The Fossil Record of Star Formation from Galaxy Spectra

BEN PANTER

Institute for Astronomy
School of Physics

A thesis submitted to the University of Edinburgh
for the degree of Doctor of Philosophy

September 2004

Abstract

In this thesis I present work using the MOPED algorithm to extract in a non-parametric fashion star formation histories and galaxy masses from the spectra of galaxies in the Sloan Digital Sky Survey. The recovered parameters for all galaxies are combined to give insight into the processes of star and galaxy formation on both individual galaxy and cosmic scales.

The MOPED algorithm allows use of the entire spectral range, rather than concentrating on specific features, and can be used to estimate the complete star formation history without prior assumptions about its form. By combining the star formation histories of 96,545 galaxies in the redshift range $0 < z < 0.34$ the cosmic star formation rate is determined from the present day to $z \sim 6$. The results show that the peak of star formation occurred at $z \sim 0.6$, and that 26% of the mass of stars in the present-day Universe was formed at $z > 2$. The average metallicity rises from $\frac{Z}{Z_\odot} = 0.44$ at high redshift to a peak of 0.8 at $z \sim 1$ before declining to a level around 0.25 at the present day. Although the peak in star formation is more recent than previously thought, the sample used includes galaxies with a range of masses not accessible to traditional studies, down to a limit of $L \sim 2 \times 10^{-3} L_*$.

By cutting the sample into ranges of mass it can be seen that the redshift at which star-formation activity peaks is an essentially monotonically increasing function of final stellar mass. The time of the peak in star formation ranges from $z > 2$ for the highest mass galaxies ($M_S \gtrsim 10^{12} M_\odot$) to $z \sim 0.2$ for the lowest ($M_S \lesssim 10^{10} M_\odot$). A typical L_* galaxy appears to have its peak at around $z \sim 0.8$. These differences in star formation with mass reconcile the redshift of the peak found in this work with the previous estimates, generally deep surveys only probe the SFR of galaxies with $M_S \gtrsim M_{L_*}$.

The stellar mass calculated using the reconstructed spectra eliminates contamination from either emission lines or AGN components. Using these masses it is possible to construct the mass function for the stellar mass component of galaxies which give excellent agreement with previous works, but extend their range by more than two decades in mass to $10^{7.5} < M_s/h^{-2} M_\odot < 10^{12}$. I present both a standard Schechter fit and a fit modified to include an extra, high-mass contribution, possibly from cluster cD galaxies. The Schechter fit parameters are $\phi^* = (7.8 \pm 0.1) \times 10^{-3} h^3 \mathrm{Mpc}^{-3}$, $M^* = (7.64 \pm 0.09) \times 10^{10} h^{-2} M_\odot$ and $\alpha = -1.159 \pm 0.008$. The sample also yields an estimate for the contribution from baryons in stars to the critical density of $\Omega_{\mathrm{b}*} h = (2.39 \pm 0.08) \times 10^{-3}$, in good agreement with other indicators. No evolution of the mass function in the redshift range $0.05 < z < 0.34$ is apparent, indicating that almost all stars were already formed at $z \sim 0.34$ with little or no star formation activity since then and that the evolution seen in the luminosity function must be largely due to stellar fading.

The star formation history can be interpreted as a measure of how gas was transformed into stars as a function of time and stellar mass: the Baryonic Conversion Tree (BCT). There is a clear

correlation between early star formation activity and present-day stellar mass: the more massive galaxies have formed about 80% of their stars at $z > 1$, while for the less massive ones the value is only about 20%. Comparing the BCT to the dark matter merger tree indicates that star formation efficiency at $z > 1$ had to be high (as much as 10%) in galaxies with present-day stellar mass larger than $2 \times 10^{11} M_{\odot}$, if this early star formation occurred in the main progenitor. The LCDM paradigm can accommodate a large number of red objects; it is the high efficiency in the conversion from gas to stars that needs to be explained. On the other hand, in galaxies with present-day stellar mass less than $10^{11} M_{\odot}$, efficient star formation seems to have been triggered at $z \sim 0.2$. This work shows that there is a characteristic mass ($M \sim 10^{10} M_{\odot}$) for feedback efficiency (or lack of star formation). For galaxies with masses lower than this, feedback (or star formation suppression) is very efficient while for higher masses it is not. The BCT, determined here for the first time, should be an important observable with which to confront theoretical models of galaxy formation.

