplet

The far red spectrum of galaxies is where old stellar populations emit most of their light. Contamination by light from hot, young stars and any blue “featureless” AGN continuum is reduced compared to shorter wavelengths. The impact of Galactic reddening is also much reduced (Rutledge et al. 1997). The Ca II triplet ($8500.4, 8544.4, 8664.5\,\text{Å}$), the strongest absorption feature in this region, is visible in stellar spectra of all but the hottest spectral types. The Ca II triplet EW is considered a good estimator of luminosity class in high metallicity objects and a useful metallicity indicator in metal-poor systems (e.g. Diaz et al. 1989). The relatively narrow intrinsic widths of the lines ($\sigma \approx 50\mathrm{km\,s^{-1}}$) and the increased velocity resolution per Å, twice that at blue optical wavelengths, makes the Ca II triplet ideal for studying the internal kinematics of galaxies (Dressler 1984; Terlevich et al. 1990). In spectra of intermediate S/N, the principal limitation to the use of the Ca II triplet at low redshift is the presence of the large number of strong OH emission lines in the same spectral region.

The SDSS spectra allow the measurement of the the Ca II triplet in galaxies out to redshifts $z \lesssim 0.06$ and the lines are ideally placed to investigate any potential bias in line feature properties caused by the sky-residual subtraction procedure. Furthermore, performing our own line search on low S/N SDSS spectra allows an immediate demonstration of the potential benefits afforded by the sky-residual subtraction technique. A large sample of galaxies with significant Ca II triplet absorption is readily identified and measured velocity dispersions are available from the SDSS spectroscopic data reduction pipeline for a substantial fraction of the galaxies of interest. As a simple integration over pixels is used to estimate EWs, the velocity dispersions are required in order to define the wavelength range over which the integration is performed.

*Equivalent width line ratios*

Galaxy velocity dispersion ($\sigma_v$) is only measured in the standard SDSS data reduction pipeline for objects with ECLASS $< -0.02$, $z < 0.4$ and SPECCLASS = “GALAXY” [4]. A sample of 7760 galaxies is selected according to the criteria:

1. $z < 0.054$, to provide enough spectrum for continuum estimation

2. $70 < \sigma_v < 420\,\mathrm{km\,s^{-1}}$

3. $\mathrm{S/N} > 10$ in the $r$-band

4. Ca II $\lambda\lambda 8544.4, 8664.5$ absorption lines detected in the SDSS catalogue at $> 4\sigma$ significance.

The second and third criteria are as recommended on the SDSS website. Those galaxies which are untouched by the sky subtraction procedure (their non-sky rms is equal to or greater than their sky rms) are removed from this sample. As high S/N objects have been selected through the requirement of significant Ca II triplet line detection, this removes a further $\sim 1000$ galaxies, leaving a sample of around 5500[5].

[4] see http://cas.sdss.org/astro/en/help/docs/algorithm.asp

[5] This number is slightly dependent on the width of the mask used for the Ca II triplet lines as the non-sky rms changes as a greater or lesser proportion of the wings of the lines are masked.

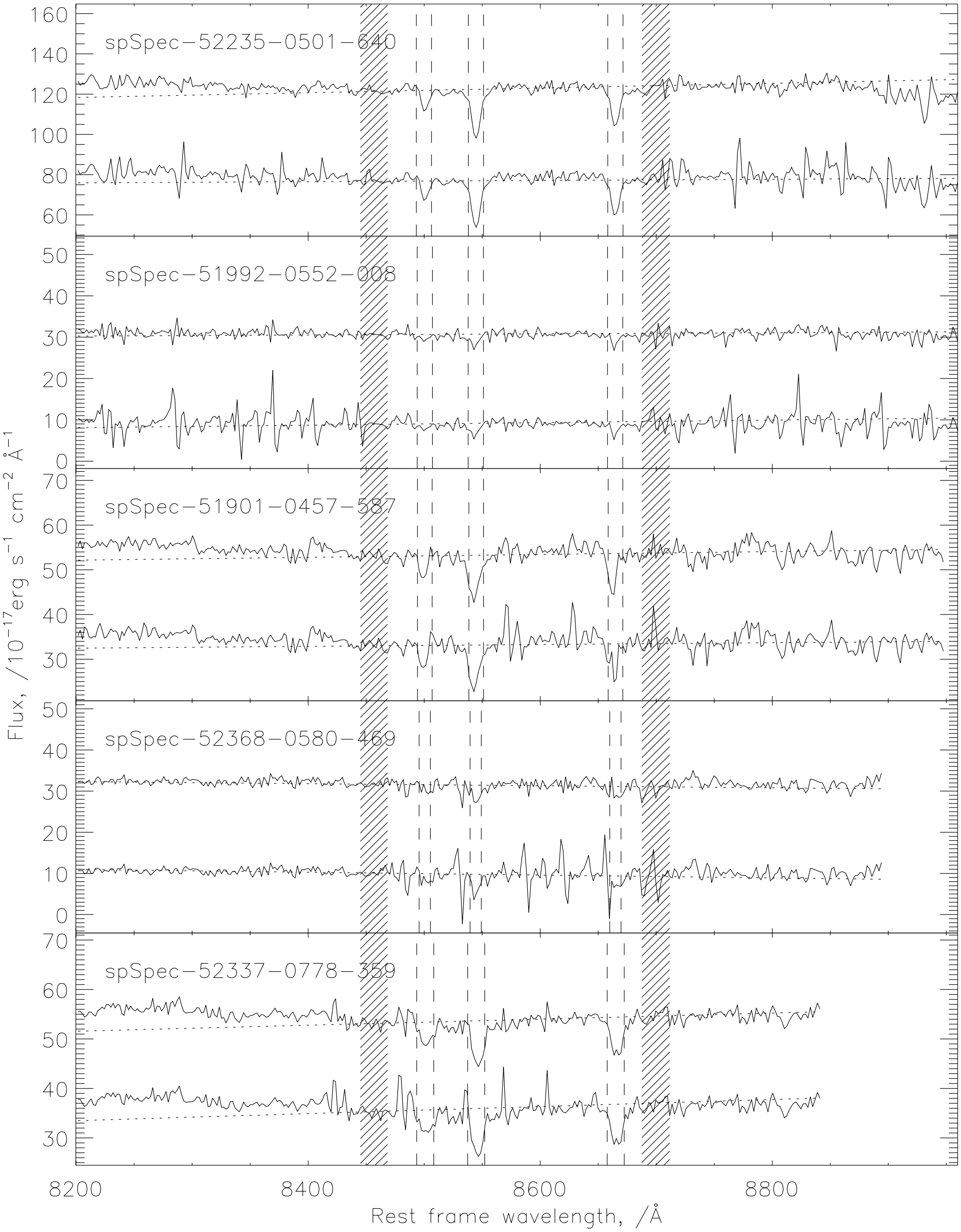

**Figure 4.16:** Example SDSS spectra in the region of the Ca II triplet, before (lower) and after (upper) sky-residual subtraction. Vertical lines indicate where equivalent widths of lines are measured, and shaded regions where the continuum is estimated. The inferred continuum is overplotted (dashed line). In each case the upper corrected spectrum is offset for clarity.

The centre of each line is given by the line wavelength in the SDSS catalogue; the feature mask width and the region over which the line EW is measured are set as $\pm m\sigma_v$ and $\pm n\sigma_v$ respectively. A value of $n = 2$ is used which generally spans the widths of the lines well, $m$ is left to take different values to investigate any bias caused by the sky-residual subtraction procedure. The continuum level at each wavelength in the region of the absorption lines is determined using a linear interpolation between the median flux contained in two bands, (8444.3:8469.3 Å and 8687.4:8712.4 Å). Fig. 4.16 shows examples of SDSS galaxy spectra in the Ca II triplet region before and after sky-residual subtraction. The regions defining the absorption lines and continuum bands are indicated.

The rest-frame EW is calculated by a sum over pixels in the wavelength range defined by the galaxies velocity dispersion and known central line positions:

$$\mathrm{EW} = \sum_{i=1}^{N} \Delta(W_i)(1 - \frac{f_i}{c_i}) \tag{4.3}$$

where N is the total number of pixels, $\Delta(W_i)$ the width of pixel $i$ in Å (rest-frame), $f_i$ the observed flux, and $c_i$ the continuum.

The total EW is defined as the sum of the two strongest absorption lines at $8545$Å and $8665$ Å. Inclusion of the other, weaker, line often increases the noise (e.g. Diaz et al. 1989). While the total EW varies with galaxy type, the ratio of the line EWs remains constant. The mean total EW, and mean and variance of the ratio of first to second EWs for the galaxy sample, are practically identical before and after sky residual subtraction, and for mask widths of $m = 1$, 2 and 3. The constancy of the values indicates that, provided the feature mask is broad enough to include the wings of the absorption lines, there is no significant difference between the tests using sky-subtracted and non-sky-subtracted spectra.

On average, the sky-residual subtraction produces improvements in the S/N of spectra within the regions of the masked absorption features; the feature mask excludes the absorption line wavelengths from contributing to the PCA reconstruction of the sky residual but the sky-residual subtraction is applied to all wavelengths. The key result of the test using the properties of the Ca II triplet is that, providing the features are masked prior to the PCA reconstruction of the sky-residual signal, the properties of the features are not systematically biased.

### *Improvement in feature detection quality*

A practical illustration of the improvement afforded by the sky-residual subtraction procedure is seen in the results of a conventional matched-filter search (Hewett et al. 1985) for the presence of the Ca II triplet in relatively low S/N spectra, treating the original and sky-residual subtracted spectra identically. The DR3 release, used in this case simply to increase the sample size, contains $\sim 15000$ galaxy spectra with S/N in the $r$-band of $< 15$.

Adopting a nominal $3\sigma$ threshold for detection results in 5702 and 5662 detections for the original and sky-residual spectra respectively, 5119 of which are common to both. Close to the $3\sigma$ S/N threshold, twice as many spurious detections of the Ca II triplet ”in emission” exist in the original sample, providing evidence that the $\sim 10\%$ of lines unique to each sample are statistically more reliable in the sky-subtracted data set. However, the most significant difference apparent in the properties of the detected features is in the quality of the fit between the matched-filter template and the data. For each

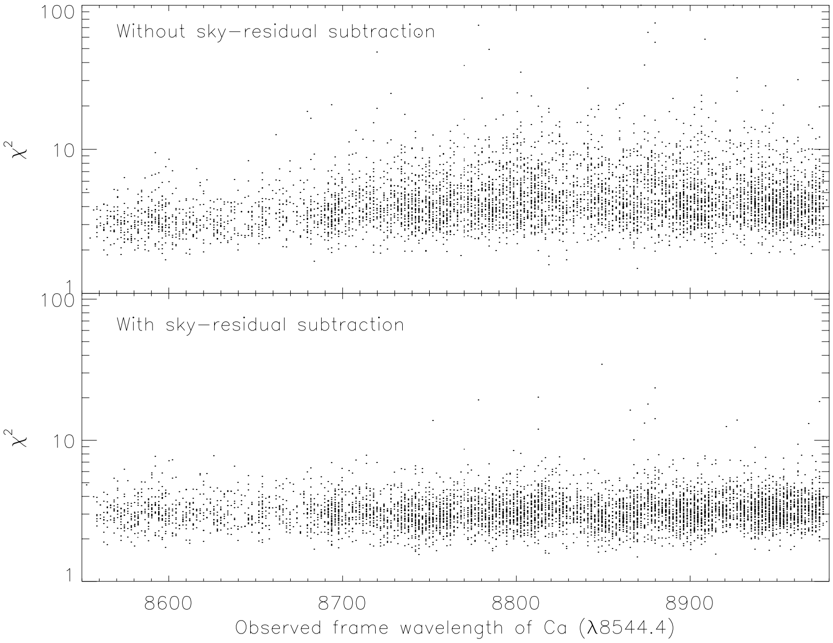

**Figure 4.17:** Resulting reduced-$\chi^2$ between a scaled matched-filter and $\sim$ 6000 spectra with detected Ca II triplet, as a function of the observed wavelength of the 8544.4 Å line. Upper (lower) panel is for spectra without (with) sky-residual subtraction. All galaxies in SDSS DR3 with S/N in the $r$-band of $< 15$ and redshifts $z < 0.05$ are searched for the Ca II triplet at the known redshift of the galaxy. The sky-residual subtracted spectra are treated identically to the original dataset. The samples plotted result from adopting a $3\sigma$ detection threshold for the presence of the Ca II triplet.

Ca II triplet detection a goodness-of-fit is calculated based on the sum of the squared deviations between the data and the scaled template. In Fig. 4.17 this is plotted versus the observed frame wavelength of the 8544 Å line. The overall scatter in the $\chi^2$ distribution at wavelengths where OH lines are present (>8650 Å) is significantly reduced in the case of the sky-residual subtracted spectra (the median $\chi^2$ for detections with wavelength >8650 Å decreases by 27%). The systematic trend as a function of wavelength is also essentially removed. A significant fraction of the original spectra present extremely poor fits where the observed frame wavelength of the triplet coincides with a particularly strong OH line. At wavelengths >8650 Å, the fraction of detections with $\chi^2$ values exceeding twice the median value decreases from 8% to just 1% in the case of the sky-residual subtracted galaxies.

### 4.5.2 Absorption features in damped Lyman-$\alpha$ systems

The detection of intervening absorption features in quasar spectra is an example of an investigation in which it is not possible to simply mask the wavelength regions containing features before the application of the sky residual subtraction. That is not to say that the sky-residual subtraction scheme is not of potential interest. The improvement in the quality of the spectra is such that the S/N of specific features can increase, albeit at the potential cost of modification of the feature properties. In practice, a hybrid scheme, involving one or more iterations, with detected features masked in a second application of the sky-residual subtraction, is straightforward to implement (see proceeding chapters).

The impact of the sky-residual subtraction on unmasked absorption features can be illustrated through

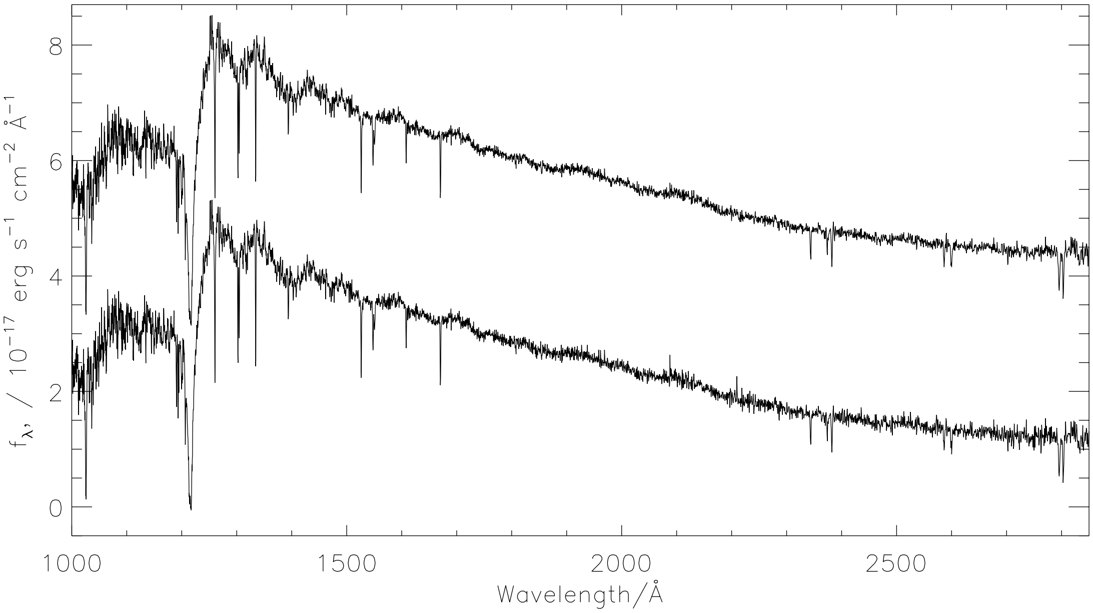

**Figure 4.18:** Composite spectrum of 68 quasars containing DLAs identified in the SDSS-DR1 by Prochaska & Herbert-Fort (2004) combined in the rest-frame of the DLA: before (lower) and after (upper) sky-residual subtraction. The metal absorption features were masked prior to sky-residual subtraction. No attempt was made to remove the underlying quasar spectra.

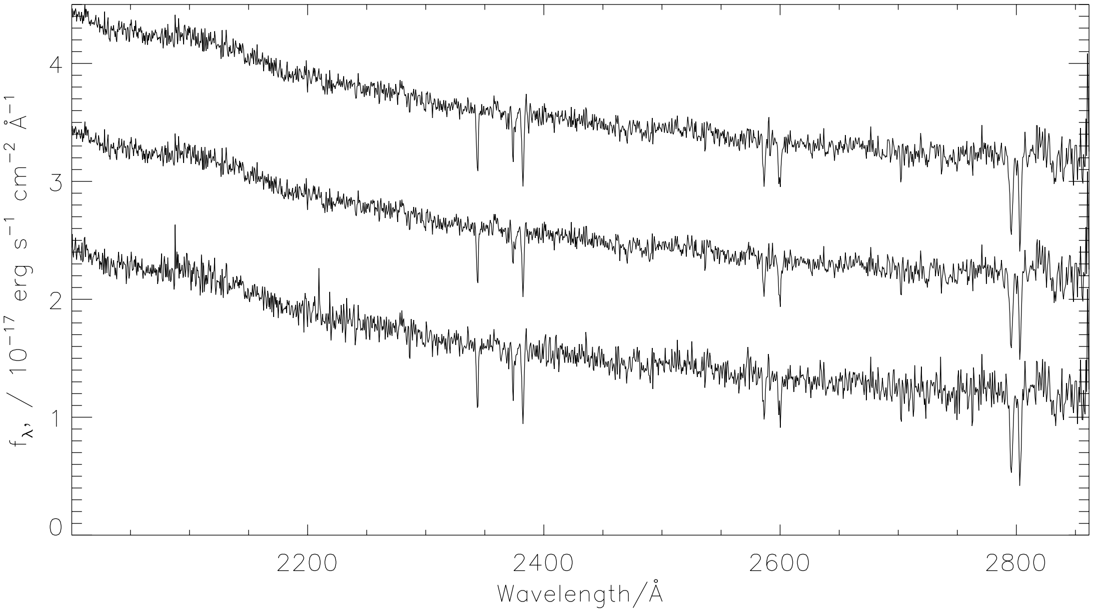

**Figure 4.19:** Composites of 63 quasar spectra containing DLAs combined in the absorber rest-frame over the wavelength range $2000 - 2850$ Å that includes strong absorption features due to Mg and Fe. The three spectra show the composite before sky-residual subtraction (lower), after sky-residual subtraction with absorption features unmasked (middle), and with absorption features masked (upper).

**Table 4.2:** Equivalent widths of metal absorption features in a composite of 68 SDSS-DR1 quasar spectra with DLAs combined in the absorber rest-frame, before and after sky residual subtraction: (a) absorption features unmasked; (b) $14$ Å around absorption features masked. Wavelengths of continuum and measurement regions are given, together with the number of spectra contributing to the composite (where the wavelength of the line falls between 6700 and $9180$ Å).

| Line ID | N | cont1 | cont2 | line | before | $\mathrm{after}^{\mathrm{a}}$ | $\mathrm{after}^{\mathrm{b}}$ |
|---|---|---|---|---|---|---|---|
| FeII[2344.9] | 47 | 2327:2337 | 2353:2363 | 2340.9:2347.9 | -0.65 | -0.58 | -0.66 |
| FeII[2375.2] | 46 | 2357:2367 | 2392:2402 | 2371.2:2378.7 | -0.50 | -0.44 | -0.54 |
| FeII[2383.5] | 45 | 2357:2367 | 2392:2402 | 2379.0:2386.5 | -0.68 | -0.62 | -0.66 |
| MgII[2796.4] | 11 | 2760:2790 | 2810:2825 | 2792.0:2799.5 | -1.89 | -1.71 | -1.81 |
| MgII[2803.5] | 10 | 2760:2790 | 2810:2825 | 2799.5:2807.0 | -1.75 | -1.55 | -1.73 |

creating composite spectra of damped Lyman-$\alpha$ (DLA) systems identified by Prochaska & Herbert-Fort (2004). Each quasar spectrum is shifted to the rest-frame of the DLA absorber and normalised to contain the same signal in the absorber rest-frame wavelength interval $1250-1800$Å. The composite spectrum is then constructed using an arithmetic mean of all the spectra with signal at each wavelength, taking care to account for the flux/Å term in the SDSS spectra. Fig. 4.18 shows the resulting composite spectra, calculated with and without the application of the sky-residual subtraction to the constituent quasar spectra. As in Section 4.4.2 the centres of prominent broad emission lines in the quasars are masked during the sky-residual subtraction.

Fig. 4.19 shows three versions of the composite spectrum, focusing on the rest-frame wavelength region that derives from observed-frame wavelengths $\lambda > 6700$Å. The three versions of the composite were calculated using quasar spectra without sky-residual subtraction (bottom), with sky-residual subtraction (middle) and with sky-residual subtraction *and* the absorption features masked. The absorption-line mask consisted of 14Å intervals, in the absorber rest-frame, centered on each absorption line.

Absorption-line EWs are measured in the absorber rest-frame, using the method of Section 4.5.1, integrating over a fixed wavelength range (between 7 and 8Å for the isolated lines). Table 4.2 includes the EWs of the absorption features in the three composite spectra, together with the wavelength intervals used in the line and continuum measurement. For each composite the same continuum is used for measuring EWs, in this case calculated from the spectrum with lines masked, although in practice any consistent continuum would suffice. The sky-residual subtracted composite, with masking of absorption lines, provides the best reference. As expected, the EWs of several of the absorption lines in the unmasked composite are systematically reduced. However, the illustration is a “worst case” example of the sky-residual subtraction scheme involving only a single iteration and in any case the effect is small. A second iteration, in which the identified features are then masked during the sky-subtraction allows an accurate measurement of the features. This indicates that an iterative procedure of applying the sky-residual subtraction twice offers the prospect of significant improvements in both the identification and the measurement of features at initially unknown wavelengths.

### 4.5.3 High redshift quasar composites

Poor sky subtraction affects mostly those spectra with low S/N, and the fainter quasars, including those at high redshift ($z > 4$), suffer from significant contamination from sky subtraction residuals. The extended wavelength range of the SDSS spectra combined with a redshift range for the quasars of $z \simeq$

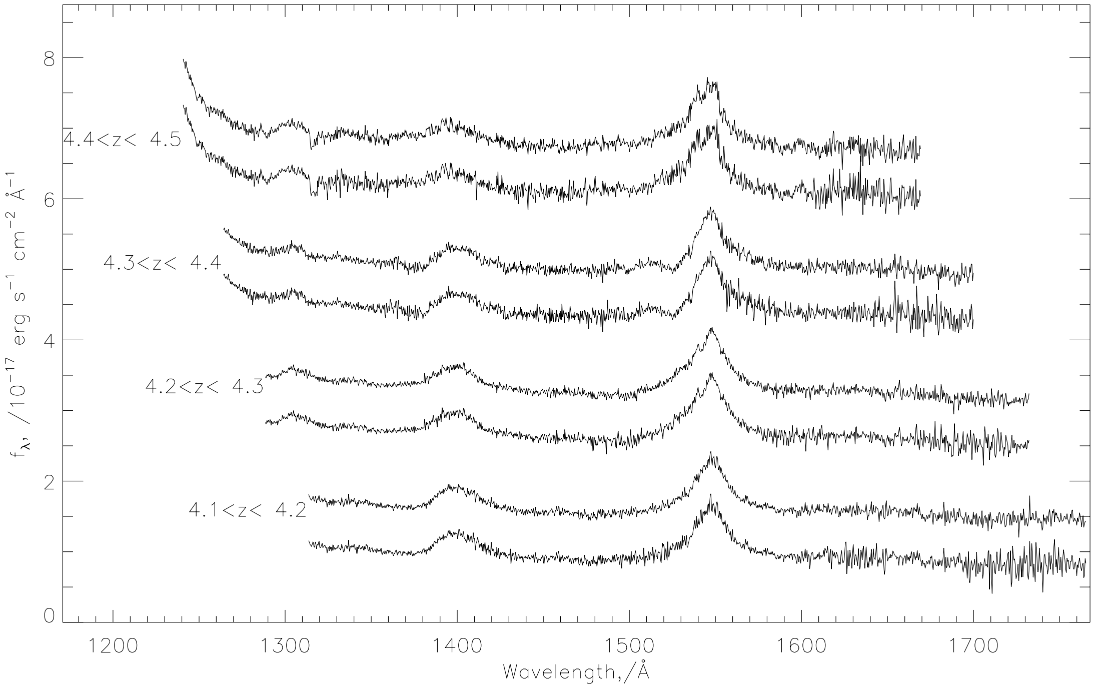

**Figure 4.20:** Composite spectra of $z \sim 4$ quasars including the SiIV+OIV $\lambda 1400$ and CIV $\lambda 1549$ emission lines. The observed-frame spectra are cut such that only pixels with wavelengths $\lambda > 6700$ Å are retained. For each pair of spectra, the lower composite is created directly from the SDSS spectra, and the upper composite from the same spectra after subtracting sky residuals. From bottom to top the composites contain 41, 35, 24 and 20 spectra and the improvement in error weighted absolute deviation from the underlying spectrum (see text) is 23%, 17%, 17% and 18%.

$0-5$, allows the construction of composite spectra over unprecedented ranges in rest-frame wavelength and luminosity (vanden Berk et al. 2004) so providing the basis upon which quantitative studies of the evolution of AGN can be performed. However, the number of quasars that occupy the extremes of the wavelength and luminosity ranges is relatively small and the sky subtraction residuals at $\lambda > 6700$ Å limit both the determination of the continuum properties and the detection of weak emission features. Composite rest-frame spectra are created of all quasars in the SDSS with $z > 4.1$, in redshift slices of $\Delta z = 0.1$. Each input quasar spectrum is normalised to have a mean flux of unity between $1250-1600$ Å, prior to combination using an arithmetic mean. Fig. 4.20 illustrates how the sky-residual subtraction technique improves the quality of these composites. Subtracting the underlying spectrum using a median filter and removing the centers of emission lines from the calculation, the error weighted absolute deviation from zero is calculated for the composites with and without sky-residual subtraction. The average improvement in S/N over the entire wavelength region due to the sky subtraction is about 20% for all redshift bins.

## 4.6 Summary

The SDSS spectra represent a vast improvement in data quantity, quality and wavelength coverage over previous surveys. The sky-subtraction carried out by the SDSS reduction pipeline has in general achieved excellent results, however the common problem of the subtraction of undersampled OH emission lines results in substantial systematic residuals over almost half the spectral range. In this chapter a technique is presented to remove these residual OH features, based on a PCA of the observed-frame sky spectra which have been sky subtracted along with all other spectra in the SDSS data releases. The scheme takes advantage of the high degree of correlation between the residuals in order to produce corrected spectra whose noise properties are very close to the limit set by counting statistics.