Declaration

I hereby declare that this thesis entitled *The Fossil Record of Star Formation from Galaxy Spectra* is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University. I further state that no part of my thesis has already been or is being concurrently submitted for any such degree, diploma or other qualification.

Parts of the work contained in this thesis have been published, or are due to be published, in refereed scientific journals.

> On the star formation rate – '*Star Formation and Metallicity History of the SDSS Galaxy Survey: Unlocking the Fossil Record*', B. Panter, A.F. Heavens, R. Jimenez, Monthly Notices of the Royal Astronomical Society, Volume 343, Issue 4, pp. 1145-1154
>
> On the star formation rate – '*The Star-Formation History of the Universe from the Stellar Populations of Nearby Galaxies*', A.F. Heavens, B. Panter, R. Jimenez, J. Dunlop, Nature, Volume 428, Issue 6983, pp. 625-627 (2004)
>
> On the stellar mass function – '*The Mass Function of the Stellar Component of Galaxies in the Sloan Digital Sky Survey*', B. Panter, A.F. Heavens, R. Jimenez, Monthly Notices of the Royal Astronomical Society, in press.
>
> On the baryonic conversion tree – '*Baryonic Conversion Tree: The Global Assembly of Stars and Dark Matter in Galaxies from the SDSS*', R. Jimenez, B. Panter, A.F. Heavens, L. Verde, Monthly Notices of the Royal Astronomical Society, in press.

This thesis is the outcome of my own work except where specifically indicated in the text.

Ben Panter
Edinburgh,
September 2004.

Acknowledgements

This thesis is the cumulation of three years work at the Institute for Astronomy in the Royal Observatory, Edinburgh. Many people have helped me on my way and I'd like to mention a few of them here. Prime amongst them is my supervisor, Alan Heavens, who I'd like to thank for being an exceptionally nice chap and generally sorting out whatever fine messes I managed to get us into, while simultaneously managing to teach me cosmology, statistics and rudimentary grammar. He was aided and abetted by Raul Jimenez, my co-supervisor in the states, who was able to drive my project forward by phenomenal effort and encouragement.

My collaborators, Jim Dunlop and Licia Verde, deserve thanks – as do all those who have offered me advice when I needed information beyond the expertise of my supervisors. I'm indebted to the referees of our publications and my thesis examiners for providing insights into my work which would otherwise have been missed. I'd like to thank Liz Gibson for greasing the gears of any part of University bureaucracy I was unfortunate enough to have to interact with.

At the observatory I've been surrounded by a great bunch of people – particular thanks to Dave for elevating the clicky-ball game to an art form, Mairi for being far too easy to wind up, Ian and Tara for programming hints, Olivia for keeping things in perspective, Jess for supporting my astronomy pub relocation efforts, Rachel for believing in the Astronomer Royal's ashes, Michael for introducing me to Zebras, Niall for his fear of everything and Martin for having an unquenchable thirst. All the students at Edinburgh, both undergraduate and postgraduate, who have helped make the ROE what it is – thank you.

Outside of my work life I've had some great times in Edinburgh. I'm indebted to all of EUSAC for giving me three years of incredible diving, to my Hippo teammates for helping to start something brilliant, to Mike and Peter for introducing me to sailing and to Lucy, Ian, Annemarie, Aimee, Sarah, James, Keff and all the others who have been there when I needed someone to talk to or just get drunk with. Helen deserves a special mention for putting up with me for so long, for always being there to listen when I felt the need to rant about something, and for bringing me up if ever I felt down. Of course, without the support of Mum, Dad and Pootle I would never have started, let alone finished, this work.