In this chapter the sky-residual subtraction scheme has been applied to the SDSS DR2 catalogue that was available at the time the work was carried out; it is immediately applicable to the more recent SDSS data releases. The precise form of the implementation depends on the spectral energy distributions of the target objects and the scientific goal of any analysis. As a consequence, the main classes of target objects, galaxies, quasars and stars, are treated slightly differently. Constructing composites of high-redshift quasars before and after sky-residual subtraction illustrates the significant increase in the overall S/N of the spectra that is achieved. The application of the procedure when narrow absorption or emission features are present requires the use of a mask to prevent the sky-subtraction scheme suffering from bias. Both the calcium-triplet in low redshift galaxies and metal absorption lines in damped Lyman-$\alpha$ systems at high redshifts fall in the wavelength range of the SDSS spectra affected by the OH-forest $(6700 - 9200\,\text{Å})$. By employing these two rather different examples it is shown that masking absorption and emission features prior to the reconstruction of the sky-residual signal in each spectrum ensures that the properties of the features themselves are not systematically biased in the resulting sky-residual subtracted spectra.

It is hoped that the significant improvement in S/N achieved over some $2000\text{Å}$ of a large fraction of SDSS spectra, particularly for the fainter objects such as the high-redshift quasars, should benefit a wide range of scientific investigations - well beyond those presented in this Thesis. As part of this project we have made freely available on the web datafiles, IDL code and a comprehensive guide on using the code, with which the procedure can be applied to all SDSS spectra[6]. Further improvements should be possible if the sky-residual subtraction procedure were to be adapted to run on the SDSS spectra using the original noise arrays based on counting statistics alone. However, as the SDSS pipeline currently stands it seems unlikely that this will be easily achievable. The nature of the sky-residual subtraction scheme is such that application to other large samples of spectra should be relatively straightforward, although the code created during this project is applicable solely to the SDSS datasets.

[6] http://www.ast.cam.ac.uk/research/downloads/code/vw/

*Milk's a really good source of Calcium*

Nicole

# 5

# Dust in quasar absorption line systems

*The SDSS catalogues contain a wealth of information on the physics of the galaxy population. Of particular relevance to this thesis is the vast number of high quality spectra, in contrast to the 2dF catalogues. While the SDSS galaxy catalogue provides direct observations of hundreds of thousands of galaxies in the local Universe, the quasar catalogue allows the indirect detection of galaxies out to very high redshifts through the absorption of the quasar light by metals in their ISM. This chapter investigates the reddening effect of the dust in these systems, presenting a new sample of absorption line systems at $z \sim 1$ with strong Ca* II $\lambda\lambda 3935, 3970$ *lines which are found to contain significant quantities of dust.*

## 5.1 Introduction

The sheer statistical power of the SDSS DR4 allows both the identification of significant numbers of rare objects and good estimates of the average properties of common objects. In this chapter, 37 rare Ca II absorption line systems are found in the spectroscopic quasar catalogue within a narrow redshift range, $0.84 < z_{\rm abs} < 1.3$. To measure any difference in colour between those quasar spectra containing Ca II absorption and those without, control templates are created by combining large numbers of quasar spectra. In the following subsections a brief summary is given of the topic of quasar absorption line systems, damped Lyman-$\alpha$ systems (DLAs) and previous attempts to measure their dust content.

### 5.1.1 Quasar absorption line and Damped Lyman-$\alpha$ systems

Quasar absorption line systems are galaxies intervening a quasar line of sight, the metals, hydrogen and helium in their ISM causing absorption of the quasar light at particular wavelengths (see Fig. 1.4). DLAs are a subset of these absorbers, defined through their high column densities of neutral hydrogen gas, $N(\mathrm{H\,I}) > 2\times10^{20}\mathrm{cm}^{-2}$ or $\log[N(\mathrm{H\,I})] > 20.3$. Other common classes of quasar absorption line systems are Mg II and C IV absorbers and so called Lyman Limit systems defined to have $10^{17} < N(\mathrm{H\,I}) < 2\times10^{20}$.

Each of these classes of absorbers has an observed number density per unit redshift, $n(z)$, which measures their typical cross-section on the sky. A simple interpretation for the different cross-sections of different classes of absorbers is that they refer to the gas structure in the interstellar medium of a typical galaxy. By assuming these "typical galaxies" have a constant space density, that each class of absorber is associated with the same galaxies and that the ISM of the galaxies varies smoothly with radius, a typical size for the absorbing region can be obtained. DLAs, Mg II absorbers, Lyman Limit systems and C IV absorbers would be found in sightlines passing within 15, 40, 40 and $70h^{-1}$ kpc of the galaxy center respectively (e.g Steidel 1993).

Although DLAs have been studied extensively for twenty years (Wolfe et al. 1986), their precise nature, evolution and relation to the population of galaxies as a whole continue to be the subject of much discussion (Wolfe et al. 2005). Despite the lower number densities of DLAs compared to other classes of absorbers, DLAs are arguably the most important class for understanding galaxy evolution at high redshifts. They are the only class of absorber known to contain predominantly neutral gas and they dominate the neutral gas content of the Universe between $0 < z < 5$. It is possible that through DLAs, we are observing the precursors to the regions in galaxies in which most stars are formed.

### 5.1.2 Finding intermediate redshift DLAs

Large samples of DLAs are now known at high redshift, primarily through the SDSS quasar survey, however the defining diagnostic of a DLA, the Lyman-$\alpha$ absorption line does not enter the optical wavelength range until $z_{\rm abs} \sim 1.8$. At low redshifts DLAs can again be observed through 21 cm emission (e.g. Zwaan et al. 2005). The need for space based ultra–violet (UV) observations has made it difficult to obtain large samples of intermediate redshift DLAs, crucial both to clarify the relation of DLAs to luminous galaxies by direct imaging of the DLA hosts (which is much easier at redshifts $z \lesssim 1$), and to follow the evolution of their properties over the course of time.

Rao & Turnshek (2000) proposed an important prognostic for intermediate redshift DLAs which circumvents to some extent the need for UV spectroscopy: the equivalent widths of the strong Mg II $\lambda2796$ and Fe II $\lambda2600$ absorption lines can be used to identify DLAs with a ∼40% success rate (Rao 2005). The SDSS quasar catalogue provides the ideal database for obtaining large samples of such metal absorption line systems over a large redshift range. A complementary method to that of Rao & Turnshek for identifying DLAs at intermediate redshift is through the detection of intrinsically weak absorption lines, potentially allowing the identification of DLAs with ∼100% certainty. The primary advantage of this latter method is that UV spectroscopy would not be required to quantify the fraction of objects which are DLAs, and metal depletion patterns could be studied without the added difficulty of unknown ionisation states. With the SDSS, identification of statistically significant samples of rare absorption systems has been made possible, such as the sample of Ca II absorbers presented in this chapter.

### 5.1.3 Dust in DLAs and obscuration bias

One puzzling aspect of DLA properties is their apparent small dust content. At high redshift this may be explained by their lower metallicities, however abundances and element dust depletion patterns appear to show little or no evolution in the population down to $z \sim 1$. Quantifying the amount of dust in DLAs is important not only for understanding the chemical evolution of galaxies over large lookback times, but also has implications for the biases involved in DLA selection in optical magnitude-limited quasar surveys. It is still possible that we are missing a population of dusty, metal rich DLAs because of the obscuration they would cause to the background quasar light (e.g. Ostriker & Heisler 1984; Fall & Pei 1989, 1993).

Dust has three observable effects in the ISM of galaxies: it selectively depletes metals, reddens their spectral energy distributions, and causes an overall extinction of the light. Due to the difficulty in defining suitable samples, there are relatively few reports of evidence for dust in DLAs via the reddening effect on background quasar spectra. Pei, Fall & Bechtold (1991), following up from the study by Fall, Pei & McMahon (1989), found the spectral energy distributions (SEDs) of 20 quasars with DLAs to be significantly redder than those of 46 quasars without DLAs. Recently however, Murphy & Liske (2004) found no evidence for dust reddening in a much larger, homogeneous sample of DLAs from the SDSS, finding a limit, $E(B-V)_{\rm SMC} < 0.02$, that is inconsistent with the results of Pei et al.. Whilst the optical selection of the SDSS sample could potentially introduce some bias into this result, Ellison, Hall & Lira (2005) find an $E(B-V)_{\rm SMC} < 0.04$ ($3\sigma$) using a smaller sample of radio selected quasars with DLAs.

Estimates of reddening from differences in the average SEDs of quasar samples must rely on statistical studies, due to the intrinsic range in the SEDs of quasar spectra. A clear indication of the presence of dust can come, however, from the identification of spectroscopic features caused by the dust grains. The strongest such feature in the Milky Way is at ∼2175Å, but the feature is weak or absent in the Large and Small Magellanic Clouds (LMC and SMC). The 2175Å feature has been detected in the spectrum of a BL Lac object at the absorption redshift of a known intervening DLA by Junkkarinen et al. (2004); Wang et al. (2004) have presented possible detections of the feature in SDSS quasar spectra with strong intervening metal absorption systems.

A further diagnostic of the presence of dust in DLAs and metal absorption line systems comes from relative abundances of elements, such as Cr to Zn, which are depleted by differing amounts onto dust grains (Savage & Sembach 1996). Results indicate that dust depletion is far less severe than in the Galactic interstellar medium today (Pettini et al. 1994, 1997) which, combined with results from the radio-selected CORALS survey (Akerman et al. 2005), has further strengthened the argument that DLAs are relatively dust free compared with modern galaxies. The effect of dust on element abundances is discussed more extensively in Chapter 6.

The importance of dust-induced selection effects in quasar samples when searching for DLAs or metal absorption line systems is clear, a small amount of dust present in an intervening galaxy could cause enough extinction for the quasar to fall below the detection limit of optical magnitude-limited surveys. The effect is known as “dust obscuration bias”. By selecting quasars at radio wavelengths, which are uneffected by dust, the CORALs survey detected no significant obscuration bias at $z > 1.8$

(Ellison et al. 2001), although with only 22 DLAs in 66 quasars small number statistics still allow ∼50% of systems to be undetected in optical surveys. By making use of the criteria for identifying potential DLAs from the strengths of metal absorption lines (Rao & Turnshek 2000), Ellison et al. (2004) extend this result to absorbers at lower redshift. Their results permit up to a factor of 2.5 underestimate of the incidence of DLAs in optical–, as opposed to radio–selected quasar samples. In another study, Vladilo & Péroux (2005) used the observed distribution of column densities of Zn II to argue that between 30 and 50% of systems would be obscured, although this is not borne out by the study of Akerman et al. (2005). It is certain that dust obscuration bias will be important at some level, however the statistics are not currently good enough to assess its overall impact.

### 5.1.4 The aim of this chapter

In this chapter a sample of Ca II absorption line systems are studied and an analysis of the reddening of the background quasar SEDs caused by dust in the absorbers is carried out. The results are compared to a sample of absorbers which fulfil the criteria of Rao & Turnshek (2000) for intermediate redshift DLAs. Before the analysis can be started a subsample of quasars must be removed from the quasar catalogue: Broad Absorption Line quasars (BALs) and quasars with associated absorbers have both been identified as having significantly redder SEDs than average. By removing these objects from both the absorber and control samples, any reddening due to intervening absorbers will be more easily detected.

## 5.2 Identifying Broad Absorption Line Quasars

BAL quasars are known to make up around 15% of all active galactic nuclei at $z_{\rm em} > 1$ (Hewett & Foltz 2003; Reichard et al. 2003b). They are characterised by broad absorption troughs blueward of emission lines of many ionisation states—presumably arising in outflowing material—and an overall reddening of the quasar spectral energy distribution. While much effort is being devoted to understanding the origin of these spectral features and the position of BAL quasars in the unified model of active galactic nuclei, their generally redder SEDs are a potential source of additional uncertainty in studies of reddening caused by intervening absorbers. Associated absorption systems, in which hydrogen and metal absorption lines seen in the spectra of quasars are associated with galaxies near to the quasar host galaxy, also appear in quasars with generally redder SEDs (Carilli et al. 1998). Thus, quasars whose spectra show evidence of BALs, or strong associated absorption, need to be removed from the sample prior to the reddening analysis. Due to the large number of quasar spectra in the SDSS, the problem becomes one of developing suitable automatic pattern recognition techniques capable of coping with the wide range in strength and shape of the absorption features.

The main difficulty lies in defining the quasar continuum around the emission lines. Traditionally, this has been achieved by fitting power laws to the spectra away from regions containing emission lines. Reichard et al. (2003a) compared this method to the use of composite spectra in the SDSS early data release; they also provide a history of BAL detection and an extensive discussion of the different methods. In this chapter a third method is presented which makes use of a principal component analysis (PCA, Section 2.3) of the spectra to reconstruct the quasar continua and emission lines, without the broad absorption features. The main advantage of PCA over the use of composite spectra for reconstructing the quasar continua is the tailoring of the fit to each individual object; all the detailed information which is lost through power law fits can also be retained.

### 5.2.1 Details of the method

Given the size of the SDSS catalogue, it is possible to create high SNR eigenspectra for different sub-samples of quasars. Splitting the sample into redshift bins can greatly simplify the PCA analysis by minimising the 'missing data' caused by differing rest wavelength coverage. Three redshift bins are defined such that all the spectra contained within them have less than 25% of their pixels missing when placed onto a common rest frame wavelength array: $0.823 < z_{\rm em} < 1.537$, $1.405 < z_{\rm em} < 2.179$ and $2.172 < z_{\rm em} < 3.193$. A slight overlap is allowed between the redshift bins in order to increase the SNR at either end of the eigenspectra; a small fraction of the spectra therefore contribute to two bins. Throughout this chapter wavelengths shortwards of 1250Å are discarded i.e. the Lyman-$\alpha$ forest and the peak of the Lyman-$\alpha$ emission in the quasars.

Several steps are involved before the final sample of BAL quasars was produced. The first step was to identify strong BAL quasar candidates using a simple "colour cut", based on the difference in the flux contained within bandpasses on either side of the C IV $\lambda1550$ emission line (or the Mg II $\lambda2796$ line when C IV is not covered); these obvious BAL quasars were then temporarily removed from the sample[1]. The remaining spectra were shifted to the rest-frame (without rebinning) and a PCA performed on the spectra in each redshift bin using the "gappy-PCA" method introduced in Section 2.3.2. This was necessary primarily because of the different intrinsic wavelength coverage of the input spectra in any one redshift bin. The large number of pixels covering the range in rest-frame wavelength within each redshift bin necessitated the use of the expectation maximisation algorithm for PCA given in Section 2.3.2.

A 41-pixel median filter was run through each quasar spectrum to identify narrow absorption features. The spectra were then reconstructed from the eigenspectra with these narrow features masked to prevent them from biasing the reconstructed continua low[2]. Reconstructed continua were created for the entire sample of quasars, including those excluded during the creation of the eigenspectra by the colour cuts around the C IV $\lambda1550$ and Mg II $\lambda2796$ lines. For those interested in quantifying the "BALnicity" of individual objects (e.g. Weymann et al. 1991), the quasar spectra would be best reconstructed using only those eigenspectra visually identified as containing no BAL features, with the regions in which BAL features are expected masked during reconstruction so as not to bias the continua too low. Fig. 5.1 shows examples of such reconstructions for both BAL and non-BAL quasar spectra.

For the purposes of identifying and removing unusual spectra, as required for the analysis of this chapter, the precise number of eigenspectra used during reconstruction is not critical. BAL spectra and quasar spectra with strong associated absorption possess large $\chi^2$ values between the true and reconstructed spectrum in the regions either side of quasar emission lines even when as many as 10-15 eigenspectra are used for reconstruction. This reflects the large variations in size and shape of BAL features which the PCA is unable to reconstruct even with relatively large numbers of components. As precise classification of each object was not of interest, the distributions of $\chi^2$ between the true and reconstructed spectra in wavelength regions blueward of the C IV $\lambda1550$, Mg II $\lambda2796$ and/or Fe II $\lambda2600$ lines were

[1] The colour cuts were applied by Paul Hewett.

[2] Masks were applied during the gappy-PCA procedure by giving the pixels zero weight.

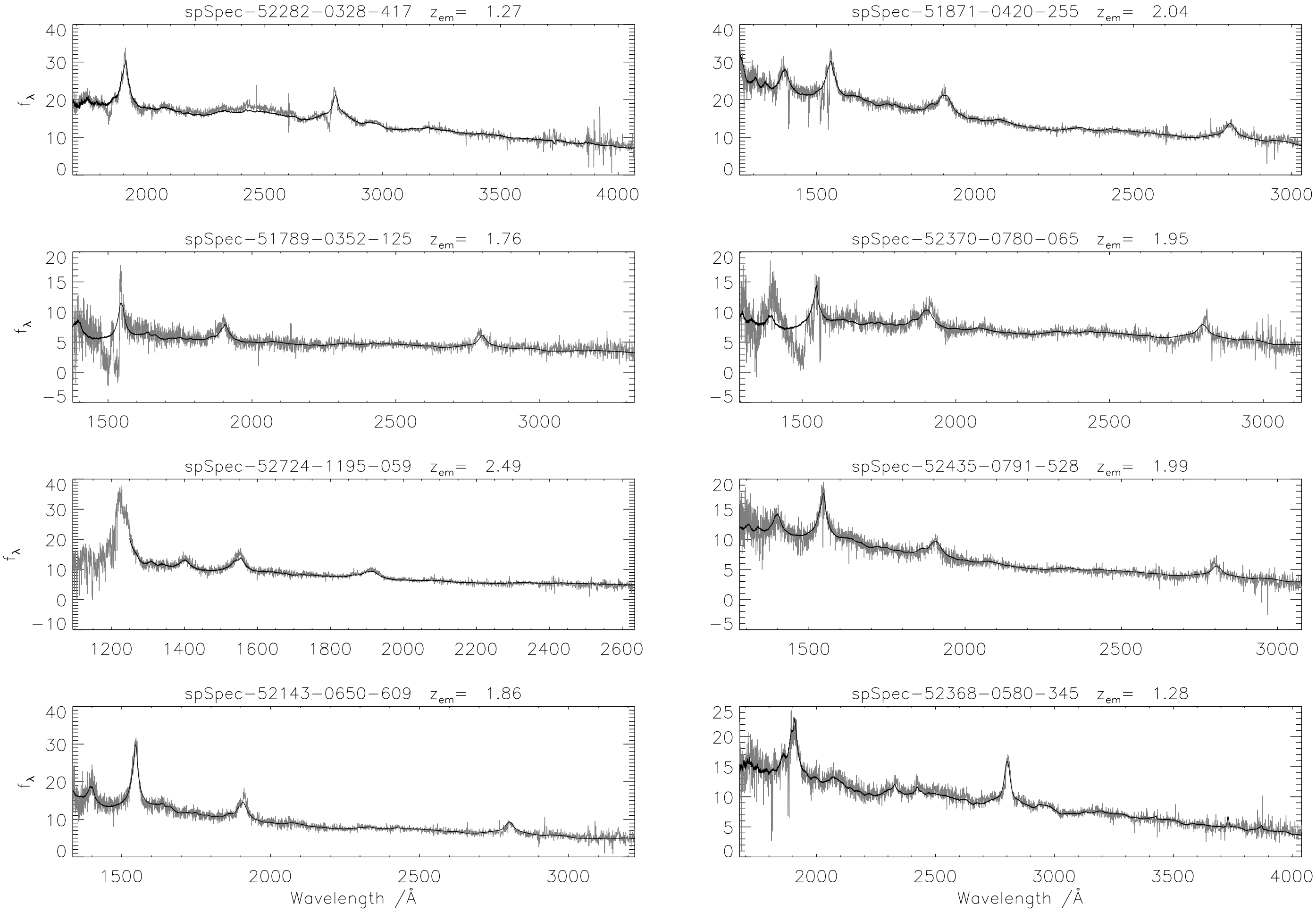

**Figure 5.1:** Examples of quasars with (upper four panels) and without (lower four panels) broad absorption troughs and/or intrinsic absorption lines; the PCA reconstructions of the spectra are overplotted in black. Note the failure of the reconstructions to follow the majority of the BAL features due to the iterative removal of these objects from the sample used to create the eigenspectra. Unusual quasars not used in this study are identified by the large $\chi^2$ between the true and reconstructed spectra in wavelength regions around the C IV $\lambda1550$, Mg II $\lambda2796$ and Fe II $\lambda2600$ lines. The $y$-axis is in units of $10^{-17}$ erg cm$^{-2}$ s$^{-1}$ Å$^{-1}$. The SDSS spectroscopic filename (modified julian date, plate number, fibre number) and quasar redshift are given above each panel.

plotted. All objects which fell into the high $\chi^2$ tail of each distribution were removed, and a new input sample produced for a second iteration of the PCA. The entire PCA and reconstruction procedure was repeated twice to ensure that as few BALs as possible were present when creating the final eigenspectra and the eigenspectra were therefore most representative of ordinary quasars. In total, 13% of quasars, with redshifts $0.85 \leq z \leq 3.2$, were removed from the input sample.

Ultimately, the aim was to create a subsample of quasars suitable for: (a) searching for intervening absorption systems and (b) creating control spectra with which to compare the SEDs of quasars with absorbers. The final results of this chapter are insensitive to the details of the scheme adopted to identify and exclude BAL quasars. Inadvertently removing non-BALs from the sample reduces the sightlines available for the search, but this is a small effect given the size of the input sample. On the other hand, any BAL quasars which remain in the sample will increase the variance among the quasar SEDs, and the ability to isolate any systematic differences between the SEDs of quasars with intervening metal absorption line systems and the quasar population as a whole will be reduced. This effect is included in the Monte Carlo simulations used to estimate the mean and variance of SED colours for random samples of quasars (see Section 5.5.2).

## 5.3 A sample of Ca II and Mg II absorption systems

The sample of quasar spectra used during the absorption line search and creation of control quasar composite spectra was restricted to those with spectroscopic SNR$>10$ in the $i$-band and with Galactic extinction-corrected point spread function (PSF) magnitudes $i < 19.0$ in order to minimise the number of false detections during the Ca II line searches. The magnitude cut is very similar to that used in constructing the main spectroscopic quasar sample from the SDSS photometric catalogue ($i \leq 19.1$). After exclusion of BAL quasars, the final sample consists of 11 371 quasars from the DR3 quasar catalogue (Schneider et al. 2005) and a further 3 153 from the quasars observed on the 226 additional spectroscopic plates in DR4.

The 14 524 quasars were searched independently for both Ca II $\lambda\lambda3935, 3970$ and Mg II $\lambda\lambda2796$,$2804$ absorption lines at redshifts $0.84 < z_{\rm abs} < 1.3$. Residual sky OH emission features were subtracted from all spectra using the method of Chapter 4, which greatly reduces residual sky noise in the wavelength region around the Ca II lines. The search[3] used a matched-filter technique (e.g. Hewett et al. 1985) with two template Gaussian doublets of the appropriate wavelength separation and full width at half maximum FWHM = 160 km s$^{-1}$ (the resolution of the SDSS spectra) and 240 km s$^{-1}$. Three values of the doublet ratio were incorporated into the search: 2:1 (corresponding to unsaturated lines on the linear part of the curve of growth), 1:1 (for saturated lines on the flat part of the curve of growth) and an intermediate case, 4:3.

The redshift range of the absorber sample, $0.84 < z_{\rm abs} < 1.3$, is set by the appearance of the 2175Å feature of the Milky Way reddening curve beyond 4000Å in the SDSS spectra, and by the Ca II doublet moving beyond the red limit of the spectra. The samples are further restricted by requiring the 2175Å feature to fall redward of 1250Å in the quasar rest-frame to avoid confusion with the Lyman $\alpha$ forest.

[3] The matched-filter search for metal absorption lines was carried out by Paul Hewett.

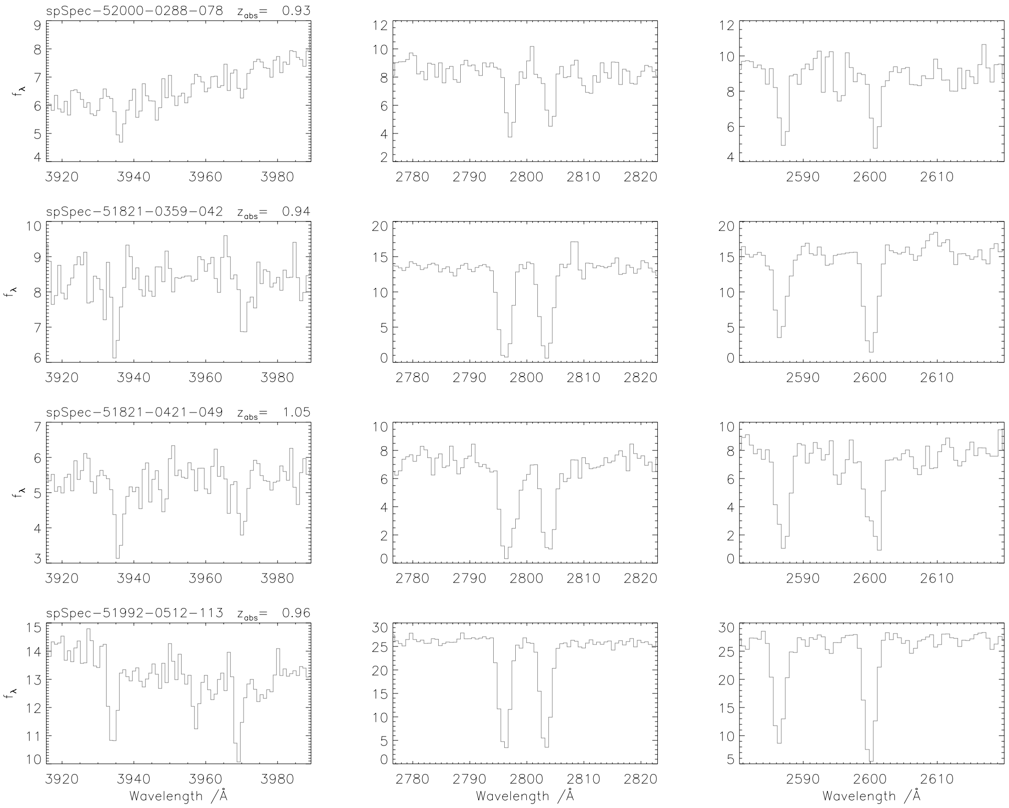

**Figure 5.2:** Examples of Ca II systems (left panels) found in the SDSS quasar sample, along with associated Mg II (centre) and Fe II lines (right). The units of the $y$-axis are $10^{-17}$ erg cm$^{-2}$ s$^{-1}$ Å$^{-1}$. The SDSS spectroscopic filename (MJD, plate number, fibre number) and absorber redshift are given above each panel in the left-hand column.