Contents

List of Figures

Chapter 1

Introduction

A fundamental question for astronomy is "How did the Universe come to be as we observe it today?". In this chapter I summarize the current understanding of the evolution of the Universe from the first moments after the Big Bang to the present day. I then focus on the methods used to investigate the most recent stage of the evolution - the conversion of gas to stars in galaxies, referred to as the *star formation rate*. Also contained in this chapter is an overview of how stars are modelled, a description of the data used in the rest of the thesis and the tools required to place the results in a cosmological context.

1.1 Understanding the Universe

1.1.1 The Early Universe

The Universe as we observe it today is dotted with stars, galaxies and clusters of galaxies. Until the beginning of the 20th Century, it was thought that the Universe had always been in this state, and that it always would be - this condition is known as a static universe. Theoretical solutions to the Einstein field equations which were not static but evolved with time were given independently by Friedmann and Lemaître, but it was only with the discovery of redshift in galaxy spectra by Slipher in the 1920's, and the understanding that this meant that almost all observed galaxies were receding from us by Hubble & Humason (1931), that these alternative theories were credited. Hubble, and later many other workers, attempted to measure the rate of the expansion using distance ladders to work out the distance to distant galaxies. This work is extremely difficult and although estimates from ground based observations settled on two quite precise measurements, they were irreconcilably a factor of two apart. An accurate value for the expansion could not be obtained until the Hubble Space Telescope (HST) was used for the Hubble Key Project on the Extragalactic Distance Scale (HKPEDS, Freedman, 1994). This project used the relationship between the period and luminosity of Cepheid stars to determine the distance to galaxies, and obtained a value of the Hubble constant of $H_0 = 80 \pm 17$ km sec^{-1} Mpc^{-1}. With more observations the project was able to refine this estimate, and finally set the value as $H_0 = 72 \pm 8$ km sec^{-1} Mpc^{-1} (Freedman et al., 2001).

Theorists suggested two different models to account for the expansion observed. The steady state model explained the expansion as the constant creation of a tiny amount of matter, whereas the big bang model suggested that in fact the Universe came into being some billions of years ago, and has been expanding ever since. Although there was considerable argument between these models for many years all current evidence points to the big bang model. The balance between the expansion and the force of gravity determines the eventual fate of the Universe – a choice between expanding forever and eventually recollapsing to a "big crunch". The border between these two extremes, which in a Universe dominated by matter is spatially flat, has a density equal to the *critical density*. The ratio of the density of the Universe to the critical density is known as Ω.

Gamow (1946) suggested that in order to explain the abundances of metals in the present day in the context of a Big Bang universe, there must be some background radiation in the Universe, observable in the microwave region. This radiation was first discovered by Penzias & Wilson (1965), and explained in the context of the cosmic microwave background (CMB) by Dicke et al. (1965). These measurements showed some variation of temperature over the sky, initially limited to a broad dipole. The Penzias & Wilson (1965) experiments were consistent with a black body curve peaking at a temperature around 3K, and subsequent studies at different wavelengths reinforced this idea. The Cosmic Background Explorer (Boggess et al., 1992, COBE,) satellite was able to refine these measurements showing that not only was the radiation an exact black body curve, but that the temperature corresponding to the peak in the curve was by no means constant across the sky. The satellite observed ripples superimposed on the dipole. These tiny fluctuations in the temperature over the sky carry information about the distribution of matter in the early Universe. The magnitude and distribution of the ripples (measured by their power spectrum) can supply information on the formation of structure, and the underlying cosmology of the Universe. The map of the sky produced by COBE was insufficiently resolved to directly determine Ω, but

subsequent experiments, culminating in the balloon missions BOOMERANG (Lange et al., 2001) and MAXIMA(Hanany et al., 2000) were able to constrain the Universe to being almost certainly flat de Bernardis et al. (2000). A further satellite mission, the Wilkinson Microwave Anisotropy Probe (WMAP, Bennett et al., 2003), allowed an even more precise determination of both Ω and the contributions to this density of the different flavours of energy in the Universe (Spergel et al., 2003). The concordant cosmology reached allows us to determine that the age of the Universe is approximately 13.7 billion years and that its density is very close to the critical density, inferring that it is flat. The results of the WMAP experiment, when combined with all those preceding it, set the contributions to Ω from baryons, cold dark matter and Λ as $\Omega_b \sim 0.04$, $\Omega_{CDM} \sim 0.23$ and $\Omega_\Lambda \sim 0.73$. They also give further confirmation on the value of the Hubble constant, with $H_0 = 71 \pm 4$ km sec^{-1} Mpc^{-1}.