### 5.3.1 Ca II systems

At the redshifts of interest the Ca II lines fall in the red portion of the optical spectrum, beyond 7250Å, where the SDSS quasar spectra become progressively noisier due to the increasing sky background and decreasing instrumental sensitivity. Thus, candidate Ca II systems were required to possess corresponding Mg II absorption within ±200 km s$^{-1}$. This list of Ca II candidates was subjected to further scrutiny by a fully parameterised fit of the absorption lines. A continuum was fitted to the regions around the Mg II and Ca II lines using the IRAF routine CONTINUUM and the corresponding portions of the quasar spectra were normalised by dividing by the continuum level. Gaussian doublets were then fitted to the normalised spectra using a maximum-likelihood routine. Rest equivalent widths ($W$) were calculated from the parameters of the fitted Gaussian doublets and errors estimated by propagation of the parameter errors derived during the maximum likelihood fit. Absorption redshifts were redefined using the centre of the fitted Mg II $\lambda 2796$ line. A small number of Ca II doublets for which the fit proved to be visually unsatisfactory were removed from the sample.

The final catalogue consists of 37 Ca II systems with $W_{\lambda 3935} \gtrsim 0.35$ Å and $\langle z_{\rm abs} \rangle = 0.95$. Fig. 5.2 shows some examples of spectra with Ca II systems in regions around the Ca II $\lambda\lambda 3935, 3970$, Mg II $\lambda\lambda 2796, 2804$ and Fe II $\lambda\lambda 2587, 2600$ doublets. The rest frame equivalent widths for these lines and Mg I $\lambda 2853$ for all the Ca II systems can be found in Table 5.1.

**Table 5.1:** Name and spectroscopic identification of each quasar in the Ca II absorber sample, together with rest frame equivalent widths of Ca II $\lambda\lambda 3935, 3970$, Mg II $\lambda\lambda 2796, 2804$, Mg I $\lambda 2853$ and Fe II $\lambda 2600$. The final column gives the reddening of each quasar SED caused by the Ca II absorber, calculated assuming a Large Magellanic Cloud extinction curve. For one object, found in a quasar spectrum not well fit by the control template, an individual colour excess estimate was not possible and a value is not given.

| SDSS ID | MJD[a],plate,fibre | $i$[b] | $z_{\rm em}$ | $z_{\rm abs}$[c] | $W_{\lambda 3935,3970}$ | err($W$) | $W_{\lambda 2796,2804}$ | $W_{\lambda 2853}$ | $W_{\lambda 2600}$ | $E(B-V)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| J002133.36+004301.2 | 51900,0390,537 | 17.46 | 1.245 | 0.942 | 0.34, 0.22 | 0.10, 0.07 | 1.80, 1.66 | 0.55 | 0.99 | -0.027 |
| J010332.40+133234.8 | 51821,0421,049 | 18.32 | 1.660 | 1.048 | 1.07, 0.80 | 0.22, 0.19 | 3.06, 2.66 | 1.44 | 2.25 | 0.077 |
| J014717.76+125808.4 | 51820,0429,215 | 17.82 | 1.503 | 1.039 | 0.50, 0.17 | 0.12, 0.07 | 4.28, 4.26 | 1.15 | 3.05 | 0.043 |
| J074804.08+434138.4 | 51885,0434,340 | 18.54 | 1.836 | 0.898 | 0.53, 0.27 | 0.11, 0.11 | 1.71, 1.19 | 0.30 | 0.69 | 0.030 |
| J080736.00+304745.6 | 52319,0860,601 | 18.62 | 1.255 | 0.969 | 0.56, 0.79 | 0.16, 0.20 | 2.83, 2.70 | 1.24 | 2.23 | 0.003 |
| J080958.56+515118.0 | 53090,1780,532 | 18.56 | 1.289 | 0.902 | 0.64, 0.39 | 0.13, 0.11 | 2.25, 2.22 | 1.14 | 1.92 | -0.007 |
| J081054.00+352226.4 | 52378,0892,106 | 18.40 | 1.304 | 0.877 | 0.56, 0.31 | 0.08, 0.08 | 2.17, 2.11 | 0.99 | 1.78 | 0.030 |
| J081930.24+480827.6 | 51885,0440,007 | 17.75 | 1.994 | 0.903 | 0.75, 0.36 | 0.07, 0.07 | 1.69, 1.57 | 1.03 | 1.36 | *** |
| J083157.84+363552.8 | 52312,0827,001 | 17.98 | 1.160 | 1.126 | 0.73, 0.41 | 0.22, 0.16 | 2.52, 2.49 | 0.78 | 1.53 | -0.028 |
| J085221.36+563957.6 | 51900,0448,485 | 18.67 | 1.449 | 0.844 | 0.65, 0.49 | 0.23, 0.22 | 3.32, 3.03 | 1.24 | 2.45 | 0.016 |
| J085556.64+383231.2 | 52669,1198,100 | 17.62 | 2.065 | 0.852 | 0.45, 0.16 | 0.08, 0.07 | 2.65, 2.50 | 0.71 | 2.07 | 0.038 |
| J093738.16+562837.2 | 51991,0556,456 | 18.53 | 1.798 | 0.978 | 1.23, 0.62[d] | 0.07, 0.08 | 4.90, 4.34 | 2.35 | 3.21 | 0.296 |
| J095352.80+080104.8 | 52734,1235,465 | 17.47 | 1.720 | 1.023 | 0.48, 0.35 | 0.07, 0.08 | 0.91, 0.80 | 0.46 | 0.64 | 0.031 |
| J100000.96+514416.8 | 52400,0903,258 | 18.72 | 1.235 | 0.906 | 0.81, 0.58 | 0.26, 0.25 | 4.47, 3.90 | 1.41 | 2.47 | -0.017 |
| J100145.12+594008.4 | 52282,0770,087 | 17.85 | 1.186 | 0.899 | 0.47, 0.33 | 0.14, 0.10 | 0.64, 0.58 | 0.50 | 0.41 | 0.050 |
| J103024.24+561832.4 | 52411,0947,179 | 17.83 | 1.288 | 1.000 | 0.68, 0.33 | 0.21, 0.12 | 1.91, 1.84 | 0.95 | 1.58 | 0.031 |
| J112053.76+623104.8 | 52295,0775,455 | 17.41 | 1.130 | 1.072 | 0.57, 0.44 | 0.12, 0.11 | 2.02, 1.94 | 0.92 | 1.50 | 0.070 |
| J112932.64+020422.8 | 51992,0512,113 | 17.36 | 1.193 | 0.965 | 0.56, 0.53 | 0.10, 0.10 | 2.08, 2.03 | 0.71 | 1.64 | -0.002 |
| J113357.60+510845.6 | 52367,0880,288 | 18.30 | 1.576 | 1.029 | 1.25, 0.67 | 0.41, 0.30 | 2.66, 2.72 | 0.82 | 1.97 | 0.117 |
| J115244.16+571203.6 | 52765,1311,631 | 17.96 | 1.603 | 0.847 | 0.54, 0.35 | 0.13, 0.11 | 3.35, 3.19 | 1.15 | 2.33 | 0.217 |
| J120300.96+063440.8 | 53089,1623,209 | 18.47 | 2.182 | 0.862 | 1.42, 0.89 | 0.29, 0.24 | 5.57, 4.97 | 2.98 | 3.75 | 0.417 |
| J122144.64-001142.0 | 52000,0288,078 | 18.58 | 1.750 | 0.929 | 0.58, 0.23 | 0.18, 0.12 | 0.97, 0.82 | 0.72 | 0.74 | -0.008 |
| J122756.40+425631.2 | 53112,1452,505 | 17.17 | 1.310 | 1.045 | 0.39, 0.17 | 0.07, 0.08 | 1.61, 1.39 | 0.25 | 1.15 | 0.026 |
| J124659.76+030307.2 | 52024,0522,531 | 18.86 | 1.178 | 0.939 | 1.07, 0.60 | 0.38, 0.22 | 2.91, 2.78 | 1.30 | 2.15 | 0.196 |

[a] Modified Julian date

[b] PSF $i$-band magnitude, corrected for Galactic extinction

[c] Measured from Mg II $\lambda 2796$

[d] Multiple absorption line system. All lines are fit with double Gaussians and the quoted EW is the total of the two systems. When fitting Ca II , the velocity-separation of the systems is taken to be that determined from the associated Fe II ($\lambda$2600) lines.

**Table 5.2:** Continuation of Table 5.1

| SDSS ID | MJD,plate,fibre | $i$ | $z_{\rm em}$ | $z_{\rm abs}$ | $W_{\lambda 3935,3970}$ | err($W$) | $W_{\lambda 2796,2804}$ | $W_{\lambda 2853}$ | $W_{\lambda 2600}$ | $E(B-V)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| J130841.28+133130.0 | 53089,1772,122 | 18.60 | 1.954 | 0.951 | 0.87, 0.91 | 0.27, 0.29 | 1.82, 1.61 | 0.68 | 1.16 | 0.080 |
| J131058.08+010824.0 | 51985,0295,325 | 17.88 | 1.389 | 0.862 | 0.74, 0.49 | 0.13, 0.10 | 2.18, 2.27 | 1.33 | 1.46 | 0.208 |
| J140444.16+551637.2 | 53088,1324,421 | 18.46 | 1.589 | 1.070 | 1.33, 0.54 | 0.34, 0.22 | 1.98, 2.11 | 0.94 | 1.44 | 0.195 |
| J144104.80+044348.0 | 52026,0587,329 | 18.49 | 1.112 | 1.039 | 0.97, 0.64 | 0.25, 0.20 | 2.27, 2.38 | 1.05 | 1.93 | 0.154 |
| J145633.12+544832.4 | 52353,0792,242 | 17.97 | 1.518 | 0.880 | 0.47, 0.45 | 0.17, 0.14 | 4.02, 3.72 | 1.71 | 3.11 | 0.057 |
| J151247.52+573842.0 | 52079,0612,438 | 18.72 | 2.135 | 1.044 | 0.98, 0.59 | 0.26, 0.20 | 2.04, 2.28 | 0.75 | 1.57 | 0.142 |
| J153503.36+311832.4 | 53119,1388,068 | 17.79 | 1.510 | 0.904 | 0.84, 0.36 | 0.15, 0.12 | 2.11, 1.80 | 0.82 | 1.03 | -0.008 |
| J153730.96+335837.2 | 52823,1355,633 | 17.50 | 1.024 | 0.913 | 0.38, 0.50 | 0.09, 0.09 | 1.82, 1.78 | 0.70 | 1.16 | -0.002 |
| J160932.88+462613.2 | 52354,0813,070 | 18.70 | 2.361 | 0.965 | 0.65, 0.36 | 0.26, 0.19 | 1.07, 0.92 | 0.65 | 0.60 | -0.035 |
| J172739.12+530227.6 | 51821,0359,042 | 18.04 | 1.442 | 0.945 | 0.62, 0.53 | 0.15, 0.15 | 2.71, 2.57 | 0.98 | 2.17 | -0.028 |
| J173600.00+573104.8 | 51818,0358,529 | 18.31 | 1.824 | 0.873 | 0.81, 0.60 | 0.18, 0.15 | 2.01, 1.80 | 0.87 | 1.55 | 0.074 |
| J224511.28+130903.6 | 52520,0739,030 | 18.66 | 1.546 | 0.860 | 2.03, 0.93[4] | 0.38, 0.23 | 3.99, 3.67 | 1.77 | 3.39 | 0.013 |
| J233917.76-002942.0 | 51877,0385,229 | 18.33 | 1.344 | 0.966 | 0.74, 0.64 | 0.19, 0.19 | 2.71, 2.44 | 0.85 | 1.80 | 0.106 |

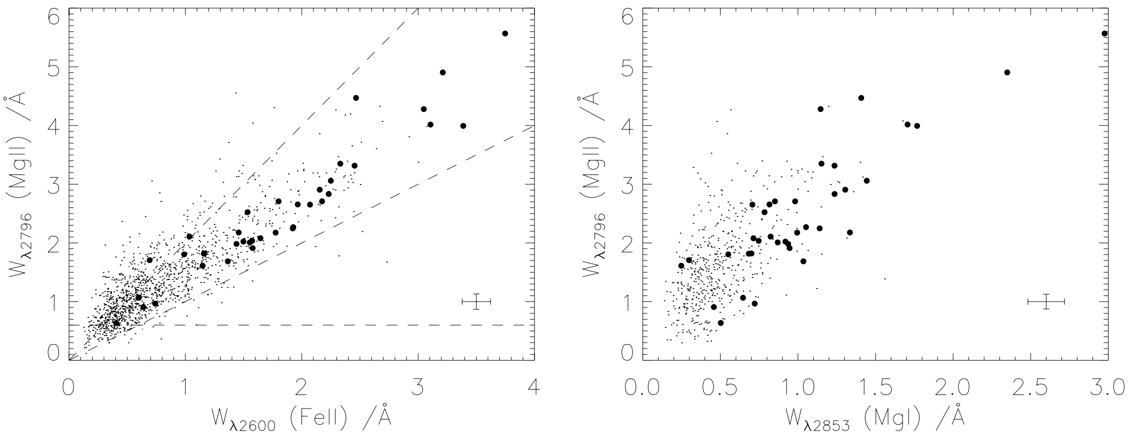

**Figure 5.3:** The rest frame equivalent widths of Mg II $\lambda2796$ vs. Fe II $\lambda2600$ on the left and Mg I $\lambda2853$ on the right for the Ca II absorbers (filled symbols) and Mg II absorbers (small points). The dashed lines in the left panel indicate the limits within which DLAs have been found in Mg II absorbers by Rao (2005). The error bars in the bottom right-hand corners show the typical errors on the equivalent widths.

### 5.3.2 Mg II systems

It is interesting to compare the properties of the rare Ca II absorbers with those of other classes of quasar absorption line systems. To this end, a catalogue of 2 338 Mg II absorbers was compiled from the sample used for confirmation of the DR3 Ca II absorption systems. The same fully parameterised fit was carried out on the Mg II absorbers as described above for the Ca II sample and, because the large number of systems made visual inspection unappealing, absorbers with poor continuum fits, unphysical line ratios and detection significances (for the Mg II doublet) of less than $4\sigma$ were culled automatically from the sample. Line significance was defined by

$$\mathcal{S} = \frac{\sum(f_i - f_{c,i})(m_i - f_{c,i})}{\sqrt{\sum \sigma_i^2 (m_i - f_{c,i})^2}} \tag{5.1}$$

where the sum is over i pixels, $f$ is the flux, $f_c$ the continuum, $m$ the fitted model and $\sigma$ the error on the flux and a full derivation of this equation can be found in Bolton et al. (2004)

Among these Mg II systems the subset satisfying the criteria for likely DLA candidates is of particular interest. Specifically, Rao (2005) found that 43% of strong ($W_{\lambda2796} > 0.6$Å) Mg II systems with associated Fe II $\lambda2600$ absorption such that $1 < W_{\lambda2796}/W_{\lambda2600} < 2$ are confirmed to be bona-fide DLAs by subsequent UV spectroscopy. 789 of our Mg II systems satisfy these requirements and we shall refer to this sample as the “Mg II-selected DLAs”.

### 5.3.3 Trends of strong metal line strengths

In Fig. 5.3, the equivalent widths of Mg II $\lambda2796$, Fe II $\lambda2600$ and Mg I $\lambda2853$ in the Ca II absorbers (filled symbols) are compared with those of the Mg II absorbers (small dots). In general, Ca II systems tend to have strong associated Mg II, Fe II and Mg I lines. Referring to the left-hand panel, it can be seen that all but two of the Ca II systems fall within the defining criteria of Mg II-selected DLAs and their average equivalent widths are significantly higher than the average for the Mg II-selected DLAs

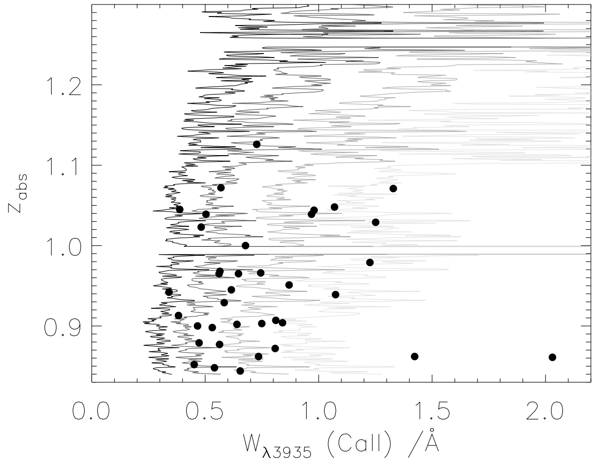

**Figure 5.4:** The probability of detecting a Ca II absorption system in the absorption line survey as a function of absorption redshift and rest equivalent width of the stronger member of the Ca II doublet, calculated from a series of Monte Carlo trials. From left to right, the contours are drawn at probabilities $P_{\rm Ca} = 0.1, 0.3, 0.7$ and 0.95. The example shown here assumes a Ca II doublet ratio $W_{\lambda3935} : W_{\lambda3970} = 4 : 3$ and that a Mg II doublet is detected for each Ca II system recovered, except when blended with the strong sky lines at $\lambda5579$ and $\lambda6302$. These sky lines give rise to the conspicuous gaps in detection efficiency at $z_{\rm abs} \simeq 0.99$ and 1.25. Filled circles correspond to the 37 Ca II absorbers detected in this survey.

in general. Turning to the right-hand panel, there is a hint that the Ca II absorbers may have stronger Mg I for a given equivalent width of Mg II than conventional Mg II absorbers, possibly indicating that they tend to arise in regions of high gas density where the recombination rate is higher (Hobbs 1974).

## 5.4 Selection function and total redshift path

In order to assess the statistical properties of the Ca II absorption line systems and compare them with other classes of absorbers, it is necessary to quantify the sensitivity of the absorption line survey as a function of Ca II equivalent width and wavelength. This question was addressed with Monte Carlo simulations by placing artificial Ca II and Mg II absorption lines in the SDSS quasar spectra and determining the fraction recovered using the same search techniques as in the real survey[5]. The simulations were run for doublets with ratios of 1:1, 2:1 and 4:3 for Ca II, and of 1:1 and 2:1 for Mg II, and the effect of varying the line width was investigated. The general conclusion from these different trials was that for the parameter ranges considered there was a variation in detection efficiency of at most 5-10%.

Fig. 5.4 shows the probability of detection, $P_{\rm Ca}$, of a Ca II system with doublet ratio $W_{\lambda3935} : W_{\lambda3970}$ =4:3 as a function of redshift $z_{\rm abs}$ and equivalent width of the stronger member of the doublet, $W_{\lambda3935}$. In this example it has been assumed that a Mg II doublet is detected for every artificial Ca II system recovered (thus setting $P_{\rm Mg} = 1$), except at redshifts which place the Mg II lines close to strong sky emission lines. Values of $z_{\rm abs}$ and $W_{\lambda3935}$ for the 37 Ca II systems are overplotted as filled circles. The probability of detection drops sharply at redshifts $z_{\rm abs} > 1.2$, as the Ca II doublet moves beyond 8500 Å in the observed frame; the pathlength available for finding absorbers also decreases with

[5] The Monte Carlo simulations were carried out by Paul Hewett.

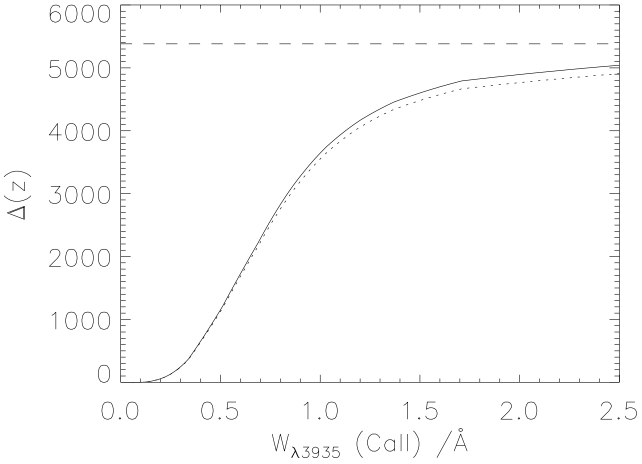

**Figure 5.5:** The total redshift path of the Ca II survey calculated from Monte Carlo trials. In the two examples shown here, the artificial Ca II lines used in the simulations had a fixed doublet ratio of 4:3, and the corresponding Mg II lines were either always detected (continuous curve) or assumed to have $W_{\lambda2796} = W_{\lambda2804} = 1.0\,$Å (dotted curve). The horizontal long-dash line shows the maximum path-length available if the probability of detection were $P_{\rm detect} = 1$ over the entire redshift path covered by each input quasar spectrum.

increasing redshift. Even so, it is perhaps surprising that *no* Ca II systems have been detected at these redshifts. The significance of this result is tested below.

By multiplying the number of quasar spectra in which a Ca II absorption system at a given redshift could be detected, $\mathcal{N}_{\rm quasar}$ ($z_{\rm abs}$), by the probability of detection given its values of $z_{\rm abs}$, $W_{\lambda3935}$ and $W_{\lambda2796}$, an estimate is obtained for the *effective* number of sightlines over which such a system could be expected to be found:

$$\mathcal{N}_{\rm eff}(z_{\rm abs}, W_{Ca}, W_{Mg}) = \mathcal{N}_{\rm quasar}(z_{\rm abs}) \times P_{\rm detect} \tag{5.2}$$

where

$$P_{\rm detect} = P_{\rm Ca}(W_{\rm Ca}, z_{\rm abs}) \times P_{\rm Mg}(W_{\rm Mg}, z_{\rm abs}) \tag{5.3}$$

and the notation for the rest frame equivalent widths of Ca II $\lambda3935$ and Mg II $\lambda2796$ has been shortened for convenience.

Integrating over redshift gives the total redshift path covered by the absorption line survey as a function of rest frame equivalent width:

$$\Delta Z(W_{\rm Ca}, W_{\rm Mg}) = \int_{z_{\rm abs,min}}^{z_{\rm abs,max}} \mathcal{N}_{\rm eff}\,(z_{\rm abs}, W_{\rm Ca}, W_{\rm Mg})\, d\,z_{\rm abs} \tag{5.4}$$

This function is plotted in Fig. 5.5 for two cases: one calculated assuming $R_{\rm Mg} = 1$ throughout (except when the Mg II lines are blended with sky lines) and the other adopting $W_{\lambda2796} = 1.0\,$Å (a conservative lower limit for the Ca II absorbers) and a Mg II doublet ratio of 1:1. The difference between the two cases is minimal.

It was noted above that no absorption systems are found between $1.2 < z_{\rm abs} < 1.3$ and we are now in a position to calculate the significance of this result. The total pathlength available for finding absorbers over this redshift range is only 0.064 times that available over the entire $0.84 < z_{\rm abs} < 1.3$ range. By adopting a binomial distribution with 37 trials and P($z_{\rm abs} > 1.2$) = 0.064, the probability of detecting 0/37 absorbers in this redshift range is 0.086, i.e. the significance of the lack of detection is $< 90$%. Perhaps with future SDSS catalogue releases it will be possible to confirm that the effect is indeed a low significance statistical fluctuation.

By combining the observed equivalent width distribution with equation (5.4), an estimate can be made of the intrinsic equivalent width distribution of the Ca II absorbers. This is shown alongside the Mg II equivalent width distribution in the two top panels of Fig. 5.6. Completeness corrections are only shown above 0.5 Å and 0.3 Å for the Ca II and Mg II samples respectively; below these values the corrections become large and uncertain due to small number statistics.

Finally, a statistic commonly used in absorption line surveys is the number of systems per unit redshift with rest equivalent width greater than some threshold value: $n(W^{\rm lim})$. While in general $n(W^{\rm lim})$ is a function of redshift, negligible redshift dependence is assumed here since the redshift interval is small. By setting the minimum equivalent width of Mg II $\lambda2796$ to be 0.6Å, smaller than measured for any of the Ca II systems in the survey, the joint dependency in the cumulative probability function is removed:

$$n(W^{\rm lim}_{\rm Ca}, W^{\rm lim}_{\rm Mg} = 0.6) = \sum_i \frac{1}{\Delta Z_i (W_{\rm Ca} \geq W^{\rm lim}_{\rm Ca})} \tag{5.5}$$

where the sum is over all absorbers in the sample with $W_{\lambda3935}$ greater than the limit. The function $n(W^{\rm lim}_{\rm Ca}, W^{\rm lim}_{\rm Mg} = 0.6)$ is shown in the bottom left-hand panel of Fig. 5.6; for $W^{\rm lim}_{\rm Ca\,II\,\lambda3935} = 0.5$ Å, the Ca II systems are found to have a number density per unit redshift $n(z) = 0.013$ (at a mean redshift $\langle z_{\rm abs} \rangle = 0.95$).

$n(W^{\rm lim})$ was also calculated for the Mg II absorption line systems; the results are shown in the bottom right-hand panel of Fig. 5.6. The results from the survey presented here are in good agreement with those determined by Nestor et al. (2005) from the SDSS EDR (dashed line) over a very similar redshift range.

## 5.5 Measuring the average reddening

This section turns to the estimation of the dust content of quasar absorption line systems through the reddening effect of the dust on the SED of the background quasar. Traditionally the study of reddening in quasar samples has been a difficult one due to the intrinsic variation in the shape of quasar SEDs. The large sample of quasars in the SDSS at similar redshifts to those with the identified absorption line systems, allows a good estimate of the average quasar spectrum to be obtained. Any reddening signal in the spectra of quasars with an intervening absorber can then be found by comparing their SEDs with that of the average quasar SED at the appropriate redshift.

### 5.5.1 Quasar reference spectra and dust extinction curves

Quasar reference spectra were created by combining spectra from the DR3 quasar catalogue, obeying the same SNR and magnitude criteria as for the the absorption line catalogue, in redshift bins of $\Delta z = 0.1$

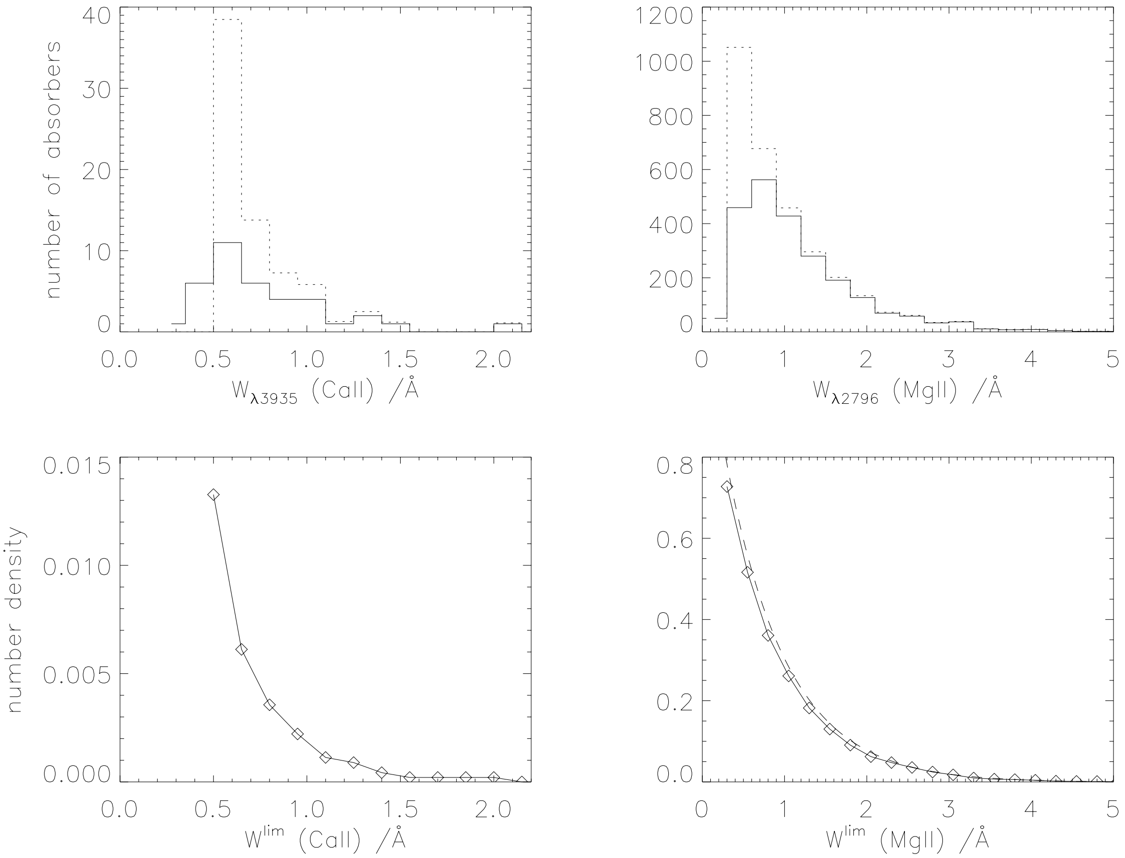

**Figure 5.6:** Top two panels: Equivalent width distributions of Ca II and Mg II absorption systems before (continuous histograms) and after (dashed histograms) completeness corrections. No corrections are shown for the smallest equivalent width bins where the statistics are too poor for reliable estimates. Bottom two panels: Number density of absorbers per unit redshift interval as a function of minimum equivalent width limit, $n(W^{\rm lim})$. The values deduced for the Mg II systems are in good agreement with the fit to the distribution of SDSS EDR Mg II systems with $0.871 < z_{\rm abs} < 1.311$ (dashed line) reported by Nestor et al. (2005).

staggered in redshift by $z = 0.05$. Each individual spectrum was first corrected for Galactic reddening using the quoted extinction in the SDSS photometric catalogue and the Galactic extinction curve from Cardelli et al. (1989), as extended by O'Donnell (1994). The spectra were then shifted to the quasar rest frame, being careful to allow for the flux/Å term in the SDSS spectra, and normalised by dividing by the median flux in a wavelength range common to all spectra in the particular redshift bin (avoiding the main quasar emission lines, strong sky lines and bad pixels). Finally, the spectra were combined into a composite spectrum using an arithmetic mean. Different methods of combining spectra were tested, and the details of the procedure were found to make no significant difference to the final composites. The statistical power of the SDSS is made clear by the sheer number of quasar spectra available for such an analysis. Up to a redshift of $z_{\rm em} = 1.9$ more than 500 quasars contribute to each composite and only for one quasar in our Ca II sample does the appropriate composite depend on less than 100 quasars.

Turning now to the spectra of quasars *with* absorbers: each one of these was shifted to the rest frame of the quasar, scaled by the median flux as above, and then divided by the reference composite quasar spectrum closest in redshift. The resulting spectrum, which contains any reddening signal, was then shifted to the rest frame of the absorber, $z_{\rm abs}$, before being combined with others to form the composite spectra to be analysed for reddening.

Several different composite spectra were constructed for both the Ca II absorbers and the Mg II-selected DLAs (i.e. those Mg II absorption line systems fulfiling the criteria of Rao (2005)). For the former, three composites were produced: the first combines all 37 absorbers together; the other two split the sample into two by $W_{\lambda3935}$ (at $W_{\lambda3935} \sim 0.65$). Henceforth these samples are termed 'All', 'High-$W_{\lambda3935}$', and 'Low-$W_{\lambda3935}$'. For the Mg II-selected DLAs four composites were created: the first combines all 789 absorbers; the other three contain subsamples of absorbers with increasing absorption line strength. The subsamples were defined by drawing diagonal lines across Fig. 5.3 from $(W_{\lambda2796}, W_{\lambda2600}) = (m, 0)$ to $(W_{\lambda2796}, W_{\lambda2600}) = (0, m)$, where $m = 1.5$, 2 and 3 Å, and taking only the systems that lie above each line and are within the Rao (2005) limits.

The reddening of each composite was measured by fitting extinction curves of three forms, appropriate for dust in the Milky Way, the Large Magellanic Cloud (LMC) or the Small Magellanic Cloud (SMC) to find the best fitting value of *E(B−V)*. The extinction curve tabulations of Cardelli et al. (1989), as extended by O'Donnell (1994), were used for the Milky Way and those of Pei (1992) were used for the Magellanic Clouds and a total-to-selective extinction ratio $R_V = 3.1$ was assumed throughout.

### 5.5.2 The significance of detection

Before discussing the results of the reddening analysis, the possibility of any systematic effects which may produce a spurious reddening signal need to be assessed. This was investigated with a series of Monte Carlo simulations in which artificial composites were constructed by drawing quasar spectra at random from the reference quasar sample *without* Ca II and Mg II absorption systems. The selection probability of each spectrum was modified by the redshift path $\Delta z_{\rm quasar}$ over which the spectrum was searched for intervening absorbers:

$$\begin{aligned} \Delta z_{\rm quasar} &= min[1.3, z_{\rm em} - 0.01] \\ &- max[0.84, 1250(1 + z_{\rm em})/2175 - 1] \end{aligned} \tag{5.6}$$

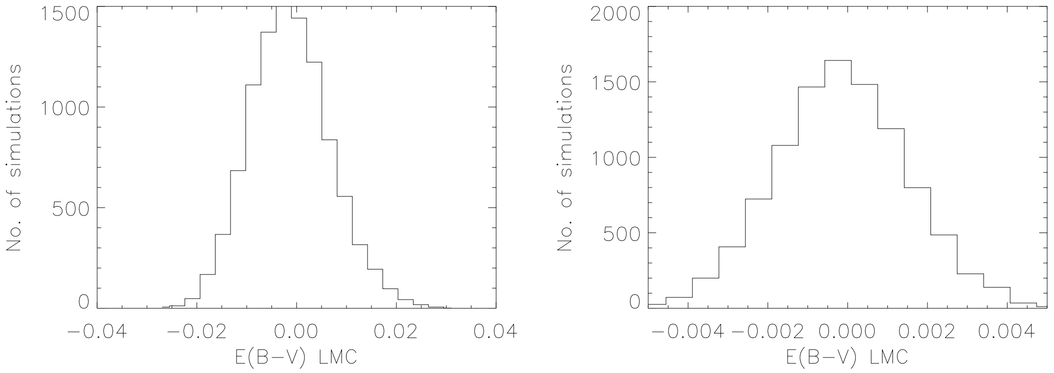

**Figure 5.7:** The distribution of apparent values of *E(B−V)* produced by Monte Carlo simulations of 10 000 sets of 37 (left) and 789 (right) randomly chosen quasar spectra when shifted randomly to rest frames corresponding to the range in redshift of our absorption systems. See text for additional details. The range of *E(B−V)* covered by the central 68% of simulations provides an estimate of the error on the reddening measured in the absorber composite spectra.

where $min$ and $max$ indicate the minimum and maximum of the values enclosed in the brackets. On selection, each quasar was assigned a random absorption redshift (within the range over which real absorbers were sought), and shifted to the rest frame of this "absorber". The process was repeated until artificial samples of the required size had been assembled (e.g. 37 for the entire sample of Ca II absorbers).

Composite spectra were then constructed for these artificial absorber samples and analysed for reddening as described above, assuming an LMC extinction curve, and the whole process was repeated 10 000 times. The final product of this exercise are the distributions of apparent *E(B−V)* values which arise simply from variations in the SED of quasars. The two distributions from the simulations designed to mimic the entire Ca II and Mg II-selected DLA samples, are reproduced in Fig. 5.7. The distributions have a mean *E(B−V)* close to zero and possess a small dispersion, extending to only a few thousandths of a magnitude for the large Mg II sample and to a few hundredths of a magnitude for the Ca II absorbers. Values of *E(B−V)* greater than 0.031 and 0.0065 are never found in the 10 000 simulations performed for the 37 Ca II absorbers and 789 Mg II-selected DLAs respectively. Errors on the reddening measured in the absorber samples are obtained by taking the 68th percentiles of these distributions.

## 5.6 Average dust content of metal absorption line systems

The results of the reddening analysis of the real absorbers are collated in Table 5.3 and illustrated in Figs. 5.8 and 5.9.

### *The* Ca II *absorbers*

Not only is there an unequivocal detection of reddening in the Ca II absorbers, a considerable range in dust content is seen within the sample, with a very pronounced difference between the 'High-' and 'Low-$W_{\lambda3935}$' subsamples (see lower panel of Fig. 5.8). For both the 'All' and 'High-$W_{\lambda3935}$' samples reddening values of the observed magnitude are never measured in the corresponding Monte Carlo simulations, making these results significant at the $> 99.99\%$ confidence level. The low value of *E(B−V)*

**Table 5.3:** Estimates of the reddening *E(B−V)* introduced by the Ca II absorbers and Mg II–selected DLAs in the spectra of background quasars assuming dust extinction curves suitable for dust in the Milky Way (MW), SMC and LMC. The final row gives the estimated error for each sample derived from Monte Carlo simulations using an LMC dust curve (see Section 5.5.2).

| | *E(B−V)*, Ca II absorbers | | | *E(B−V)*, Mg II-selected DLAs | | | |
|---|---|---|---|---|---|---|---|
| Dust law | All | High-$W_{\lambda3935}$ | Low-$W_{\lambda3935}$ | All | $m = 1.5$ | $m = 2$ | $m = 3$ |
| MW | 0.057 | 0.092 | 0.023 | 0.006 | 0.009 | 0.014 | 0.021 |
| LMC | 0.065 | 0.103 | 0.026 | 0.007 | 0.011 | 0.017 | 0.025 |
| SMC | 0.066 | 0.105 | 0.026 | 0.007 | 0.011 | 0.017 | 0.025 |
| err | 0.008 | 0.011 | 0.011 | 0.0016 | 0.0020 | 0.0024 | 0.0041 |

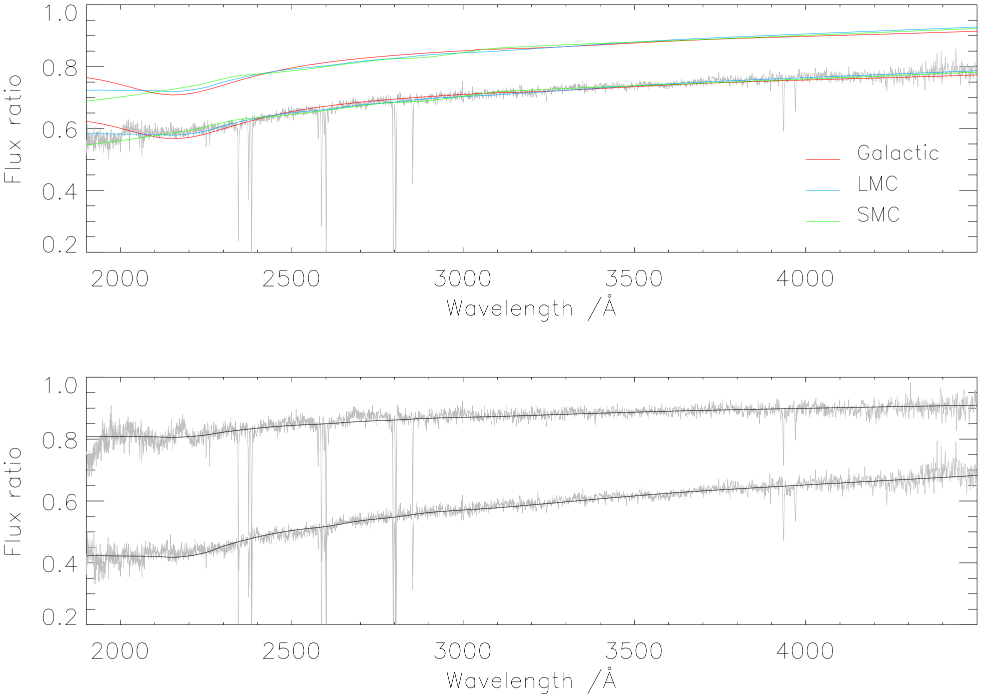

**Figure 5.8:** Top: Composite spectrum of the 37 quasars with Ca II absorbers after an estimate of the unabsorbed quasar SED has been divided out and each spectrum has been shifted to the rest frame of the absorber. The resulting composite is fitted with extinction curves appropriate to dust in the Milky Way and the Magellanic Clouds (as indicated) to deduce the values of colour excess, *E(B−V)*, listed in Table 5.3. The fits to the SED are shown superposed on the spectrum, and are also plotted above it, for clarity. Bottom: Same as for the upper panel, but with the Ca II sample now split into the 'High-' and 'Low-$W_{\lambda3935}$' subsamples. The former shows the higher extinction (lower spectrum in the bottom panel). Only the best-fit LMC extinction curves are shown.

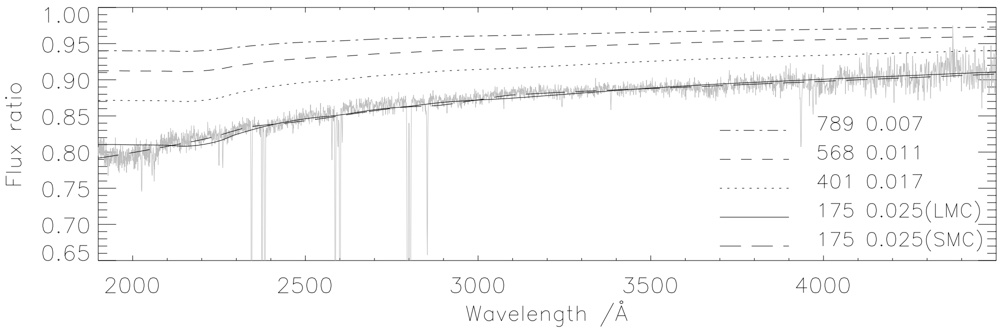

**Figure 5.9:** As for Fig. 5.8 but for Mg II-selected DLA candidates. The curves shown refer to different subsamples of Mg II-selected DLAs ordered by increasing values of $W_{\lambda 2796}$ and $W_{\lambda 2600}$, as described in the text. The full sample includes 789 quasar spectra and shows an average *E(B−V)*$= 0.007$ (for an LMC extinction law), as indicated in the lower right-hand corner of the figure. Corresponding values for the other subsamples are also given. The composite spectrum of the subsample of 175 Mg II-selected DLAs with the highest equivalent widths is shown in grey. Superposed on the spectrum are the two best-fitting LMC and SMC curves; both give *E(B−V)*$= 0.025$.

deduced for the 'Low-$W_{\lambda 3935}$' subsample, *E(B−V)*$\simeq 0.026$, is only marginally significant ($< 3\sigma$) given the results of the Monte Carlo simulations.

### *The* Mg II*-selected DLAs*

Turning to the Mg II-selected DLAs, it can be seen from Fig. 5.9 that they too exhibit a definite, but substantially lower, reddening signal. For the full sample of 789 absorbers (dot-dash line) a value of *E(B−V)*$= 0.007$ (LMC dust) is found which, while very low, is still significant at the $> 99.99$% confidence level according to the results of the Monte Carlo simulations.

The best-fit LMC extinction curves for the three subsamples split according to the strength of their strong lines are also shown in Fig. 5.9 with dashed, dotted and continuous lines corresponding to $n = 1.5, 2$ and $3$ Å respectively (see Section 5.5.1). The number of absorbers and value of *E(B−V)*$_{\rm LMC}$ for each subsample are listed in the bottom right-hand corner of the figure. For the composite of 175 absorbers with the strongest Mg II $\lambda 2796$ and Fe II $\lambda 2600$ absorption lines ($n = 3$Å; plotted in grey), the best fitting SMC extinction curve is also shown (long dash). It is evident from Fig. 5.9 that, as is the case for the Ca II absorbers, Mg II-selected DLA candidates exhibit a clear trend of increasing dust content with increasing absorption line equivalent width. Even accounting for the $\sim 57$% of sub-DLAs in the sample, the average dust content of DLAs at these redshifts is tiny. However, for the 20% of absorbers with the highest absorption line equivalent widths, the average *E(B−V)* of the DLAs in the sample is similar to the average of the Ca II absorbers, assuming that sub-DLAs contribute negligible dust reddening.

### *Difference in dust curves?*

From Figs. 5.8 and 5.9 it can be appreciated that at these mid-UV wavelengths and low levels of extinction there are only subtle differences between the three extinction laws fitted to the data. Even so, a 2175 Å bump as strong as that produced by Galactic dust can probably be excluded for all of the compos-

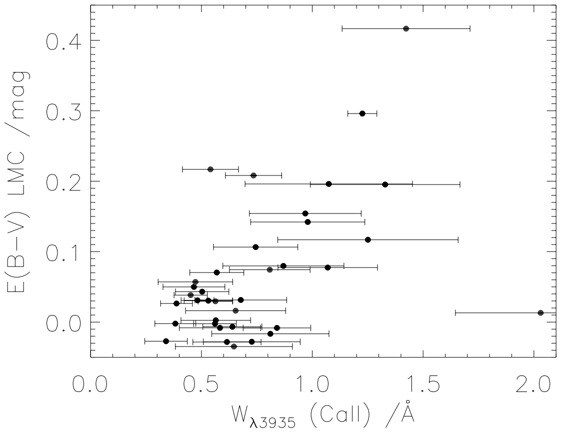

**Figure 5.10:** Colour excess caused by individual Ca II absorbers in SDSS quasar spectra plotted as a function of the equivalent width of Ca II $\lambda3935$.

ites. The LMC extinction curve, which flattens near 2250Å, seems to give the best fit to the reddening produced by the Ca II absorbers. On the other hand, the steeply rising SMC curve appears to be a better fit to the DLA candidates selected via Mg II and Fe II. It will be of interest to establish, with more extensive samples, whether these differences are significant, since they are ultimately related to differences in the size distribution and composition of the dust particles associated with different classes of quasar absorbers.

### *Individual reddening results*

Finally, while the results of the reddening analysis are most secure when applied to ensembles of absorbers, so that variations due to the intrinsic SEDs of the background quasars are mitigated, it is nevertheless of interest to examine the absorbers' reddening on a case by case basis. The results are also used in Section 5.7.1 for calculating the obscuration bias inherent in the absorber sample. Each quasar with a Ca II system was divided by the relevant quasar reference spectrum and an LMC extinction curve fitted to the quotient in the rest frame of the absorber to provide individual values of *E(B−V)*. These values are listed in the last column of Tables 5.1 and in Fig. 5.10 they are plotted against the equivalent width of Ca II $\lambda3935$. Despite the scatter, the trend of increasing reddening with equivalent width highlighted by the analysis of the 'High-' and 'Low-$W_{\lambda3935}$' subsamples can be discerned.

## 5.7 Discussion

The detection of dust in a class of quasar absorption line systems allows for the first time a direct estimate of the dust obscuration bias against such systems, up to a certain *E(B−V)* limit. By comparing with gas-to-dust ratios observed in the local Universe estimates can be obtained of the H I column densities of the systems, important for relating them to other classes of absorption line systems.

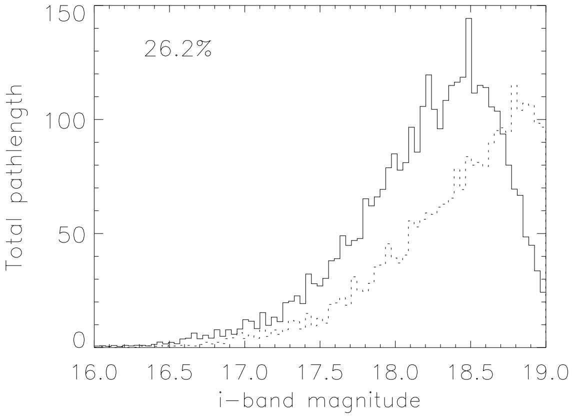

**Figure 5.11:** The effect on the total survey pathlength, were each quasar line of sight to be intercepted by an absorber with *E(B−V)* = 0.065, $z_{\rm abs} = 0.95$ and $W_{\lambda 3935} = 0.76$ (the mean values of the 37 Ca II absorbers). The continuous histogram shows the true pathlength as a function of $i$-band magnitude, while the dotted line shows the observed distribution. Quasars which fall to the right of the $i = 19.0$ survey limit would not be included in the survey if an "average" Ca II absorber was present along their line of sight. The remaining quasar spectra would have lower SNRs, decreasing the probability of line detection. The value in the top left is the overall percentage of pathlength lost.

### 5.7.1 Dust obscuration bias

There has been much discussion in the literature concerning possible selection bias in magnitude-limited quasar samples against dusty, metal-rich DLAs at high redshift because of the obscuration they would cause to the background quasar light (see Section 5.1.3). This discussion is begun with a calculation of the dust obscuration bias against the Ca II systems, necessary for comparison of their true number density to that of DLAs. The attenuation which a sample of similar objects would cause at higher redshifts is also investigated.

The overall degree of obscuration bias against a population of absorbers with a well sampled *E(B−V)* distribution in the magnitude limited SDSS quasar sample can be estimated as follows:

1. Calculate the $i$-band extinction A$_i$ (in magnitudes) caused by each observed absorber individually, using its value of *E(B−V)*$_{\rm LMC}$.

2. Determine the effective pathlength available in the SDSS quasar sample for finding each absorber, given its values of $z_{\rm abs}$, $W_{\lambda 3935}$ and A$_i$.

3. Correct the number of absorbers observed in the survey for these reduced pathlengths in order to obtain the number of absorbers that would be found in an equivalent, but unbiased, survey.

For the case of the Ca II absorbers the primary uncertainty in the final result arises from step 1 above: the individual *E(B−V)* values are estimates only and the precise distribution towards the larger values of *E(B−V)* is unconstrained due to small number statistics. The calculation is therefore restricted to the region of the distribution that *is* well sampled, and the obscuration bias quoted only for systems with *E(B−V)*$\lesssim 0.25$.

In the calculation of total pathlength as a function of magnitude it was necessary to include the SNR, and therefore magnitude, dependency of the detection probability, which in turn depends upon the redshift and $W_{\lambda3935}$ of each absorber. For an absorber with small equivalent width the total pathlength available in the faint quasars is smaller than would naively be expected, due to the lower SNRs of the spectra. A similar method to that of Section 5.4 was applied, using Monte Carlo simulations to find the recovery probability of artificially placed lines, but in this case as a function of quasar magnitude.

The procedure for determining the dust bias is illustrated in part by Fig. 5.11. The dotted histogram shows the pathlength of the survey available to find an absorber with $z_{\rm abs} = 0.95$, *E(B−V)* = 0.065 (corresponding to A$_i$ $= 0.30$), and $W_{\lambda3935} = 0.76$ (these being the mean values of the full Ca II absorber sample), compared to the pathlength available to find the same absorber if it contained no dust (continuous histogram). Those quasars which are dimmed beyond the survey magnitude limit ($i$=19.0) would no longer be included, and the overall pathlength in all the other quasar spectra is reduced because of their lower SNR; as the figure indicates, 26.2% of the original pathlength is lost. In the simplified case that all 37 absorbers have this single redshift and dust content, statistically it would be expected to find 37/(1-0.262) $\approx$ 50 absorbers in an equivalent survey that was not subject to the effect of extinction.

More realistically, the fact that the correction factor for absorbers with a given *E(B−V)* depends non-linearly on the value of *E(B−V)* must be accounted for. Absorbers with large values of *E(B−V)* are responsible for significantly larger correction factors and simply adopting the average *E(B−V)* of the whole sample results in an underestimate of the fraction of absorbers missed. However, as stated above, this calculation requires a well sampled *E(B−V)* distribution and it was therefore decided to restrict the obscuration bias calculation to Ca II absorbers with *E(B−V)*$\lesssim 0.25$. This naturally results in a lower limit for the population as a whole, but was felt to be more robust than attempting an estimate including the two dustiest absorbers. Following through the steps outlined above for the 35 absorbers with *E(B−V)* below this limit, an unbiased number of 58 absorbers is found. Therefore, the obscuration bias towards Ca II absorbers with *E(B−V)*$\lesssim 0.25$ is 40% $(1 - 35/58)$. This value is insensitive to binning of the data within sensible ranges of bin size. These 58 absorbers would have an average *E(B−V)*$\sim 0.1$.

In the following subsections of the discussion it is shown that the dust content of the Ca II absorbers is consistent with their identification as DLAs, and their number density is around 20-30% of that of DLAs at similar redshifts. In Section 5.6 it was also seen that the majority of intermediate redshift DLAs (that is those selected from their Mg II and Fe II lines) contain negligible quantities of dust. Based on these results, obscuration in the intermediate redshift DLA population as a whole will cause around 8–12% of DLAs with *E(B−V)*$\lesssim 0.25$ to be missed from optical magnitude limited surveys. For this *E(B−V)* limit the results are well within the limits on obscuration bias placed by the radio selected survey of Ellison et al. (2004), although an increasing fraction of DLAs with dust contents greater than this limit would be obscured.

### *Bias at higher redshifts*

To illustrate the impact of extinction caused by intervening dust at higher redshifts on the SDSS quasar sample, Fig. 5.12 shows the observed frame $i$-band extinction as a function of $z_{\rm abs}$ caused by an intervening galaxy with LMC reddening *E(B−V)* = 0.01, 0.05, 0.1 and 0.15. It can be seen that an absorption line system at $z_{\rm abs} = 1$ with *E(B−V)*~0.1 mag and LMC type dust causes just under 0.5 mag of extinction of the quasar light. The same system at $z_{\rm abs} = 2.5$ would cause 0.9 mag of extinction; the effect of dust obscuration is clearly considerably greater at the redshifts probed by traditional DLA studies ($z_{\rm abs} > 2.2$

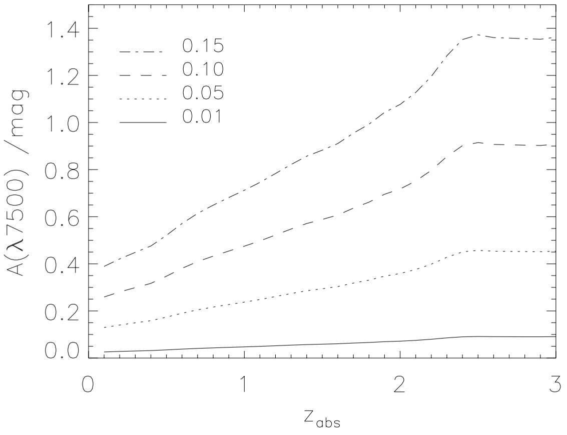

**Figure 5.12:** The observed frame ($i$-band) extinction of quasar light caused by LMC type dust in an intervening galaxy as a function of $z_{\rm abs}$. The different curves are for different columns of dust as parameterised by the *E(B−V)* values given in the top left.

in the SDSS).

The dust obscuration analysis was repeated for a population of imaginary Ca II absorption systems with the same dust properties as the 35 intermediate redshift systems with *E(B−V)*$\lesssim 0.25$, but with randomly assigned redshifts of between 2.2 and 3. The quasar sample is taken from the DR3 catalogue of Schneider et al. (2005) with $2.2 < z_{\rm em} < 4$ and $i < 20.2$. Assuming that the identification of DLAs is independent of quasar magnitude (a reasonable approximation when considering strong absorption features such as a damped Lyman-$\alpha$ line rather than weak Ca II absorption), leads to the prediction that $\sim 75$% of the Ca II systems with *E(B−V)* $\lesssim 0.25$ observed at intermediate redshifts would be missed at these higher redshifts. The calculation strongly suggests that were a subset of high redshift DLAs to contain small amounts of dust similar to that found in the Ca II absorbers, the vast majority would not be included in current magnitude-limited quasar samples. With the reddening statistics currently available at higher redshifts (e.g. Murphy & Liske 2004), a scenario in which 20–30% of DLAs have an average *E(B−V)*$\gtrsim 0.1$, as found here for intermediate redshifts, would not be distinguishable from one in which DLAs are dust free.

### 5.7.2 The number density of Ca II absorbers and DLAs

The number density per unit redshift of a class of quasar absorption line systems, $n(z)$, is a measure of their total cross-section on the sky. By assuming a space density for the galaxies responsible for the absorption a radius can be deduced within which a sightline must pass through a typical galaxy to show absorption (Section 5.1.1). $n(z)$ for the Ca II absorbers will therefore place them within the broader context of this simple model for absorption line systems in general. Specifically, in Section 5.4 the number of Ca II absorbers per unit redshift with a minimum equivalent width of 0.5Å, $n(z, W^{\rm lim} = 0.5)$ was shown to be $\sim 0.013$. Correcting this value for the incompleteness due to reddening of the background quasars as discussed above, $n(z, W^{lim} = 0.5) \gtrsim 0.022$ is obtained, where the lower limit arises from the unconstrained obscuration bias for systems with *E(B−V)*$\gtrsim 0.25$. Comparing with the values of $n(z)$ for DLAs given by Rao (2005) of $0.079 \pm 0.019$ at $0.11 < z_{\rm abs} < 0.9$ and $0.120 \pm 0.025$

at $0.9 < z_{\rm abs} < 1.25$, strong Ca II absorption systems are found to have a number density of about 20–30% that of DLAs. While the $n(z)$ calculation for the Ca II absorbers does not account properly for the dustiest systems, it is possible that a fraction of DLAs may be similarly missed due to dust, introducing an unavoidable uncertainty into these results.

Finally, in the simple model of quasar absorption line systems of Section 5.1.1, the estimated $n(z)$ confines Ca II absorbers—and their associated reddening—to the innermost 7–8 $h^{-1}$ kpc of galaxies, where the neutral gas column density and/or metallicity are expected to be highest.

### 5.7.3 Constraining N(H I) through dust-to-gas ratios

Ultimately, the neutral hydrogen column densities of the Ca II absorbers are required in order to fully understand their relation to other classes of quasar absorption systems. Unfortunately, this will not be possible for at least several years, given the current lack of space-borne UV spectrographs. Consequently, indirect arguments must be relied upon based on known relations in the ISM of the Milky Way and local galaxies. In this section, gas-to-dust ratios, $\langle N$(H I) $\rangle / \langle$*E(B−V)*$\rangle$, observed in the local Universe are used to indicate possible $N$(H I) column densities for the Ca II absorbers.

The gas-to-dust ratios of the Milky Way, Large and Small Magellanic Clouds are $0.5\times10^{22}$, $2.0\times10^{22}$ and 7.9 $\times$ $10^{22}{\rm cm}^{-2}{\rm mag}^{-1}$ respectively. The Milky Way value will give a lower limit to $N$(H I) for a given reddening *E(B−V)* if, 1) the gas-to-dust ratio increases with decreasing metallicity, as is suggested by the trends between these three galaxies, and 2) the Ca II absorbers have lower metallicities than the Milky Way, as would generally be assumed for $z \sim 1$ galaxies. Taking the *E(B−V)* values from the fits of the Milky Way dust extinction curve (first row of Table 5.3), values of log[$N$(H I)] $\geq$20.45, 20.66 and 20.06 are found for the three samples (‘All’, ‘High-$W_{\lambda3935}$’ and ‘Low-$W_{\lambda3935}$’). Correcting for the incompleteness caused by dust obscuration, as discussed in Section 5.7.1, would lead to log[$N$(H I)] $\geq 20.69$ for Ca II absorption line systems in general. In this case the lower limit arises both from the expectation of lower dust-to-gas ratios for the systems and the knowledge that the bias against the dustiest systems has not been properly accounted for. Irrespective of this correction, the limits imply that Ca II absorption systems with $W_{\lambda3935} > 0.5$ Å are very likely to be DLAs.

### 5.7.4 Dust in intermediate redshift galaxies

The discussion is now concluded by comparing the reddening found here for the Ca II absorption-selected galaxies with measurements for high redshift galaxies detected directly via their stellar or dust emission. Based on fits to model SEDs, Shapley et al. (2001) and Papovich et al. (2001) independently found the *E(B−V)* of $z \sim 3$ Lyman Break galaxies (LBGs) to range between 0 and 0.4 with a median of ∼0.15. The advent of SIRTF allowed the same analysis for UV–selected galaxies at $z \sim 2$: Shapley et al. (2005) found a comparative range in *E(B−V)* values to LBGs. Borys et al. (2005) performed an equivalent analysis on a sample of sub–millimetre selected galaxies, finding these particular SCUBA sources to have stellar masses similar to the most massive of the UV–selected galaxies and an average extinction in the $V$–band, A$_V$, of 1.7 equivalent to an *E(B−V)* of 0.55 assuming a LMC dust curve. Fig. 5.13 compares the observed average *E(B−V)* values of Ca II absorbers to the distribution of UV–selected $z \sim 2$ galaxies from Shapley et al. (2005) and the average value for the SCUBA sources of Borys et al. (2005).

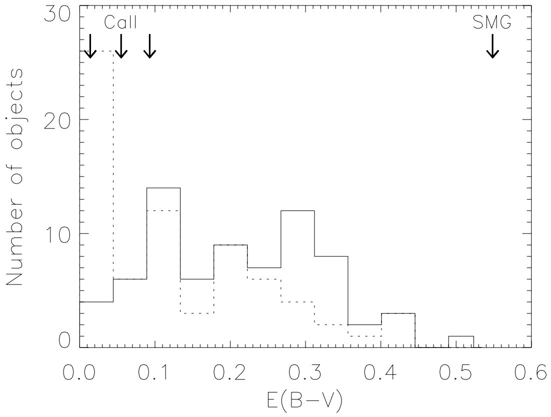

**Figure 5.13:** The distribution of *E(B−V)* for UV–selected galaxies at $z \sim 2$ from Shapley et al. (2005). The solid and dotted histograms show the best-fitting *E(B−V)* for a constant or bursting star forming model respectively. Arrows indicate the average *E(B−V)* values of the observed Ca II absorbers, uncorrected for dust obscuration effects, and the average value for the sub-millimetre selected galaxies of Borys et al. (2005).

While it is not surprising to find that the highest values of *E(B−V)* apply to galaxies that are luminous at far-IR/sub-mm wavelengths, where the emission by dust is seen directly, it is noteworthy that the more normal star-forming galaxies selected via their rest frame UV stellar light are also generally more reddened than the Ca II absorbers. An obvious question regarding the differences in the reddening between absorption- and emission-selected galaxies is the effect of dust obscuration bias on the former sample. It is clearly possible, given the calculations of dust bias presented in this chapter, that the majority of absorbers with dust contents as high as the upper end of the UV–selected galaxy distribution are obscured from sight. Whether this represents a significant population of objects depends primarily on the overall cross section of the sky with column densities of dust this large. One possible added complication to a direct comparison is that the reddening curves used in the derivation of *E(B−V)* are very different for the two cases, because of the geometrical configuration of dust and star formation in galaxies.

Until now, the two populations of absorption- and emission-selected galaxies have been differentiated by dust content, as well as metallicity and star formation rates. It is thus of potential importance that the lower limit of $\langle$*E(B−V)*$\rangle \gtrsim 0.1$ deduced for the Ca II absorbers which, after correcting for the missing systems due to dust bias, is not substantially lower than the median *E(B−V)*$\sim\ 0.15 - 0.20$ of star-forming galaxies at $z = 2 - 3$. Possibly, Ca II-selected DLAs are an intermediate link in a sequence of star formation rates, metallicities and dust content which stretches from the relatively quiescent, metal- and dust-poor DLAs to the actively star-forming, metal and dust-rich LBGs. Only further observations will provide the information needed to confirm or refute this.

## 5.8 Summary

In this chapter a sample of Ca II selected absorption line systems found in the SDSS quasar catalogue have been presented. These systems represent the first class of absorption line systems to unequivocally show reddening of the background quasar SEDs due to significant quantities of dust. By making the reasonable assumption that these galaxies at $z \sim 1$ have lower metallicities than the Milky Way, and that gas-to-dust ratios decrease with metallicity as seen in the local Universe, a lower limit of log[$N$(H I)] $\geq 20.45$ is obtained for the observed systems. This is greater than the nominal limit for DLAs, thus suggesting that Ca II absorption may provide a tool by which large samples of DLAs with $z < 1.3$ may be obtained from the SDSS quasar survey. It is noted that, after correcting for dust obscuration bias, the mean *E(B−V)* of the absorbers is not substantially lower than the median values of LBGs and UV-selected $z \sim 2 - 3$ galaxies. The question of the metal content of the Ca II absorbers is addressed in the following chapter and a more extensive summary is provided at the end of that chapter.

*Calcium....that's what*
*makes our teeth*

David from GW Sailability

# 6

# The element depletion pattern of Ca II absorption systems

*In the previous chapter a sample of Ca* II *absorption line systems found in the SDSS quasar catalogue was shown to contain significant quantities of dust, unlike average DLAs selected through their strong Mg* II $\lambda2796$ *and Fe* II $\lambda2600$ *lines. In this chapter the question of the H* I *column densities of the Ca* II *absorption line systems is further addressed by measuring the average column densities of unsaturated lines. The pattern of metal depletion in the interstellar media of the Ca* II *absorbers is presented and the dust-to-metals ratio measured directly for the first time in galaxies selected by absorption.*

## 6.1 Introduction

There is one further important piece of information that can be obtained for the Ca II absorbers from the moderate resolution and signal-to-noise ratio (SNR) SDSS spectra. Very weak metal transition lines cannot be seen in individual absorption spectra, however, by combining all the spectra into a single composite a high enough SNR can be reached to obtain average measurements. These average values can be interpreted easily if the gas in the Ca II absorbers is predominantly neutral, as for DLAs with their high column densities of H I, and the distribution of equivalent widths of the lines among the individual systems is reasonably uniform, as suggested by inspection of the individual spectra.

**Table 6.1:** The average reddening *E(B−V)* introduced by the 27 Ca II absorbers with $z_{\rm abs} > 0.88$ in the spectra of background quasars. The final row presents the errors for the LMC dust extinction curve, obtained from Monte Carlo simulations of random samples of quasars.

| | *E(B−V)* | | |
|---|---|---|---|
| Dust law | All | High-$W_{\lambda3935}$ | Low-$W_{\lambda3935}$ |
| MW | 0.048 | 0.082 | 0.012 |
| LMC | 0.055 | 0.093 | 0.014 |
| SMC | 0.055 | 0.095 | 0.015 |
| err | 0.009 | 0.013 | 0.014 |

### 6.1.1 The effect of dust on metal abundances

The dust content and metallicity of galaxies is intertwined, as more metal rich systems have the greater potential to form dust grains, which in turn selectively deplete metals in the ISM of the galaxies. Those metals which are depleted are termed “refractory”, those that do not show an affinity to dust, such as Zn II are termed “volatile”. Because Zn II is expected to be relatively undepleted it plays a crucial role in this analysis, providing a reference point against which the other metals can be contrasted (e.g Pettini et al. 1990). It is for this reason that the analysis of this chapter is restricted to those absorbers with $z_{\rm abs} > 0.88$, in whose spectra the wavelengths of the Zn II $\lambda\lambda2026, 2062$ doublet are encompassed. Table 6.1 presents the average reddening values for this subsample of 27 Ca II absorbers, equivalent to Table 5.3 for the full sample.

The selective depletion of metals leads to “depletion patterns” which are dependent upon the density, temperature and ionisation state of ISM (Savage & Sembach 1996). The impressive spectral range of the SDSS spectra allows the relative abundances of Zn II, Cr II, Fe II, Mn II, Ti II and Ca II to be measured from unsaturated lines between $0.88 < z_{\rm abs} < 1.3$. By comparing the chemical composition of the Ca II systems to those of DLAs and different interstellar media in the Milky Way further insight into the nature of the ISM in the Ca II absorbers may be obtained.

### 6.1.2 Ca II in the local Universe

Ca is an $\alpha$ element thought to be produced mainly by Type II supernovae. The convenient positioning of the ionic resonance lines of Ca II in the optical spectrum have made it accessible to astronomers for over a century with major Galactic surveys by Adams (1949) and Marschall & Hobbs (1972). More recent work has made use of high resolution spectroscopy of the Ca II lines for deciphering the structure of the ISM in the Milky Way and Magellanic Clouds (e.g. Welty et al. 1996). However, interstellar Ca II is compromised in its use as a measure of chemical abundance for two reasons. Firstly Ca is highly depleted onto dust grains: it is one of the most depleted elements in the ISM of the Milky Way with $\gtrsim$99% in the solid state. Second, the ionisation potential of Ca II is 11.9eV, just below that of H I at 13.5eV, making Ca III the dominant ion at the temperatures expected in the ISM of galaxies. Its severe depletion make it particularly interesting for understanding dust destruction/accretion processes in the ISM however (Crinklaw et al. 1994; Sembach & Danks 1994), with a small fraction of destroyed dust grains greatly increasing the column density of Ca II while little impact on the measured line-of-sight

**Table 6.2:** Rest frame equivalent widths of metal lines measured in the three Ca II absorber composites.

| | | | $W$ (mÅ) | | |
|---|---|---|---|---|---|
| Ion | Wavelength (Å) | $f$-value [a] | All | High-$W_{\lambda3935}$ | Low-$W_{\lambda3935}$ |
| Ca II | 3934.775 | 0.6267 | 725± 30 | 1011± 45 | 489± 24 |
| Ca II | 3969.590 | 0.3116 | 352± 29 | 549± 44 | 371± 29 |
| Ti II | 3242.918 | 0.232 | 121± 11 | 120± 22 | 92± 19 |
| Ti II | 3384.730 | 0.358 | 57± 10 | ... | 87± 18 |
| Mg I | 2852.963 | 1.83 | 782± 13 | 928± 30 | 717± 20 |
| Mg II | 2796.354 | 0.615 | 2256± 13 | 2472± 28 | 2191± 22 |
| Mg II | 2803.531 | 0.306 | 2106± 13 | 2238± 28 | 2067± 22 |
| Fe II | 2600.173 | 0.239 | 1722± 15 | 1799± 27 | 1598± 18 |
| Fe II | 2586.650 | 0.0691 | 1357± 17 | 1351± 26 | 1268± 19 |
| Mn II | 2606.462 | 0.198 | 147± 12 | 122± 21 | 100± 17 |
| Mn II | 2594.499 | 0.280 | 213± 14 | 231± 21 | 146± 17 |
| Mn II | 2576.877 | 0.361 | 277± 14 | 395± 29 | 227± 17 |
| Fe II | 2382.765 | 0.320 | 1582± 16 | 1606± 28 | 1554± 19 |
| Fe II | 2374.461 | 0.0313 | 929± 16 | 1014± 30 | 891± 20 |
| Fe II | 2344.214 | 0.114 | 1325± 16 | 1360± 29 | 1385± 24 |
| Fe II | 2260.781 | 0.00244 | 128± 17 | 129± 30 | 135± 20 |
| Fe II | 2249.877 | 0.00182 | 109± 15 | 95± 27 | 87± 17 |
| Cr II | 2066.164 | 0.0512 | 46± 12 | 80± 16 | 84± 17 |
| Cr II (+Zn II) | 2062.236 | 0.0759 (0.246) | 114± 15 | 195± 22 | 111± 17 |
| Cr II | 2056.257 | 0.103 | 92± 18 | 90± 25 | 108± 19 |
| Zn II (+Mg I) | 2026.137 | 0.501 (0.113) | 174± 19 | 267± 26 | 117± 19 |

[a] Rest frame wavelengths and $f$-values from Morton (2003)

reddening is noticed.

An unexplained underabundance of Ca with respect to other $\alpha$ elements by up to a factor of two has been noted in stars in early type galaxies, possibly suggesting metallicity dependent SNe yields not predicted by models (e.g. Thomas et al. 2003). While this is not of any consequence for this chapter, where Ca II line strength will reflect the dust depletion and ionisation state of the gas, it emphasises the uncertainties intrinsic in studies of elemental abundances at high redshifts due to our still imprecise knowledge of their underlying cosmic abundance from stellar models.

## 6.2 Method: obtaining relative metal abundances

In this chapter the column densities and relative abundances of different ions associated with the Ca II absorbers whose absorption lines are covered in the SDSS spectra are deduced. The unsaturated absorption lines which are required for precise abundance analyses are generally too weak to be detected and measured in the individual SDSS quasar spectra, however composites can provide sufficient SNR to obtain average column density measurements.

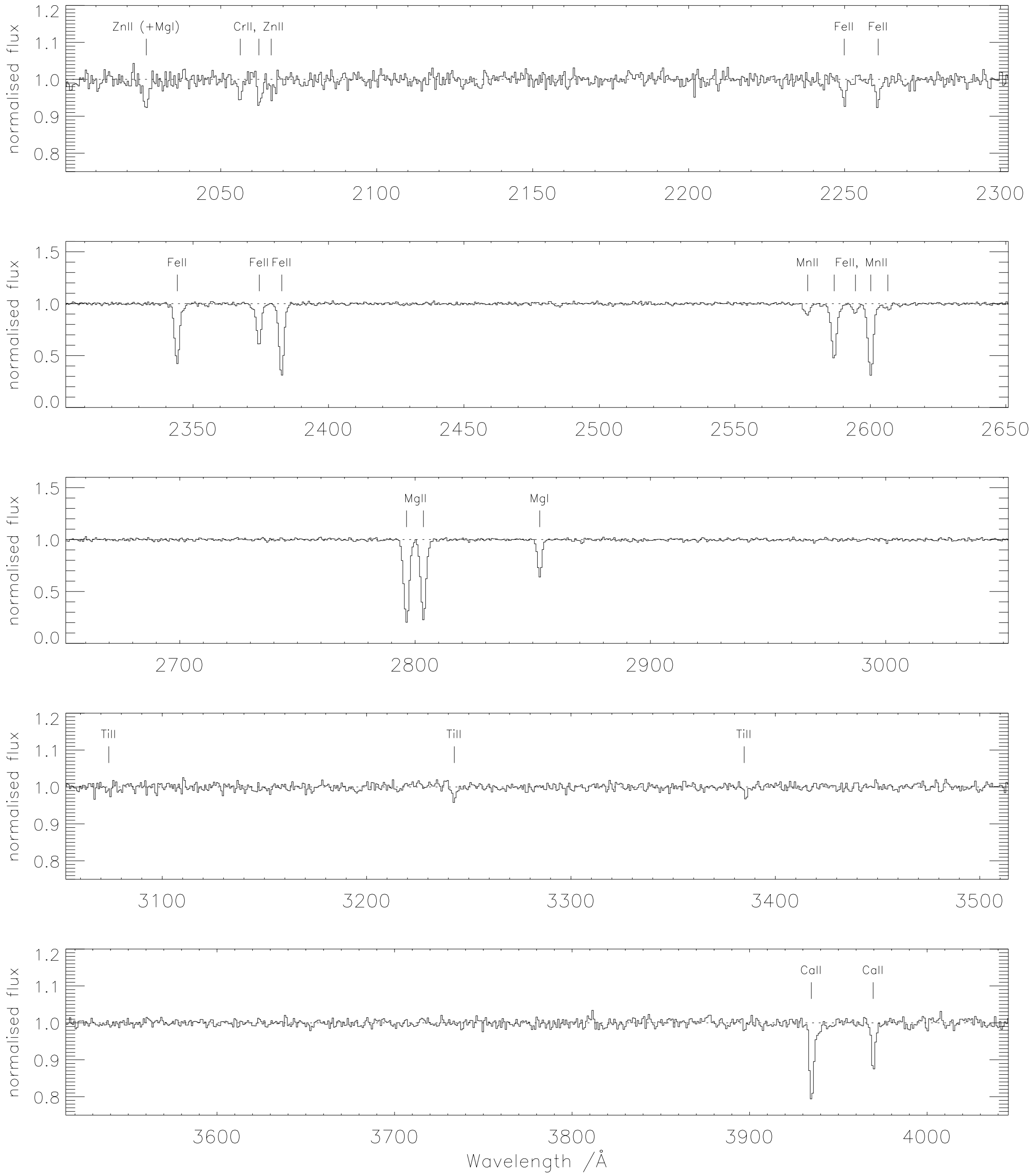

**Figure 6.1:** The composite spectrum of the 27 Ca II absorption systems with $\langle z_{\rm abs} \rangle = 0.98$; transitions listed in Table 6.2 are indicated. Note that two different $y$-axis scalings are used for the weak and strong lines respectively.

### 6.2.1 Creating composite spectra

As in Chapter 5, three composite spectra are considered: one consisting of all 27 absorbers, and two further subsets each consisting of half of the absorbers separated at the median value of the equivalent width of the Ca II $\lambda3935$ line, in this case, $W_{\lambda3935} = 0.68$ Å. As before these three samples are referred to respectively as 'All', 'High-$W_{\lambda3935}$', and 'Low-$W_{\lambda3935}$'.

To create the composite spectra, the individual spectra were shifted to the absorber rest frame without rebinning. Large scale variations in each spectrum were removed by applying a sliding median filter of 45 pixels twice, and the spectra subsequently combined into a composite using an error-weighted arithmetic mean. Alternative methods for both flattening the individual spectra and combining them were explored; in all cases the final results were found to be insensitive to the precise methods used. A second normalisation was carried out on portions of the composite spectra including absorption lines of interest by spline-fitting of the continuum, and the rest frame equivalent widths of the lines were measured using the DIPSO package[1]. Table 6.2 summarises the results, and Fig. 6.1 shows the normalised composite spectrum formed by coadding all 27 Ca II absorbers.

### 6.2.2 Column densities

Inspection of Table 6.2 and Fig. 6.1 shows that the composite spectra cover a variety of transitions from the elements Mg, Ca, Ti, Cr, Mn, Fe and Zn. To avoid potential problems with line saturation, the column density analysis is limited to weak lines most likely to fall on the linear part of the curve of growth.

Ion column densities, $N$, were derived by fitting all the weak lines in the composite spectra simultaneously with the line profile fitting package VPFIT[2] adopting $f$-values and rest wavelengths from the compilation by Morton (2003) and reproduced in Table 6.2. Essentially, for such weak lines, the column densities are related to the equivalent width of the lines via the expression:

$$N = 1.13 \times 10^{20} \cdot \frac{W}{\lambda^2 f} \ \ \mathrm{cm}^{-2} \tag{6.1}$$

where $W$ and $\lambda$ are both in Ångstroms. VPFIT was used primarily to ensure consistency in absorption redshift and width between the different lines. Columns 3-5 of Table 6.3 list ion column densities in each of the three composites together with the errors returned by VPFIT. The last three columns compare each ion to Zn II column density. These are quoted in the standard way, relative to solar abundances from the compilation by Lodders (2003) and reproduced in Table 6.2:

$$[\mathrm{X/Zn}] \equiv \log\left[\mathrm{N(X)/N(Zn)}\right] - \log\left[\mathrm{X/Zn}\right]_{\odot} \tag{6.2}$$

[1] The second normalisation and measuring of the equivalent widths was carried out by Max Pettini.

[2] http://www.ast.cam.ac.uk/∼rfc/vpfit.html

**Table 6.3:** Ion column densities and abundances relative to Zn.

| | | $\log N$ | | | [X/Zn] | | |
|---|---|---|---|---|---|---|---|
| Ion | Solar [a] | All | High-$W_{\lambda 3935}$ | Low-$W_{\lambda 3935}$ | All | High-$W_{\lambda 3935}$ | Low-$W_{\lambda 3935}$ |
| Zn II | 4.63 | 12.82±0.065 | 13.16±0.054 | 12.59±0.110 | 0.00 | 0.00 | 0.00 |
| Cr II | 5.65 | 13.39±0.068 | 13.49±0.093 | 13.42±0.065 | −0.45 | −0.69 | −0.19 |
| Fe II | 7.47 | 15.09±0.045 | 15.02±0.086 | 15.09±0.048 | −0.56 | −0.98 | −0.34 |
| Ti II | 4.92 | 12.48±0.057 | 12.31±0.123 | 12.45±0.069 | −0.63 | −1.14 | −0.43 |
| Ca II | 6.34 | 12.94±0.017 | 13.10±0.019 | 12.86±0.023 | $> -1.59$ [b] | $> -1.77$ | $> -1.45$ |
| Mn II | 5.50 | 13.09±0.023 | 13.19±0.030 | 12.96±0.032 | −0.60 | −0.84 | −0.50 |

[a] Solar abundance relative to hydrogen, in the usual logarithmic scale with H at 12.00 (Lodders 2003).

[b] The values for Ca II are lower limits because a significant but unknown fraction of Ca is doubly ionised and the line ratios of some individual absorbers suggest a degree of saturation.

where $N(X)$ is the column density of the ion in cm$^{-2}$. In all but one case (Ca II), the ions observed are the major ionisation stages of the corresponding elements in H I regions. Assuming that the observed Ca II absorbers have column densities of H I large enough for their gas to be predominantly neutral (as suggested in Section 5.7.3 from their measured dust contents), the ion ratios relative to Zn II can be taken as measures of the elements' abundances relative to Zn.

## 6.3 Results: relative abundances and dust depletions

Without a measurement for $N$(H I) overall metal abundances of the Ca II absorbers cannot be deduced, however element ratios relative to Zn can be studied for which a wealth of data exists for high redshift DLAs. It was chosen to refer the ratios to the Zn II column density because, alone among the elements available here, the depletion level of Zn onto dust grains in the Milky Way ISM is low compared to the refractory Ti, Cr, Mn and Fe which generally exhibit large and variable depletions (Savage & Sembach 1996). In reality, intrinsic differences in the nucleosynthetic history of the elements can lead to departures from solar relative abundances; the dust depletions subsequently operate on this underlying pattern. While this is a rich topic of study (e.g. Wolfe et al. 2005), it is not necessary to differentiate between the two effects in the current study which is primarily concerned with a direct comparison between the Ca II absorbers and DLAs.

To this end, the available DLA literature was searched for systems where the abundances of Zn and at least one of the other elements considered here have been measured. After adjusting the abundances to Lodder's (2003) solar scale, they are compared in Fig. 6.2 with those of the Ca II absorbers. References to the original papers are given in the figure caption. Note that in general these measurements refer to a much wider range of redshifts than that probed here.

Fig. 6.2 shows clearly that the element depletions deduced for the Ca II systems are typical of the values encountered in most DLAs, with Ti, Cr, Mn and Fe typically less abundant than Zn by factors of 3-4 ($\sim 0.5$-0.6 dex). There is a marked distinction between the two subsamples, with the 'High-$W_{\lambda3935}$' subsample in particular exhibiting some of the most pronounced depletions measured in DLAs to date. Most data are available for the Zn and Cr pair (top panel of Figure 6.2), partly for historical reasons (Pettini et al. 1990) and partly because, at the redshifts of most DLAs, the relevant absorption lines are conveniently located in the optical spectrum. The ratio [Cr/Zn] shows an approximate correlation with [Zn/H]: the depletion of Cr decreases with decreasing overall metallicity, although the scatter is considerable (Pettini et al. 1997; Akerman et al. 2005). By using this correlation as a rough indicator of metallicity for the Ca II absorbers, a range between $\sim 1/30$ and $\sim 1/3$ of solar is suggested ([Zn/H] between $-1.5$ and $-0.5$). This point is returned to in Section 6.4.1.

Fig. 6.3 presents the depletions relative to Zn of the Ca II absorbers compared to sight lines through the warm and cold neutral medium of the Milky Way. As is found in DLAs the depletion pattern is similar but the overall level of depletion is lower in the Ca II absorbers. For DLAs this is cited as being consistent with their inferred low metallicities (Vladilo 2004). However, it is of note that the depletions deduced for the 'High-$W_{\lambda3935}$' subsample approach those typical of the warm neutral medium of the Milky Way.

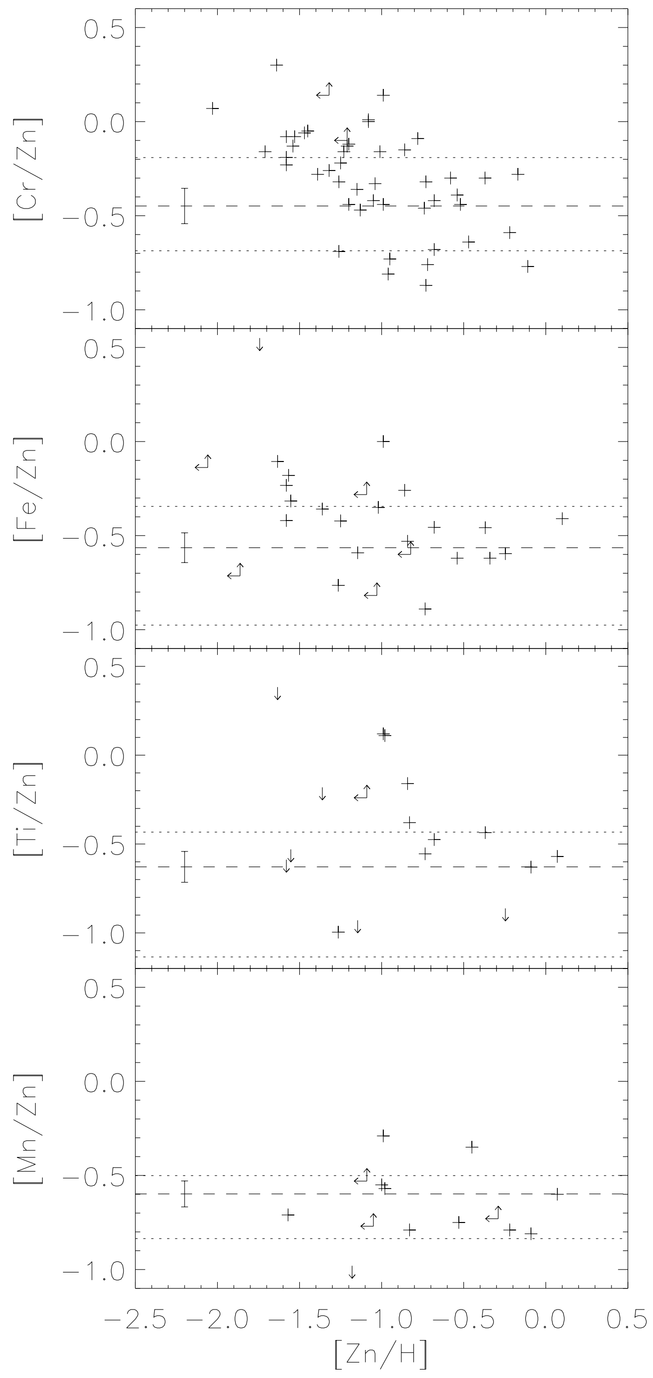

**Figure 6.2:** The crosses show the abundances of refractory elements relative to Zn (an element which shows little affinity for dust) in DLAs from measurements reported in the literature. The horizontal lines are plotted at the values of [X/Zn] in Table 6.3: the dashed lines (with error bars) are the values for the sample of all 27 Ca II absorbers, while the dotted lines are for the two subsamples of 'High-' (lower dotted line) and 'Low- $W_{\lambda3935}$' (upper dotted line). The stronger Ca II absorbers consistently exhibit a higher degree of dust depletion. The sources of the DLA data and corresponding redshift intervals are as follows. Cr: Kulkarni et al. (2005); Akerman et al. (2005) ($0.69 < z_{\rm abs} < 3.39$). Fe: Prochaska et al. (2001); Pettini et al. (1999); Pettini et al. (2000) ($0.61 < z_{\rm abs} < 3.39$). Ti: Ledoux et al. (2002); Prochaska et al. (2001) ($0.43 < z_{\rm abs} < 2.48$). Mn: Ledoux et al. (2002) ($0.43 < z_{\rm abs} < 2.14$). Many of these measurements were obtained from the HIRES DLA database at: http://kingpin.ucsd.edu/∼hiresdla/.

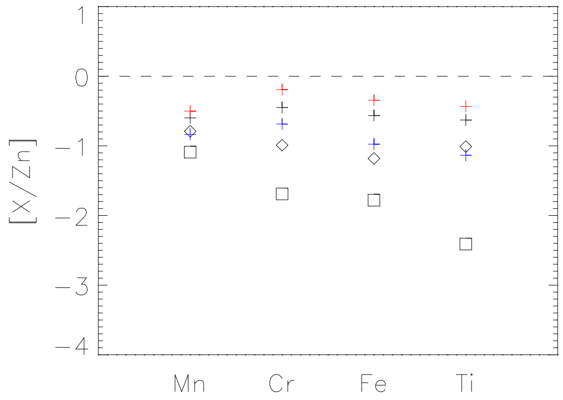

**Figure 6.3:** Abundances relative to Zn of refractory elements in the Ca II absorbers (indicated by crosses), and in the warm (diamonds) and cold (squares) neutral ISM of the Milky Way (from the compilation by Welty et al. 1999). Three sets of crosses are shown, respectively for all Ca II systems (black/middle crosses), and for the 'Low-' (red/upper) and 'High-$W_{\lambda3935}$' (blue/lower) subsamples. Elements are ordered by increasing condensation temperature (Savage & Sembach 1996) and corrected to the solar reference values used in this paper.

## 6.4 Discussion

In this section the metal column densities are combined with the dust measurements of the previous chapter to investigate the dust-to-metals ratio of the Ca II absorbers. The column density of Zn II is used to provide a further constraint on $N$(H I).

### 6.4.1 Dust-to-metals ratio

The analysis of this and the proceeding chapter allows for the first time a direct estimate of the dust-to-metals ratio in absorption-selected high redshift galaxies. This ratio measures the degree by which refractory elements are incorporated into dust grains, which in turn is determined from the balance between dust formation and destruction processes. Generally it is quoted as a ratio of particle number or mass densities, however the ratio $\mathcal{R}_{\rm DM} \equiv \langle E(B-V)\rangle / \langle N(\text{Zn II})\rangle$ provides a practical observational substitute for the purposes of this work. In order to calculate this, $E(B-V)$ values are required specific to the absorbers used in the column density analysis, these are given in Table 6.1; Zn II column densities are taken from Table 6.3. The results are for the observed sample of Ca II absorbers, without any allowance for dust obscuration bias.

For the three Ca II composites, 'All', 'High $W_{\lambda3935}$' and 'Low $W_{\lambda3935}$', values of $\mathcal{R}_{\rm DM} = 8.3^{+1.9}_{-1.8}$, $6.4 \pm 1.2$ and $3.6 \pm 3.7 \times 10^{-15}$ mag cm$^2$ are obtained respectively. The errors have been calculated by propagation of the errors quoted in Tables 6.3 and 6.1 and the LMC colour excesses have been used. Although at first glance there appears to be a discrepancy between the values of the three samples, with the mean value being larger than both the two subsamples, the errors are such that this is not significant. Future SDSS releases may improve the situation, however, for now the measurement errors do not allow any distinction between the 'High-' and 'Low-$W_{\lambda3935}$' samples.

Table 6.4 provides comparison $\mathcal{R}_{\rm DM}$ values for the Milky Way, LMC and SMC. Zn abundance in

**Table 6.4:** Gas-to-dust ratios and metal abundances of the Milky Way, LMC and SMC, along with derived dust-to-metals ratios.

| | MW | LMC | SMC |
|---|---|---|---|
| $\langle N(\text{H I})\rangle/\langle E(B{-}V)\rangle$<br>$(10^{22}\ \text{cm}^{-2}\ \text{mag}^{-1})$ | 0.493[a] | 2.0[b] | 7.9[c] |
| (Fe/H)<br>$(10^{-5})$ | 2.95[d] | 1.12[e] | 0.55[f] |
| $\langle E(B{-}V)\rangle/\langle N(\text{Zn II})\rangle$<br>$(10^{-15}\ \text{mag cm}^{2})$ | 4.7 | 3.1 | 1.6 |

[a] Diplas & Savage (1994)
[b] Koornneef (1982)
[c] Fitzpatrick & Massa (1990)
[d] Lodders (2003)
[e] B-type stars, Korn et al. (2000)
[f] A-type stars, Venn (1999)

the ISM of the LMC and SMC has not been well studied, however the abundance of Fe is expected to be similar to that of Fe-peak elements such as Zn making the observationally better quantified Fe abundance a suitable alternative. The measurement of [Zn/H]$= -0.64^{+0.13}_{-0.17}$ for a single sightline in the SMC by Welty et al. (1997) compares favourably with the Fe abundance derived from stellar spectra of [Fe/H]$= -0.73 \pm 0.07$ (Venn 1999), particularly when allowance is made for the expected small level of depletion of Zn onto dust grains ($\sim$0.1-0.2dex in the Milky Way). Thus, $\mathcal{R}_{\rm DM}$ values are calculated from

$$\frac{\langle E(B{-}V)\rangle}{\langle N(\text{Zn II})\rangle} = \left[\frac{\langle N(\text{H I})\rangle}{\langle E(B{-}V)\rangle} \times \frac{\text{Fe}}{\text{H}} \times \left(\frac{\text{Zn}}{\text{Fe}}\right)_{\odot}\right]^{-1} \qquad (6.3)$$

with each relevant quantity given in Table 6.4 and assuming a solar (Zn/Fe) ratio throughout. Few references to the dust-to-metals ratios of the SMC and LMC can be found in the literature for comparison; the lower values are in qualitative agreement with those quoted in Issa et al. (1990).

From Table 6.4 and $\mathcal{R}_{\rm DM} = 8.3^{+1.9}_{-1.8}$ for the 27 Ca II absorbers it can be seen that the Ca II absorbers have a dust-to-metals ratio similar to, or possibly slightly higher than ($< 3\sigma$) the Milky Way. This result is interesting, given the general belief that the dust-to-metals ratio of high redshift DLAs is lower than the Milky Way. Of course, a direct measurement has not been obtained for these systems and a comparison between the level of depletion of elements seen in the DLAs and the Milky Way ISM is relied upon. Pettini et al. (1997) find that refractory elements are only half depleted onto dust grains, compared to $\gtrsim 90\%$ in the Milky Way ISM, and suggest this is due to a dust-to-metals ratio 1/2 that of the Milky Way. The ratios of Cr, Fe and Ti to Zn[3] in the Ca II absorbers are typically 1/3 solar and, following the same arguments, this would suggest a dust-to-metals ratio 2/3 that of the Milky Way. While the apparent discrepancy between the two methods is intriguing, with implications for studies of dust content in DLAs and physical processes involved in dust grain destruction, the errors remain large

[3] Mn is also found to be underabundant, but this could be due to combined dust depletion and intrinsic abundance effects (Ledoux et al. 2002).

enough that further observations are required to clarify the issue.

Theoretical models for the evolution of the ISM of galaxies, in which dust grain destruction timescales evolve in a similar way to the accretion processes of metals onto grains predict constant dust-to-metals ratios with time (Dwek 1998). Detailed comparisons with such models is beyond the scope of this thesis, but in the future accurate estimates of the dust-to-metals ratio in high redshift galaxies will provide them with important constraints.

### 6.4.2 Constraints on $N$(H I) from Zn II

Similarly to the previous chapter where a constraint was placed on $N$(H I) from dust-to-gas ratios observed in the local Universe, lower limits to the values of $N$(H I) can be derived from the Zn II column densities of the Ca II absorbers by taking the solar abundance of Zn. From the values in Table 6.3 we obtain log[$N$(H I)] $\geq 20.19$, 20.53 and 19.96 for the entire sample, and ‘High-’ and ‘Low-$W_{\lambda3935}$’ sub-samples respectively. Thus, even without knowledge of their metallicity, it can be seen immediately that the strongest and, as shown in Section 5.6, dustiest Ca II systems are very likely to be damped Lyman-$\alpha$ systems. Even the lower equivalent width sample is on average only within a factor of two of the canonical DLA limit log[$N$(H I)] $= 20.3$.

The assumption of solar metallicity is likely to be overly conservative in the present context. The abundance of Zn has been measured in many DLAs at $z_{\rm abs} \simeq 1$ and found to be generally sub-solar. The recent compilation by Kulkarni et al. (2005) shows a mean and median [Zn/H]$\simeq -1.0$ at these intermediate redshifts. If the Ca II absorbers were randomly chosen from known DLA samples, then their corresponding values of log[$N$(H I)] would be ten times higher than those given above, placing them firmly at the upper end of the distribution of column densities of DLAs.

## 6.5 Summary and Conclusions: reddening and dust depletions

Chapters 5 and 6 have presented measurements of dust depletions and reddening in a sample of 37 (27) Ca II absorption line systems with $z_{\rm abs} \sim 1$. Significant reddening due to dust is found, in contrast to the majority of Mg II-selected DLAs at this redshift and DLAs at higher redshift. However, the 20% of Mg II-selected DLAs with the largest equivalent widths of strong metal lines also exhibit a similar degree of reddening assuming the 57% of sub-DLAs in the sample contribute negligible reddening.

*Are they DLAs?*

There are several lines of argument that suggest the observed Ca II absorbers are a sub-population of DLAs. The strengths of the strong Mg II and Fe II absorption lines in all but two of the Ca II absorbers fulfil the DLA selection criteria of Rao (2005) and the depletion ratios of the Ca II absorbers cover a similar range to those of higher redshift DLAs. With the assumptions of no depletion of Zn II onto dust grains and solar metallicity (presumably conservative for $z \sim 1$ systems) lower $N$(H I) limits are obtained of $\langle$log[$N$(H I)] $\rangle > 20.19$, only 0.11dex below the DLA limit. For the systems with the strongest Ca II absorption, and largest dust content as measured by their colour excesses, $\langle$log[$N$(H I)] $\rangle > 20.53$ is found, already above the DLA limit. A second, more stringent limit is placed by assuming the systems have a gas-to-dust ratio similar to the Milky Way, again conservative for their expected lower metallicities, giving $\langle \log[N(\text{H I})] \rangle > 20.43$ for the entire sample of 37 absorbers.

The number density of Ca II absorbers with $W_{\lambda3935} > 0.5$Å is around 20-30% that of DLAs at similar redshifts, after accounting for the systems missing from the sample due to obscuration of the background quasar by dust associated with the absorbers.

### *The nature of* Ca II*-selected DLAs*

Possible interpretations of the results presented in these chapters are that Ca II systems trace the subset of DLAs with:

1. The largest values of neutral hydrogen column density, allowing the detection of significant dust column densities not found in DLAs at the same redshifts.

2. The highest metallicities. If the dust-to-metals ratio is metallicity dependent as suggested by (Vladilo 2004) this could explain the high values found in the absorbers.

3. The largest volume densities, $n_{\text{H}}$ (explained below).

4. A combination of the above.

The third point arises because Ca II is a minor ionisation state of Ca at the temperatures expected in the ISM of the galaxies. The fraction of Ca II therefore grows in proportion to $n_{\text{H}}^2$: two clouds with the same total column density and same Ca abundance, but differing volume densities will have different column densities of Ca II, due to the higher recombination coefficient for Ca II in the denser cloud (Hobbs 1974). This is in accord with the tendency of most Ca II systems to have stronger Mg I $\lambda2853$ absorption for a given Mg II $\lambda2796$ equivalent width (see Fig. 5.3)—Mg I is also a minor ionisation stage. Unfortunately it is difficult to assess the relative importance of these three possibilities without additional information.

The Zn II column densities and *E(B−V)* values provide a direct measure of the dust-to-metals ratio for the first time in absorption selected galaxies: a ratio is found similar to, or slightly larger than, that of the Milky Way. However, this result is in apparent contrast with that inferred from the measured [Cr/Zn] ratio of 2/3 that of the Milky Way ISM, although at present the errors are too large to be certain of a discrepancy. This high dust-to-metal ratio may argue for high metallicities (item [ii] above), if the proposal of Vladilo (2004) is correct that dust-to-metals ratios increase with galactic chemical evolution and therefore metallicity.

The alternative hypothesis that Ca II absorption with $W_{\lambda3935} > 0.5$Å selects the highest column density DLAs suggests metallicities of [Zn/H]$= -1.07$, assuming the column density distribution for high redshift DLAs found in the SDSS DR3 to hold at $z \sim 1$ (Prochaska et al. 2005). This is in agreement with the metallicity predicted from the known trend of [Cr/Zn] with [Zn/H] in DLAs (Fig 6.2). These apparently conflicting results are intriguing, with important implications for the nature of DLAs and numerical models of evolution of galaxy ISM, however untangling them will have to await further observations. One important implication for the hypothesis that these systems have the highest column densities of all DLAs, is the effect of dust obscuration bias on the estimate of the total neutral gas density of the high redshift Universe, which is dominated by the highest column density systems (e.g. Prochaska et al. 2005) despite their insignificant number density compared to lower column density systems.

### *Trend of dust content with velocity dispersion*

A particularly interesting result of the analysis is the trend of increasing dust content with equivalent width of the strong lines, in both the Ca II absorbers and Mg II-selected DLAs. These lines are presumably saturated, and the equivalent widths of saturated lines primarily reflects the velocity dispersion of the system. This therefore suggests that the dustiest systems are either more massive or undergoing some form of disturbance. The former may tie in with the picture of Ca II absorbers being the most chemically evolved subset of DLAs. The latter has been discussed in the literature for DLAs showing strong Ca II absorption (Bowen 1991), but the number of systems was too small to allow firm conclusions. Imaging of the absorbers is required to distinguish these hypotheses; at $z_{\rm abs} \lesssim 1$ this is considerably easier than for DLAs selected traditionally through the Lyman-$\alpha$ line at $z \gtrsim 1.8$.

### *Outlook*

It is of course unfortunate that a lack of observing facilities precludes the measurement of hydrogen column densities for these particular systems, however, plenty of DLAs are known in which Ca II absorption would be detectable either in the red part of the optical spectrum or near infrared. Such measurements would tell us for certain the relation between $N$(H I), Ca II equivalent width and dust content, providing firm metallicities and gas-to-dust ratios which are lacking from the current analysis.

Not only do Ca II absorbers potentially present an important method for substantially improving the numbers of known DLAs at $z \lesssim 1.3$, the results presented here have led to many intriguing questions. It is certainly possible that the Ca II absorbers will turn out to be the key to understanding how different classes of high–redshift galaxies fit into a single unified picture.

*Enjoy your achievements as well as your plans*
*Keep interested in your own career, however humble*
*It is a real possession in the changing fortunes of time*

From Desiderata, Old St Paul's Church, Baltimore

# 7

# Further work

## 7.1 The nature of Ca II absorbers

The Ca II absorption line systems found in the SDSS quasar spectra have been shown in this thesis to have some particularly interesting properties. They contain significant quantities of dust, unlike the average DLA at similar and higher redshifts. Many weak metal transition lines are detected in the composite absorption line spectrum including Zn II, a volatile element undepleted onto dust grains. The column densities of dust and Zn II strongly suggest $N$(H I) values greater than the nominal DLA limit, and the number density per unit redshift of Ca II absorbers implies they represent ∼20-30% of all intermediate redshift DLAs.

Were Ca II absorption line systems proved to be DLAs, these results would have two main implications. Firstly, Ca II–selected DLAs would contribute enormously to the small number of DLAs currently known with $z_{\rm abs} < 1.3$, where the Lyman-$\alpha$ line is inaccessible without a UV-spectrograph. It is believed that DLAs account for the majority of cold, neutral gas in the Universe over a wide range in redshift, thus they are considered particularly important for star formation at high redshifts. Over the epoch probed by the Ca II–selected DLAs the star formation rate of the Universe is declining rapidly, making this redshift range of particular interest for directly observing the stellar populations and ISM of DLAs. The second implication for Ca II absorption line systems being a subsample of DLAs is that they are the first to show considerable quantities of dust. With average *E(B−V)*$\gtrsim$0.1 mag their dust contents approach that of the chemically evolved, star-forming UV–selected galaxies at z∼2-3 with median *E(B−V)*∼0.15-0.2. The correlation of dust content with equivalent width of Ca II possibly implies that the dustiest systems are hosted by more massive galaxies, or perhaps they are part of mergers. Detection of Ca II may identify the most chemically evolved systems, bridging the gap between DLAs and emission–selected galaxies.

However, if the Ca II absorbers are *not* any more chemically evolved than average DLAs their Zn II column densities imply $N$(H I) values at the top of the range for known DLAs. This would have

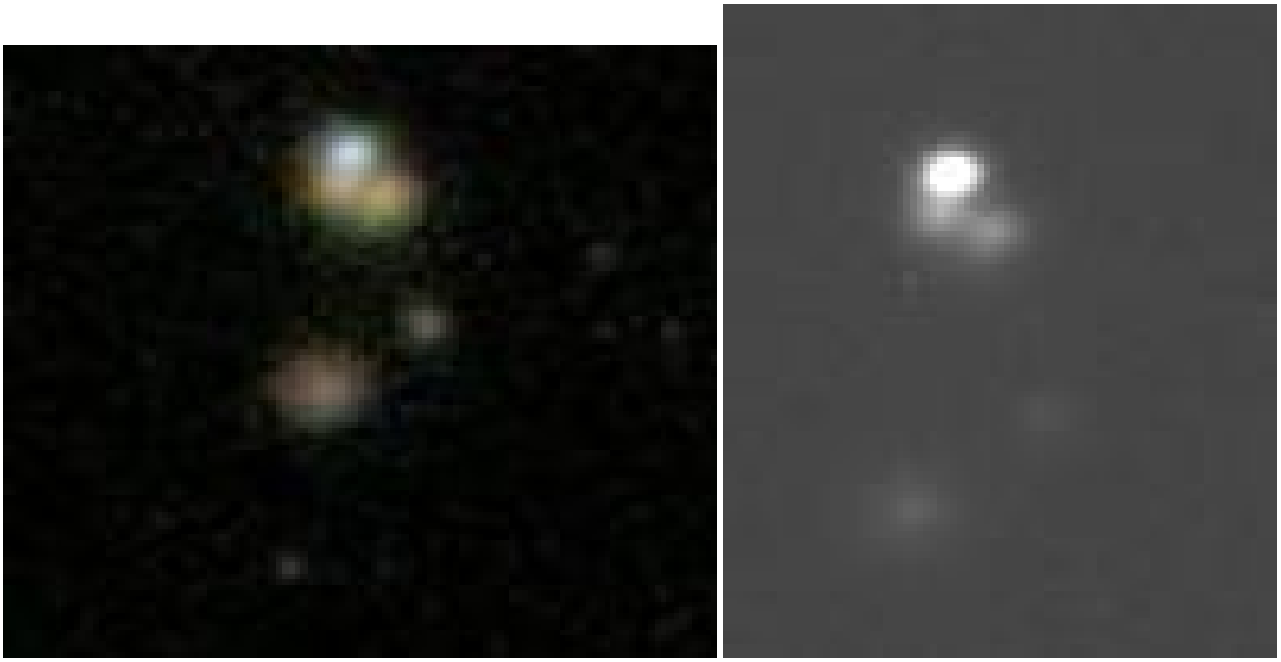

**Figure 7.1:** SDSS (left) and INT $R$-band (right) image of a quasar with a $z$=0.106 Ca II absorption line system identified in its spectrum (spSpec-52669-1159-206). The SDSS image is about 40” across. The quasar is the white object at the top and clear candidates for the host galaxy are seen, with a possible merger occurring. The quasar has dimmed by 0.5-0.6 mag since the SDSS image was obtained. INT image courtesy of Michael Murphy and Arfon Smith.

important implications for the total density of neutral gas at z$\sim$1, as current measurements may then be significantly effected by dust obscuration bias.

Extending the Ca II catalogue to lower redshifts is a simple task; once more has been learnt about the physical nature of the systems this is an obvious next step. The five-band SDSS imaging has already provided candidate host galaxies for some low-redshift systems via photometric redshifts and an extensive joint imaging and spectroscopy campaign of hosts over a large redshift range could be envisaged in the future.

### 7.1.1 Follow up observations

There are several aspects of the nature of the Ca II absorption line systems that can be immediately addressed through follow up observations. The first is the Ca II equivalent widths of known DLAs; Ca II has so far been mostly neglected in metal line absorption studies and any measurements of both $N$(H I) and Ca II absorption would answer the question of chemically evolved versus high H I column density. For high-redshift absorbers Ca II falls in the near IR, accessible to instruments such as ISAAC on the VLT. At intermediate redshifts the number of confirmed DLAs with measured $N$(H I) is fewer, but Ca II falls in the optical waveband where spectroscopy is simpler. ISIS on the WHT will allow concurrent observations in wavebands covering both the Zn II and Ca II doublets for DLAs at $0.7 < z_{\rm abs} < 1.2$. The new catalogue by (Rao et al. 2005) contains 26 such systems and given the measured $n(z)$ of Ca II absorbers (uncorrected for dust extinction bias), 2-3 of these would be expected to show Ca II absorption with equivalent widths >0.5Å. Below this threshold the SDSS Ca II sample is highly incomplete, however, it would be surprising not to find a considerable number of DLAs with Ca II absorption below this limit.

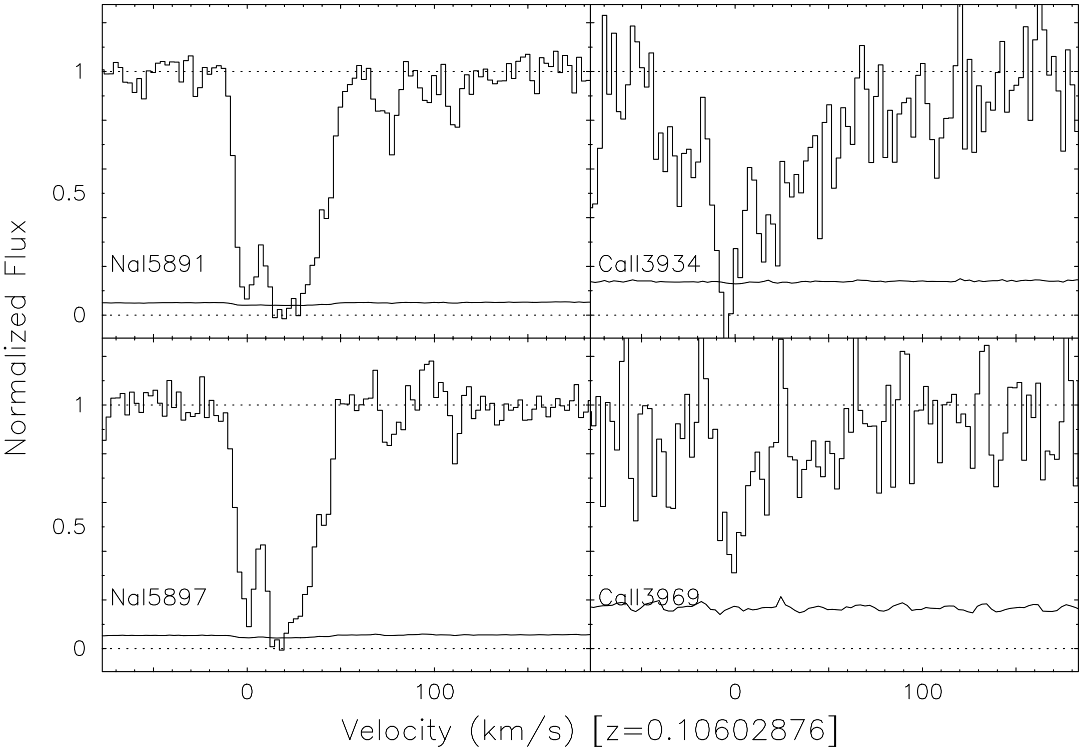

**Figure 7.2:** High resolution Keck HIRES spectrum in the region of the Ca II and Na I lines of the $z = 0.106$ absorber in SDSS quasar spSpec-52669-1159-206. Figure courtesy of Michael Murphy and Neil Crighton.

Any information on the stellar populations of the host galaxies of high-redshift DLAs has proved extremely difficult to obtain (Weatherley et al. 2005). At $z \lesssim 1$ imaging and spectroscopy has successfully identified host galaxies of around 14 systems, finding them to cover a wide range of morphologies (Rao et al. 2003; Wolfe et al. 2005). Deep imaging of the regions around all 36 SDSS quasars in which the Ca II absorption systems have been identified will contribute significantly to the current number of known hosts and answer the question of whether or not Ca II absorbers are hosted by more massive, chemically evolved galaxies than the average DLA.

### 7.1.2 Diffuse interstellar bands

Diffuse interstellar bands (DIBs) are thought to originate from transitions by complex molecules in the interstellar media of galaxies. Their strength correlates with colour excess in the Milky Way and as such their detection would indirectly imply the presence of dust in DLAs were they to be detected. Very few DIBs have been detected in extragalactic objects, only one has been detected due to a DLA, in the spectrum of a highly reddened BL Lac object also with a strong 2175Å feature (Junkkarinen et al. 2004). With their known high dust contents at $z \sim 1$ Ca II absorbers are strong candidates for the detection of DIBs in galaxies over a wide range in redshift.

One attempt has recently been made at obtaining a high SNR spectrum of a Ca II absorption line system detected in the SDSS quasar catalogue at $z = 0.1$. This system shows strong Ca II absorption ($W_{\lambda3935} = 0.82$, $W_{\lambda3970} = 0.45$), together with the Na I $\lambda\lambda5892, 5898$ doublet. The SDSS image of this system is shown in Fig. 7.1 with two clear candidates for the host galaxy, close together and possibly merging. Four 20 minute exposures were obtained with HIRES on Keck in 0.5” seeing on June 1st of this year, however the quasar appeared to have dimmed considerably and a high enough SNR was not achieved to identify DIBs. Fig. 7.2 shows the regions of spectra around the Ca II and Na I lines of the absorber. A follow up five minute $R$-band image was obtained with the INT, confirming that the quasar had dimmed by about 0.5-0.6 mag since the SDSS image was obtained.

## 7.2 Emission line galaxy survey

SDSS quasar spectra allow the detection of intervening galaxies through absorption by the galaxy’s ISM, but galaxies over a wide range in redshift can also be detected in SDSS spectra serendipitously via their emission lines. This technique has already provided a sample of spectroscopically selected strong gravitational lens candidates (Bolton et al. 2004) and this section looks at a potential new use for these objects – in a blind emission line survey.

Most spectroscopic galaxy samples are based on magnitude limited photometric catalogues meaning low surface brightness systems are under-represented. Were these systems to be undergoing significant star formation this bias of spectroscopic samples would result in an underestimate of the total star formation rate (SFR) in the Universe. The following subsections give some background into the use of emission lines as star formation rate indicators and previous emission line surveys. Some preliminary results are then presented to conclude the chapter.

### 7.2.1 Star formation rates

The origin and rate of the decline in star formation from $z \sim 1$ to the present day is currently one of the key unanswered questions in modern cosmology. Recent observational results have suggested this is not due to a decrease in major merger rate (Blanton et al. 2003; Kauffmann et al. 2004; Brinchmann et al. 2004) and theoretical simulations have suggested either a decline in minor merger rates or declining gas supply could be the dominate process (Cole et al. 2000; Somerville et al. 2001b).

There are several methods for estimating star formation rates in galaxies. UV light is a good estimator of the total light output from young stars but is compromised as a SFR indicator by the uncertainty in the effects of dust (Treyer et al. 1998). The UV luminosity can be combined with the thermal IR luminosity of a galaxy, which is dominated by the reprocessed UV light of the stars by enshrouding dust clouds, to give a more robust estimate (Bell et al. 2005). Nebula emission lines, resulting from the re-emission of stellar continuum light, are also good indicators of the instantaneous SFR of galaxies, with H$\alpha$ being the most common indicator due to its small attenuation by dust compared to UV wavelengths; [O II] may also be used, and is useful at higher redshifts, although it suffers from metallicity and reddening dependent effects (Kewley et al. 2004).

In this analysis the H$\alpha$ line is used to estimate SFR via the conversion factor of Kennicutt (1998):

$$\mathrm{SFR}(\mathrm{M}_\odot\,\mathrm{yr}^{-1}) = 7.9 \times 10^{-42} \mathrm{L}(\mathrm{H}\alpha) \tag{7.1}$$

where L(H$\alpha$) has been corrected for dust extinction and stellar absorption and is given in $\mathrm{erg\,s^{-1}}$. An important question for estimating SFRs at higher redshifts is the reliability of [O II] as a SFR indicator and this is investigated in the final subsection of this chapter; Kewley et al. (2004) advocate the following conversion for [O II]:

$$\mathrm{SFR}(\mathrm{M}_{\odot}\,\mathrm{yr}^{-1}) = 6.58 \times 10^{-42}\mathrm{L}([\mathrm{OII}]) \tag{7.2}$$

where L([O II]) has again been corrected for reddening at the wavelength of [O II]. This is derived from a comparison between H$\alpha$ and [O II] line luminosities in 97 nearby galaxies, assuming the Cardelli et al. (1989) reddening extinction law.

### 7.2.2 Previous and ongoing surveys

Several emission line surveys have been carried out to address the possibility of new line emitting populations of galaxies undetected by other techniques. Many of these have relied upon comparing broad and narrow band imaging, more recently Fabry-Perot interferometers have been used to scan the sky in a narrow wavelength range, allowing the unambiguous detection of single lines (Jones & Bland-Hawthorn 2001; Glazebrook et al. 2004; Hippelein et al. 2003). The latter two of these surveys have concluded that broad band surveys do not miss a substantial fraction of the cosmic star formation rate density and a population of weak continuum line-emitting objects does not exist.

Recently Drozdovsky et al. (2005) presented a preliminary analysis of a sample of 601 emission line objects with $z \leq 1.6$ detected with the HST/ACS Grism. This survey is the deepest so far, however, as with previous surveys, the low resolution means the H$\beta$+[O III] and H$\alpha$+N II lines are blended, and follow up spectroscopy is required to determine the nature of the emission (e.g. AGN versus star forming). A proportion of the objects have uncertain redshift identifications due to only single lines being detected. No analysis of the inferred cosmic star formation rate density from the survey has been presented.

### 7.2.3 The SDSS blind emission line survey

When an SDSS fibre is positioned on a known source there is some probability that light from another galaxy, entirely unassociated with that primary galaxy, will fall down the fibre. At low redshift, these secondary systems are likely to be low surface brightness systems with weak continua and strong lines not detected in SDSS broadband photometry. At high redshift they will be massive, star forming galaxies or, potentially, AGN and they may be lenses. In the following they are all termed emission line galaxies (ELGs). The total area of sky covered by fibres in the main SDSS DR4 catalogue is ∼1400 square arcmin, more than a factor of 3 greater than the CADIS emission line survey which has the greatest sky coverage of recent emission line surveys (Hippelein et al. 2003). The limiting depth of the survey is difficult to assess without simulations, due to the dependence of line detection on the SNR of the primary object spectrum. However, it is certainly comparable to the CADIS survey of $3 \times 10^{-17}\mathrm{erg\,s^{-1}cm^{-2}}$, although not as deep as the HST/ACS survey of Drozdovsky et al. (2005) with a median depth of $1.6 \times 10^{-17}\mathrm{erg\,s^{-1}cm^{-2}}$.

Recently Paul Hewett has obtained samples of secondary ELGs in the galaxy, quasar, sky and stellar SDSS DR4 spectroscopic samples, using a matched-filter search similar to that used for the detection

**Table 7.1:** Summary of the number of secondary ELGs with $z < 0.4$ and two confirmed lines identified in the SDSS DR4 spectroscopic samples. The final column is for a control sample of ELGs selected from the main SDSS catalogue.

| | Galaxy | Quasar | Star | Sky | Control |
|---|---|---|---|---|---|
| H$\alpha$+H$\beta$ | 390 | 43 | 13 | 0 | 65 |
| H$\alpha$+[O II] | 410 | 53 | 14 | 0 | 182 |
| H$\alpha$+[O III] | 416 | 62 | 15 | 0 | 84 |
| H$\alpha$+N II | 312 | 50 | 13 | 0 | 70 |
| O II+O III | 253 | 26 | 2 | 0 | 212 |
| total | 575 | 66 | 16 | 3 | 217 |

of the quasar absorption line systems presented in this thesis. A total of 724 candidate objects were obtained with $z_{\rm em} < 0.4$, 66 of which were discarded after fully parameterised line fits were carried out by myself. All of the confirmed systems contain two or more lines detected at $> 3\sigma$ and a velocity difference between the lines of $< 100\,{\rm km\,s^{-1}}$. Table 7.1 summarises the number of objects detected with each line pairing in each dataset. A further set of ELGs with higher redshifts have been identified, but this section concentrates on those with H$\alpha$ emission.

One of the main advantages of this survey over other surveys is that the resolution of the spectra is high enough to clearly distinguish the H$\beta$+[O III] and H$\alpha$+N II lines. Additionally many spectra have both H$\alpha$ and [O II] emission lines, the two primary lines used as SFR diagnostics. Neither of these facts are true of the HST/ACS survey. It was therefore decided to focus at first on those objects with significant H$\alpha$ and H$\beta$ line detections to allow reddening measurements via the Balmer decrement. Accurate SFR estimates can then be obtained from the line luminosities of [O II] and H$\alpha$. Additionally the sample of the present analysis has been restricted to only those secondary objects detected in the SDSS galaxy catalogue as it was found that an accurate continuum for the primary object was required. A PCA was used to reconstruct the galaxy spectra and a little more work is required to apply the same technique to the other primary object samples.

### 7.2.4 Results so far

Fig. 7.3 shows regions of the composite spectra of secondary ELGs with $0.001 < z < 0.1$ and $0.3 < z < 0.35$ found in the galaxy sample. Prior to combining the ELG spectra, the spectra of the primary galaxies were reconstructed with a PCA continuum and subtracted. The residual spectra were then weighted by the squared luminosity distance of the ELG and combined using an error weighted mean. In the low redshift sample the Ca II $\lambda\lambda 3935, 3970$ stellar absorption lines are visible, the order of magnitude difference in line luminosities is clear, as is a difference in line ratios. Fig. 7.4 provides a comparison composite of 155 ELGs selected directly from the SDSS catalogue with $0.001 < z < 0.15$ to have emission lines of similar detection significance to the secondary objects, a sliding median filter has been used to remove the large continuum variations for better comparison with the secondary ELG composite spectra. The main noticeable difference between the emission line properties of the secondary ELGs and control sample is the small N II and large [O III] line strengths.

The composite spectra can be used to observe average trends in the emission line fluxes and ratios

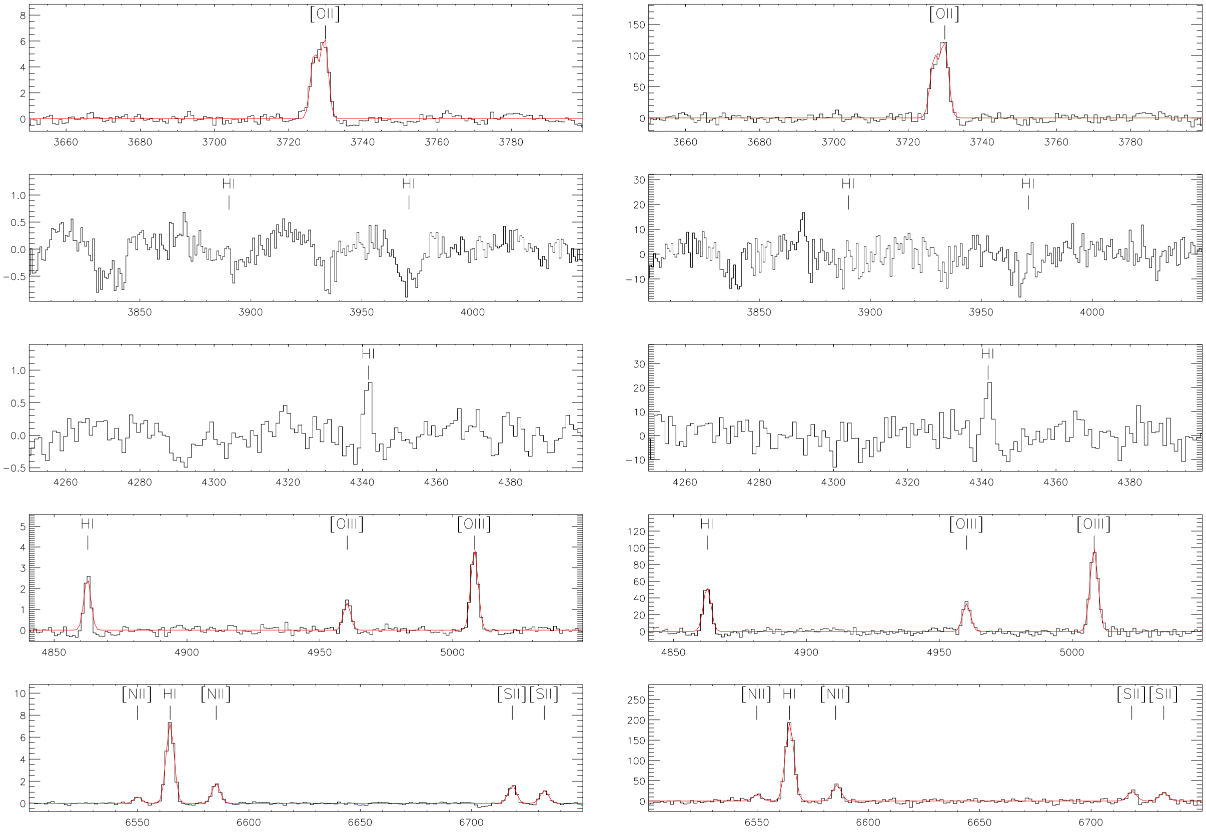

**Figure 7.3:** Composite spectrum of $0.001 < z < 0.05$ (left) and $0.3 < z < 0.35$ (right) ELGs detected in the SDSS spectroscopic galaxy catalogue, after the primary galaxy spectrum has been subtracted using PCA reconstructions. Line fluxes are in $10^{38}\mathrm{erg\,s^{-1}cm^{-2}\AA^{-1}}$ and the fully parameterised model fit is overplotted for the primary emission features.

with redshift, Fig. 7.5 shows the changing mean H$\alpha$ line luminosity, [O II] line luminosity, and H$\alpha$/H$\beta$ ratio as a function of redshift. The average H$\alpha$ line luminosity is also plotted against that of [O II] and compared to the theoretically expected value were both lines to measure the same SFRs. In general the agreement is encouraging although the slight offset needs further investigation. The observed trends of line luminosity with redshift are qualitatively expected, with the higher redshift objects being more luminous and more massive.

The line fluxes of these composites may also be used to classify the objects and measure their metallicities. Fig. 7.6 shows the position of the average ELG spectra on some well known line diagnostic plots. On the left, the figure shows the R$_{23}$ abundance indicator which can be used to estimate metallicity (Kewley & Dopita 2002). On the right, the line ratios can be used to determine that on average the emission line objects are dominated by star formation and not AGN (Baldwin et al. 1981; Brinchmann et al. 2004).

One caveat to these results is that in some of the composite spectra stellar absorption of H$\alpha$ and H$\beta$ is noticeable and this has not been accounted for in the measurement of the line fluxes. It is expected that correcting for this will tighten some of the observed trends.

Alongside improving the results above, by extending to the other samples and correcting for stellar absorption for example, there are many interesting questions that can be addressed with this survey. It is hoped that an estimate of the scatter between the H$\alpha$ and [O II] SFRs will be obtained between

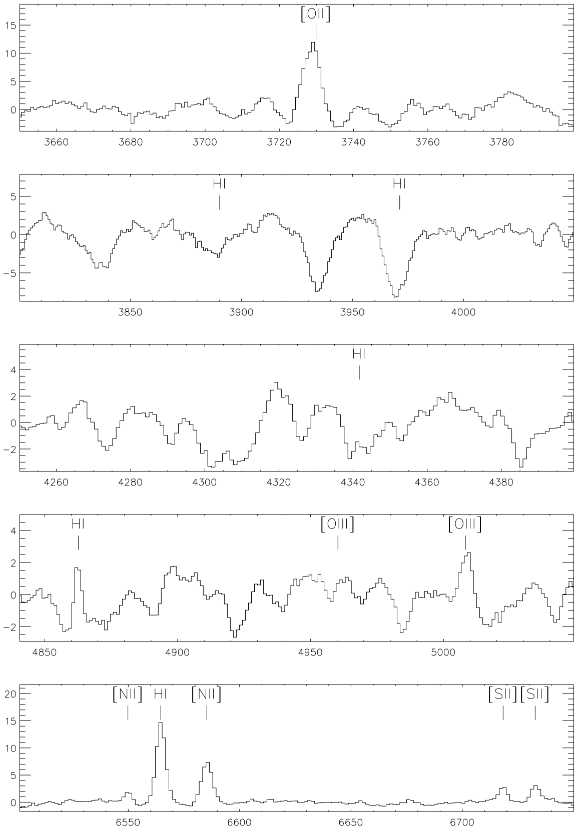

**Figure 7.4:** Composite of 155 ELG spectra in the primary SDSS galaxy catalogue with $0.001 < z < 0.15$.

$0 \lesssim z \lesssim 0.4$, although some effort will be required to improve the continuum estimates for the primary objects. The H$\alpha$/H$\beta$ ratios will be used to investigate empirically the dust attenuation in ELGs. By estimating the fraction of light from the galaxies contributing to the line fluxes, a problem that is simpler to model than aperture bias, an estimate of the SFR density of the ELGs will be obtained and compared to the results from the main SDSS sample (Brinchmann et al. 2004). The large number of objects out to $z \sim 1$ will allow a consistent measure of the SFR density with redshift, in contrast to the narrow redshift ranges observed by most H$\alpha$ and [O II] line surveys.

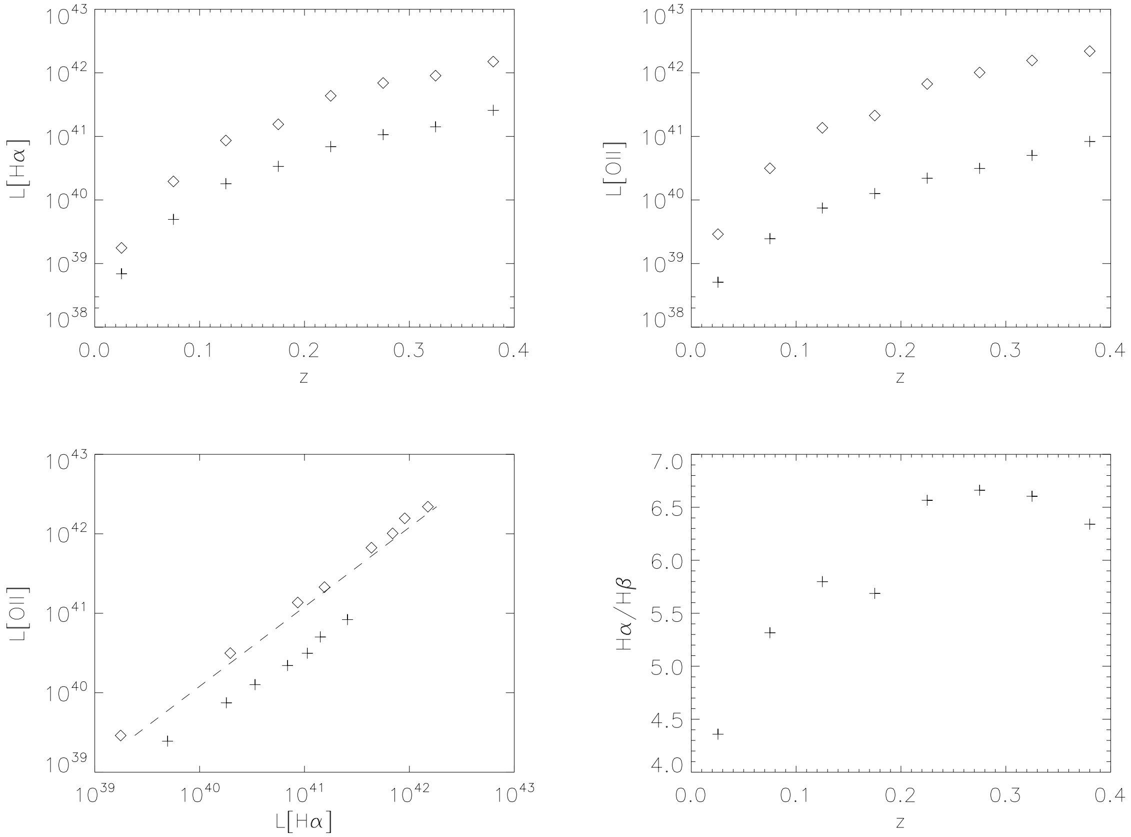

**Figure 7.5:** Average trends of line luminosities in secondary ELGs found in the spectra of SDSS galaxies. The crosses and diamonds show quantities pre- and post-application of dust attenuation corrections respectively.

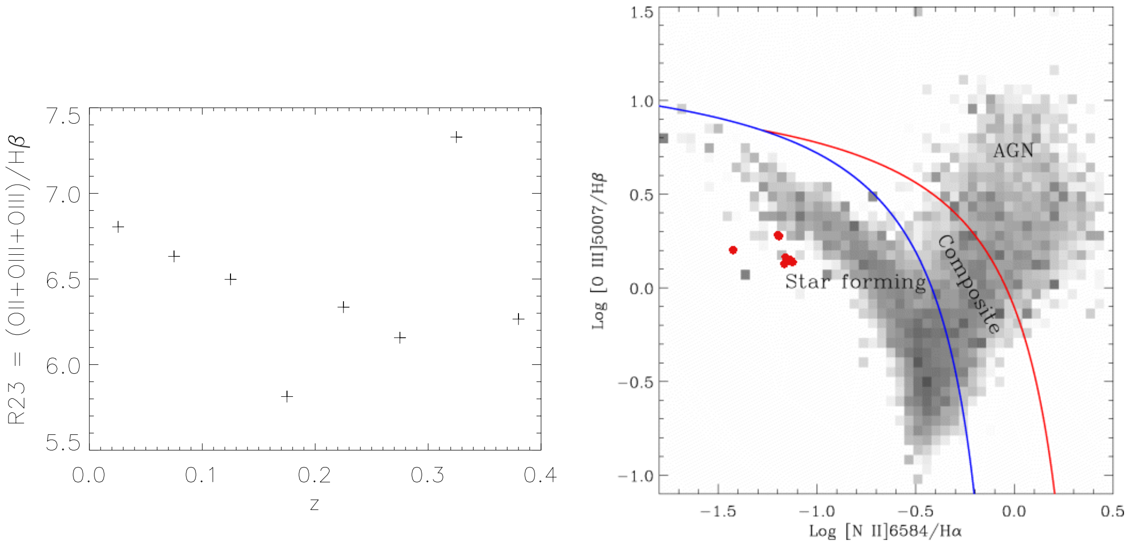

**Figure 7.6:** Left: $R_{23}$, a common metallicity indicator, as a function of redshift for the ELG composite spectra. Right: figure reproduced from Brinchmann et al. (2004) to indicate the position of the ELGs (red dots) in relation to star forming galaxies and AGN in the primary SDSS spectroscopic catalogue.

# A
# Air and vacuum wavelengths

**Table A.1:** Air and vacuum wavelengths of transitions of species commonly seen in nebula emission and/or absorption in galaxy spectra

| Air (Å) | Vacuum (Å) | Name |
|---|---|---|
| 3726.03 | 3727.09 | [O II] |
| 3728.82 | 3729.88 | [O II] |
| 3889.05 | 3890.15 | H I |
| 3970.08 | 3971.20 | H I |
| 4340.46 | 4341.68 | H I |
| 4861.32 | 4862.68 | H I |
| 4958.91 | 4960.29 | [O III] |
| 5006.84 | 5008.24 | [O III] |
| 6548.04 | 6549.85 | [N II] |
| 6562.79 | 6564.61 | H I |
| 6583.46 | 6585.28 | [N II] |
| 6716.43 | 6718.29 | [S II] |
| 6730.81 | 6732.67 | [S II] |

**Table A.2:** Air and vacuum wavelengths of transitions of species commonly seen in absorption in galaxy spectra (hydrogen lines given in previous table).

| Air (Å) | Vacuum (Å) | Name |
|---|---|---|
| 2025.49 | 2026.14 | Zn II |
| 2055.60 | 2056.26 | Cr II |
| 2061.58 | 2062.24 | Cr II |
| 2062.00 | 2062.66 | Zn II |
| 2065.50 | 2066.16 | Cr II |
| 2249.18 | 2249.88 | Fe II |
| 2260.08 | 2260.78 | Fe II |
| 2343.50 | 2344.21 | Fe II |
| 2373.74 | 2374.46 | Fe II |
| 2382.04 | 2382.76 | Fe II |
| 2576.11 | 2576.88 | Mn II |
| 2585.88 | 2586.65 | Fe II |
| 2593.72 | 2594.50 | Mn II |
| 2599.40 | 2600.17 | Fe II |
| 2605.68 | 2606.46 | Mn II |
| 2795.53 | 2796.35 | Mg II |
| 2802.71 | 2803.53 | Mg II |
| 2852.12 | 2852.96 | Mg I |
| 3057.39 | 3058.28 | Ti II |
| 3241.98 | 3242.92 | Ti II |
| 3383.76 | 3384.73 | Ti II |
| 3933.66 | 3934.78 | Ca II |
| 3968.49 | 3969.61 | Ca II |
| 5889.95 | 5891.58 | Na I |
| 5895.92 | 5897.56 | Na I |
| 8498.06 | 8500.40 | Ca II |
| 8542.05 | 8544.40 | Ca II |
| 8662.12 | 8664.50 | Ca II |

# BIBLIOGRAPHY

Abazajian K., et al. (The SDSS Collaboration), 2003, AJ, 126, 2081

Abazajian K., et al. (The SDSS Collaboration), 2004, AJ, 128, 502

Abazajian K., et al. (The SDSS Collaboration), 2005, AJ, 129, 1755

Adams W. S., 1949, ApJ, 109, 354

Adelman-McCarthy J. K., et al. (The SDSS Collaboration), 2005, AJsubmitted

Akerman C. J., Ellison S. L., Pettini M., Steidel C. C., 2005, A&A, in press (astro-ph/0506180)

Baldry I. K., Glazebrook K., Baugh C. M., et al. (The 2dFGRS Team), 2002, ApJ, 569, 582

Baldry I. K., Glazebrook K., et al., 2003, in IAU Symposium The Color-Magnitude Distribution of Galaxies from the SDSS

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

Baugh C. M., Croton D. J., Gaztañaga E., et al. (The 2dFGRS Team), 2004, MNRAS, 351, L44

Bell E. F., Papovich C., Wolf C., et al., 2005, ApJ, 625, 23

Benson A. J., Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 2000, MNRAS, 316, 107

Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38

Bernardeau F., Kofman L., 1995, ApJ, 443, 479

Blanton M., 2000, ApJ, 544, 63

Blanton M., Cen R., Ostriker J. P., Strauss M. A., 1999, ApJ, 522, 590

Blanton M. R., Hogg D. W., Bahcall N. A., et al., 2003, ApJ, 594, 186

Bolton A. S., Burles S., Schlegel D. J., Eisenstein D. J., Brinkmann J., 2004, AJ, 127, 1860

Borys C., Smail I., Chapman S. C., Blain A. W., Alexander D. M., Ivison R. J., 2005, ApJ, in press (astro-ph/0507610)

Bowen D. V., 1991, MNRAS, 251, 649

Bower R. G., Coles P., Frenk C. S., White S. D. M., 1993, ApJ, 405, 403

Boyle B. J., Shanks T., Croom S. M., Smith R. J., Miller L., Loaring N., Heymans C., 2000, MNRAS, 317, 1014

Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151

Broadhurst T. J., Taylor A. N., Peacock J. A., 1995, ApJ, 438, 49

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

Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245

Carilli C. L., Menten K. M., Reid M. J., Rupen M. P., Yun M. S., 1998, ApJ, 494, 175

Cen R., Ostriker J. P., 1992, ApJ, 399, L113

Cen R., Ostriker J. P., 2000, ApJ, 538, 83

Chamberlain J. W., 1961, Physics of the aurora and airglow. International Geophysics Series, New York: Academic Press, 1961

Cole S., Hatton S., Weinberg D. H., Frenk C. S., 1998, MNRAS, 300, 945

Cole S., Kaiser N., 1989, MNRAS, 237, 1127

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

Cole S., Percival W. J., Peacock J. A., et al. (The 2dFGRS team), 2005, MNRAS, p. 681

Coles P., 1993, MNRAS, 262, 1065

Coles P., Jones B., 1991, MNRAS, 248, 1

Colless M., Dalton G., Maddox S., et al. (The 2dFGRS Team), 2001, MNRAS, 328, 1039

Colless M., Peterson B. P., Jackson C., et al. (The 2dFGRS Team), 2003, preprint (astro-ph/0306581)

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

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

Conway E., Maddox S., Wild V., et al. (The 2dFGRS Team), 2005, MNRAS, 356, 456

Crinklaw G., Federman S. R., Joseph C. L., 1994, ApJ, 424, 748

Cross N. J. G., Driver S. P., Liske J., Lemon D. J., Peacock J. A., Cole S., Norberg P., Sutherland W. J., 2004, MNRAS, 349, 576

Croton D. J., Colless M., Gaztanaga E., et al. (The 2dFGRS Team), 2004a, ApJ, 352, 828

Croton D. J., Gaztanaga E., Baugh C. M., et al. (The 2dFGRS Team), 2004b, ApJ, 352, 1232

Davis M., Geller M. J., 1976, ApJ, 208, 13

Dekel A., Lahav O., 1999, ApJ, 520, 24

Dekel A., Ostriker J. P., eds, 1999, Formation of structure in the universe

Diaz A. I., Terlevich E., Terlevich R., 1989, MNRAS, 239, 325

Diplas A., Savage B. D., 1994, ApJ, 427, 274

Dressler A., 1980, ApJ, 236, 351

Dressler A., 1984, ApJ, 286, 97

Drozdovsky I., Yan L., Chen H., Stern D., Kennicutt R., Spinrad H., Dawson S., 2005, AJ in press (astro-ph/0503592)

Dwek E., 1998, ApJ, 501, 643

Efstathiou G., Fall S. M., 1984, MNRAS, 206, 453

Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R. S., Frenk C. S., 1990, MNRAS, 247, 10P

Efstathiou G., Moody S., Peacock J. A., et al. (The 2dFGRS Team), 2002, MNRAS, 330, L29

Elgarøy Ø., Lahav O., 2003, Journal of Cosmology and Astro-Particle Physics, 4, 4

Elgarøy Ø., Lahav O., Percival W. J., et al. (The 2dFGRS Team), 2002, Physical Review Letters, 89, 1301

Ellison S. L., Churchill C. W., Rix S. A., Pettini M., 2004, ApJ, 615, 118

Ellison S. L., Hall P. B., Lira P., 2005, AJ, in press (astro-ph/0507418)

Ellison S. L., Yan L., Hook I. M., Pettini M., Wall J. V., Shaver P., 2001, A&A, 379, 393

Everson R., Sirovich L., 1995, J. Opt. Soc. Am. A, 12, 1657

Fall S. M., Pei Y. C., 1989, ApJ, 337, 7

Fall S. M., Pei Y. C., 1993, ApJ, 402, 479

Fall S. M., Pei Y. C., McMahon R. G., 1989, ApJ, 341, L5

Fan X., Hennawi J. F., Richards G. T., et al., 2004, AJ, 128, 515

Fan Z., 2003, ApJ, 594, 33

Fasano G., Franceschini A., 1987, MNRAS, 225, 155

Fischer P., McKay T. A., Sheldon E., et al., 2000, AJ, 120, 1198

Fitzpatrick E. L., Massa D., 1990, ApJS, 72, 163

Folkes S., Ronen S., Price I., et al. (The 2dFGRS Team), 1999, MNRAS, 308, 459

Folkes S. R., Lahav O., Maddox S. J., 1996, MNRAS, 283, 651

Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., 1992, ApJ, 398, 476

Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447

Gómez P. L., Nichol R. C., Miller C. J., et al., 2003, ApJ, 584, 210

Geller M. J., Huchra J. P., 1989, Science, 246, 897

Glazebrook K., Offer A. R., Deeley K., 1998, ApJ, 492, 98

Glazebrook K., Tober J., Thomson S., Bland-Hawthorn J., Abraham R., 2004, AJ, 128, 2652

Gunn J. E., Gott J. R. I., 1972, ApJ, 176, 1

Hambly N. C., MacGillivray H. T., Read M. A., Tritton S. B., Thomson E. B., Kelly B. D., Morgan D. H., Smith R. E., et al., 2001, MNRAS, 326, 1279

Hamilton A. J. S., 1985, ApJ, 292, 35

Hand D., Mannila H., Smyth P., 2001, Principles of Data Mining. The MIT Press

Hawkins E., Maddox S., Cole S., et al. (The 2dFGRS Team), 2003, MNRAS, 346, 78

Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965

Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., Pearce F. R., 2003, MNRAS, 338, 913

Hermit S., Santiago B. X., Lahav O., Strauss M. A., Davis M., Dressler A., Huchra J. P., 1996, MNRAS, 283, 709

Hewett P. C., Foltz C. B., 2003, AJ, 125, 1784

Hewett P. C., Irwin M. J., Bunclark P., Bridgeland M. T., Kibblewhite E. J., He X. T., Smith M. G., 1985, MNRAS, 213, 971

Hippelein H., Maier C., Meisenheimer K., et al., 2003, A&A, 402, 65

Hobbs L. M., 1974, ApJ, 188, L107

Hoekstra H., van Waerbeke L., Gladders M. D., Mellier Y., Yee H. K. C., 2002, ApJ, 577, 604

Hogg D. W., Blanton M. R., Eisenstein D. J., et al., 2003, ApJ, 585, L5

Hook I. M., Jørgensen I., Allington-Smith J. R., Davies R. L., Metcalfe N., Murowinski R. G., Crampton D., 2004, PASP, 116, 425

Hubble E., Humason M. L., 1931, ApJ, 74, 43

Hubble E. P., 1926, ApJ, 64, 321

Humason M. L., Mayall N. U., Sandage A. R., 1956, AJ, 61, 97

Issa M. R., MacLaren I., Wolfendale A. W., 1990, A&A, 236, 237

Jones D. H., Bland-Hawthorn J., 2001, ApJ, 550, 593

Junkkarinen V. T., Cohen R. D., Beaver E. A., Burbidge E. M., Lyons R. W., Madejski G., 2004, ApJ, 614, 658

Kaiser N., 1987, MNRAS, 227, 1

Kauffmann G., Heckman T. M., White S. D. M., et al., 2003a, MNRAS, 341, 33

Kauffmann G., Heckman T. M., Tremonti C., et al., 2003c, MNRAS, 346, 1055

Kauffmann G., Heckman T. M., White S. D. M., et al., 2003b, MNRAS, 341, 54

Kauffmann G., Nusser A., Steinmetz M., 1997, MNRAS, 286, 795

Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, MNRAS, 353, 713

Kayo I., Taruya A., Suto Y., 2001, ApJ, 561, 22

Kelson D. D., 2003, PASP, 115, 688

Kendall M. G., 1975, Multivariate Analysis. Griffen, London

Kennicutt R. C., 1998, ARA&A, 36, 189

Kewley L. J., Dopita M. A., 2002, ApJS, 142, 35

Kewley L. J., Geller M. J., Jansen R. A., 2004, AJ, 127, 2002

Kofman L., Bertschinger E., Gelb J. M., Nusser A., Dekel A., 1994, ApJ, 420, 44

Koornneef J., 1982, A&A, 107, 247

Korn A. J., Becker S. R., Gummersbach C. A., Wolf B., 2000, A&A, 353, 655

Kulkarni V. P., Fall S. M., Lauroesch J. T., York D. G., Welty D. E., Khare P., Truran J. W., 2005, ApJ, 618, 68

Kurtz M. J., Mink D. J., 2000, ApJ, 533, L183

Lahav O., Bridle S. L., Percival W. J., et al. (The 2dFGRS Team), 2002, MNRAS, 333, 961

Lahav O., Nemiroff R. J., Piran T., 1990, ApJ, 350, 119

Ledoux C., Bergeron J., Petitjean P., 2002, A&A, 385, 802

Lewis I., Balogh M., De Propris R., et al. (The 2dFGRS Team), 2002, MNRAS, 334, 673

Liddle A. R., 2004, MNRAS, 351, L49

Lissandrini C., Cristiani S., La Franca F., 1994, PASP, 106, 1157

Lodders K., 2003, ApJ, 591, 1220

Lupton R., 1993, Statistics in theory and practice. Princeton, N.J.: Princeton University Press

Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990, MNRAS, 243, 692

Madgwick D. S., 2003, MNRAS, 338, 197

Madgwick D. S., Coil A. L., Conselice C. J., et al., 2003c, ApJ, 599, 997

Madgwick D. S., Hawkins E., Lahav O., et al. (The 2dFGRS Team), 2003b, MNRAS, 344, 847

Madgwick D. S., Lahav O., Baldry I. K., et al. (The 2dFGRS team), 2002, MNRAS, 333, 133

Madgwick D. S., Somerville R., Lahav O., Ellis R., 2003a, MNRAS, 343, 871

Maino D., Farusi A., Baccigalupi C., Perrotta F., Banday A. J., Bedini L., Burigana C., De Zotti G., Górski K. M., Salerno E., 2002, MNRAS, 334, 53

Marschall L. A., Hobbs L. M., 1972, ApJ, 173, 43

Matsubara T., 1999, ApJ, 525, 543

Morton D. C., 2003, ApJS, 149, 205

Murphy M. T., Liske J., 2004, MNRAS, 354, L31

Murtagh F., Heck A., 1987, Multivariate Data Analysis. Astrophysics and Space Science Library, Dordrecht: Reidel

Narayanan V. K., Berlind A. A., Weinberg D. H., 2000, ApJ, 528, 1

Nestor D. B., Turnshek D. A., Rao S. M., 2005, ApJ, 628, 637

Norberg P., Baugh C. M., Hawkins E., et al. (The 2dFGRS Team), 2001, MNRAS, 328, 64

Norberg P., Cole S., Baugh C. M., et al. (The 2dFGRS Team), 2002a, MNRAS, 336, 907

Norberg P., Cole S., Baugh C. M., et al. (The 2dFGRS Team), 2002b, MNRAS, 332, 827

O'Donnell J. E., 1994, ApJ, 422, 158

Ostriker J. P., Heisler J., 1984, ApJ, 278, 1

Panter B., Heavens A. F., Jimenez R., 2003, MNRAS, 343, 1145

Papovich C., Dickinson M., Ferguson H. C., 2001, ApJ, 559, 620

Parry I. R., Carrasco E., 1990, in Proc. SPIE Vol. 1235, Instrumentation in astronomy VII Deep fibre spectroscopy. pp 702–708

Peacock J. A., 1983, MNRAS, 202, 615

Peacock J. A., 2003, in American Institute of Physics Conference Series Vol. 666. p. 275

Peacock J. A., Cole S., Norberg P., et al. (The 2dFGRS Team), 2001, Nature, 410, 169

Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton, N.J., Princeton University Press, 1980. 435 p.

Pei Y. C., 1992, ApJ, 395, 130

Pei Y. C., Fall S. M., Bechtold J., 1991, ApJ, 378, 6

Pen U., 1998, ApJ, 504, 601

Pen U., Lu T., van Waerbeke L., Mellier Y., 2003, MNRAS, 346, 994

Percival W. J., Baugh C. M., Bland-Hawthorn J., et al. (The 2dFGRS Team), 2001, MNRAS, 327, 1297

Pettini M., Boksenberg A., Hunstead R. W., 1990, ApJ, 348, 48

Pettini M., Ellison S. L., Steidel C. C., Bowen D. V., 1999, ApJ, 510, 576

Pettini M., Ellison S. L., Steidel C. C., Shapley A. E., Bowen D. V., 2000, ApJ, 532, 65

Pettini M., King D. L., Smith L. J., Hunstead R. W., 1997, ApJ, 478, 536

Pettini M., Smith L. J., Hunstead R. W., King D. L., 1994, ApJ, 426, 79

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical recipes in C. The art of scientific computing. Cambridge: University Press, 2nd ed.

Prochaska J. X., Herbert-Fort S., 2004, PASP, 116, 622

Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, ApJ, in press (astro-ph/0508361)

Prochaska J. X., Wolfe A. M., Tytler D., Burles S., Cooke J., Gawiser E., Kirkman D., O'Meara J. M., Storrie-Lombardi L., 2001, ApJS, 137, 21

Rao S. M., 2005, IAU Colloquium 199 (astro-ph/0505479)

Rao S. M., Nestor D. B., Turnshek D. A., Lane W. M., Monier E. M., Bergeron J., 2003, ApJ, 595, 94

Rao S. M., Turnshek D. A., 2000, ApJS, 130, 1

Rao S. M., Turnshek D. A., Nestor D. B., 2005, ApJ submitted

Reichard T. A., Richards G. T., Hall P. B., et al., 2003b, AJ, 126, 2594

Reichard T. A., Richards G. T., Schneider D. P., et al., 2003a, AJ, 125, 1711

Richards G. T., Hall P. B., vanden Berk D. E., et al., 2003, AJ, 126, 1131

Roweis S., 1997, Neural Information Processing Systems, 10, 626

Rutledge G. A., Hesser J. E., Stetson P. B., Mateo M., Simard L., Bolte M., Friel E. D., Copin Y., 1997, PASP, 109, 883

Savage B. D., Sembach K. R., 1996, ARA&A, 34, 279

Scherrer R. J., Weinberg D. H., 1998, ApJ, 504, 607

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

Schneider D. P., Hall P. B., Richards G. T., et al., 2005, AJ, 130, 367

Schneider D. P., Hall P. B., Richards G. T., et al. (The SDSS Collaboration), 2005, AJ, 130, 367

Seljak U., Warren S., 2004, preprint (astro-ph/0403698)

Sembach K. R., Danks A. C., 1994, A&A, 289, 539

Shane C. D., Wirtanen C. A., 1954, AJ, 59, 285

Shapley A. E., Steidel C. C., Adelberger K. L., Dickinson M., Giavalisco M., Pettini M., 2001, ApJ, 562, 95

Shapley A. E., Steidel C. C., Erb D. K., Reddy N. A., Adelberger K. L., Pettini M., Barmby P., Huang J., 2005, ApJ, 626, 698

Sheth R. K., Mo H. J., Saslaw W. C., 1994, ApJ, 427, 562

Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001a, MNRAS, 320, 289

Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087

Somerville R. S., Primack J. R., Faber S. M., 2001b, MNRAS, 320, 504

Spergel D. N., et al. (the WMAP Team), 2003, ApJS, 148, 175

Spitzer L. J., Baade W., 1951, ApJ, 113, 413

Steidel C. C., 1993, in Shull J. M., Thronson H. A., eds, ASSL Vol. 188: The Environment and Evolution of Galaxies Dordrecht: Kluwer, p. 263

Stoughton C., Lupton R. H., Bernardi M., et al. (The SDSS Collaboration), 2002, AJ, 123, 485

Strauss M. A., Weinberg D. H., et al. (The SDSS Collaboration), 2002, AJ, 124, 1810

Suzuki N., Tytler D., Kirkman D., O'Meara J. M., Lubin D., 2005, ApJ, 618, 592

Tegmark M., Blanton M. R., Strauss M. A., et al., 2004, ApJ, 606, 702

Tegmark M., Bromley B. C., 1999, ApJ, 518, L69

Terlevich E., Diaz A. I., Terlevich R., 1990, MNRAS, 242, 271

Thomas D., Maraston C., Bender R., 2003, MNRAS, 343, 279

Tipping M. E., Bishop C. M., 1999, Journal of the Royal Statistical Society, Series B, 61, Part 3, 611

Tremonti C. A., Heckman T. M., Kauffmann G., et al., 2004, ApJ, 613, 898

Treyer M. A., Ellis R. S., Milliard B., Donas J., Bridges T. J., 1998, MNRAS, 300, 303

Valageas P., Munshi D., 2004, preprint (astro-ph/0403593)

vanden Berk D., Yip C., Connolly A., Jester S., Stoughton C., 2004, in ASP Conference Series, Volume 311 p. 21

vanden Berk D. E., Richards G. T., Bauer A., et al., 2001, AJ, 122, 549

Venn K. A., 1999, ApJ, 518, 405

Verde L., Heavens A. F., Percival W. J., et al. (The 2dFGRS Team), 2002, MNRAS, 335, 432

Verde L., Peiris H. V., Spergel D. N., et al. (the WMAP Team), 2003, ApJS, 148, 195

Vladilo G., 2004, A&A, 421, 479

Vladilo G., Péroux C., 2005, astro-ph/0502137

Wang J., Hall P. B., Ge J., Li A., Schneider D. P., 2004, ApJ, 609, 589

Watson F., Offer A. R., Lewis I. J., Bailey J. A., Glazebrook K., 1998, in ASP Conf. Ser. 152, Fiber Optics in Astronomy III pp 50–59

Weatherley S. J., Warren S. J., Møller P., Fall S. M., Fynbo J. U., Croom S. M., 2005, MNRAS, 358, 985

Welty D. E., Hobbs L. M., Lauroesch J. T., Morton D. C., Spitzer L., York D. G., 1999, ApJS, 124, 465

Welty D. E., Lauroesch J. T., Blades J. C., Hobbs L. M., York D. G., 1997, ApJ, 489, 672

Welty D. E., Morton D. C., Hobbs L. M., 1996, ApJS, 106, 533

Weymann R. J., Morris S. L., Foltz C. B., Hewett P. C., 1991, ApJ, 373, 23

Whitney C. A., 1983, A&AS, 51, 443

Willmer C. N. A., da Costa L. N., Pellegrini P. S., 1998, AJ, 115, 869

Wolfe A. M., Prochaska J. X., Gawiser E., 2005, ARA&A, in press

Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, ApJS, 61, 249

Wyse R. F. G., Gilmore G., 1992, MNRAS, 257, 1

Yip C. W., Connolly A. J., Szalay A. S., et al., 2004, AJ, 128, 585

Yip C. W., Connolly A. J., vanden Berk D. E., et al. 2004, AJ, 128, 2603

York D. G., et al. (The SDSS Collaboration), 2000, AJ, 120, 1579

Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, MNRAS, 335, 762

Yoshikawa K., Taruya A., Jing Y. P., Suto Y., 2001, ApJ, 558, 520

Zehavi I., Blanton M. R., Frieman J. A., et al., 2002, ApJ, 571, 172

Zehavi I., Weinberg D. H., Zheng Z., et al., 2004, ApJ, 608, 16

Zwaan M. A., van der Hulst J. M., Briggs F. H., Verheijen M. A. W., Ryan-Weber E., 2005, MNRAS, in press

# Acknowledgments

There are many people who have helped this thesis to be completed, by contributing to the science or to keeping me sane during the three years I spent in Cambridge. First and foremost my thanks go to Paul, who kept my work going in almost straight lines, fed me chocolate and coffee, gave me a lot more confidence in my abilities (still not there yet!), provided interesting and challenging projects, even put up with my tempers, tried to talk sense into me at times, and made the work fun... I could go on for the rest of the page. Once someone told me it is every students right to fall out with their supervisor after three years, I am certainly one of the luckier ones in really hoping my supervisor hasn't fallen out with me. I am also extremely grateful to Ofer, John Peacock and Max Pettini who supervised me/co-authored papers, for believing in me, teaching me and giving me considerable amounts of their time.

While on the topic of work there are a collection of people around the world who have helped with some aspects of the work presented in this thesis, by providing data, paper preprints, help with data analysis packages and advice. In order of chapters, thanks to Will Sutherland for really reading my 2dF relative bias paper and Mike Irwin for explaining aspects of statistics so well; Eric Switzer, David Schlegel and Patrick MacDonald for helping understand the SDSS data reduction; Simon Morris who refereed the sky subtraction paper and provided very helpful comments; Chris Akerman, Bob Carswell, Sara Ellison, Tae-Sun Kim, Jo Liske, Michael Murphy, Sandya Rao and Emma Ryan-Weber, all for helping with various aspects of my work on Ca II absorption line systems. Also to Jon Willis for entertaining me in Chile while we watched anemometer readings; Dan Mortlock for being my matrix algebra guru; Margaret Harding for sorting travel, money and buying nice chocolate things that made their way in my direction; and all the support staff that run the IoA so smoothly.

I've spent a lot of time in Cambridge trying to escape and I don't know whether Allan, Cecile, David, Heather, Helen, Jenny, Ross, Sue and Tom really know how important those sailing weekends in Scotland and the lake district, dinners in Edinburgh and various other weekend escapes were. Thank you to you all! A skiing weekend from Munich next?? A particularly big thank you to David, for being an extra special friend, and for sharing his entire music collection with me for all the seven years we've known each other. It really, really helped the thesis writing. Sailing being mentioned, I'd better continue - those at Grafham Water Sailability, who I've seen less of than I would have liked this year, Mike and his vintage boats, the very special group of youngsters and their leaders who I sailed with in the Tall Ships race in July and all the other groups I've sailed with, and especially Bill and Nick on Excelsior – all helped me stay in touch with reality... and sailing. In Cambridge a special thanks go to Nicole, Nutan and Ziggy – its been lots of fun, and to all my fellow PhD students for the mutual encouragement and college dinners.

The final week of writing this thesis was spent at my Dad's desk overlooking Little Loch Broom in Wester Ross. I can't really thank a part of Scotland, but one day I hope I'll never have to leave. Thank you to Paula, for such wonderful cooking and putting up with me not even helping with the washing up. And to my Granny and Lis, just for being there always.

Finally, there are three people who deserve the last word. Dad, Michael and Matt – without your support I might not have got to where I am today, particularly my Dad for teaching me to be a scientist almost from the day I was born. I know you all three put up with a lot. I am forever grateful to you Matt for sticking around – sometimes I've no idea why you do – one day we *will* manage to live in the same country... or town... or even house.