1.1.2 The Cosmic Dark Ages

The big bang model gives us an excellent framework for the formation of the matter in the Universe, and the setting of the physics that govern it. The fluctuations in the CMB show that matter was not distributed evenly across the Universe, giving some areas slight overdensities and leaving others with below average density. As the plasma cools, electrons and protons recombine to form atoms, at a redshift of ~ 1000. The presence of so much ground state hydrogen gas in the Universe at during this period results in almost all radiation being absorbed. Since the Universe is no longer opaque, at some point the gas must have been reionized. The WMAP results (Spergel et al., 2003) suggest this may have started as early as $z \sim 17$, and the presence of high redshift quasars (Fan et al., 2004) suggest that by a redshift of ~ 6 the universe was reionized. Although the source of the reionizing photons is yet to be established, it is thought that some form of star formation (although probably not in a form we recognize today) is responsible. Evidence from the Lyman-α forest, the lines at shorter frequencies than the rest frame Lyman-α peak, suggest that recombination was a patchy process, although beyond this there is little information.

1.1.3 The Formation of Structure

After reionization, we are left with a Universe consisting of clumps of dark matter, into which baryons can fall. Such regions are unstable, as gravity causes matter near to the overdense regions to fall into them, causing the denser regions to get denser. The compression will in general heat the gas, which may oppose further gravitational collapse of baryons in dark matter halos, but if they can cool the pressure lowers and the clouds form high density objects. These primordial galaxies may fragment into protostellar cores in a process which is not fully understood. As the cores collapse they become hot again. When their temperature reaches the critical temperature for fusion reaction a star is created. These protogalaxies have masses around $10^8 M_\odot$, and their distribution will mimic that that of the CMB fluctuations. As time passes, the protogalaxies will merge with their neighbours, and the more dense regions will evolve far larger systems than those in less dense areas. We believe that the more dense areas go to form the clusters which are observed today, where the smaller field galaxies come from tiny overdensities embedded in large areas which were initially underdense. This process of building up larger galaxies by the accumulation of smaller building blocks is known as hierarchical formation, first suggested by Peebles (1970). The distribution of galaxy masses expected from such a process was predicted by

Press & Schechter (1974), and later tested using N-body simulations, and the current state of the art allows the buildup of a large galaxy to be simulated by a merger tree (Wechsler et al., 2002; Somerville & Primack, 1999).

Although these hierarchical clustering techniques were first considered in the absence of any non-baryonic dark matter, it became clear from the level of the CMB fluctuations that dark matter does indeed contribute the majority of the matter energy in the Universe. Peebles (1983) suggested a unified scheme for formation of galaxies, the Cold Dark Matter (CDM) model. Further studies of the acceleration of the Universe (Perlmutter et al., 1999; Hanany et al., 2000; Lange et al., 2001) pointed to the presence of so called *dark energy*. This dark energy experiences a repulsive rather than attractive gravitational force, and is represented by Einstein's cosmological constant Λ. This, of course, had to be incorporated into the model of structure formation, and led to the ΛCDM model, currently the most successful solution to the problem of formation of large scale structure in the Universe.

1.1.4 The Formation of Stars

Up to this point the evolution is well understood. We have a model which takes us from before the CMB to the formation of structure in the Universe and the evolution of galaxies and clusters, with only a small gap in our knowledge about the sources of reionization. We understand the size distribution of galaxies, and that there is an upper limit on the sizes observed set by inefficiency of cooling mechanisms. The problem now moves to the smaller scale: the formation of stars. In simple terms, clouds of cool gas collapse under gravity, gradually increasing in temperature. When the temperature is high enough, the gas will start to undergo fusion reactions, creating a star. The larger the cloud of gas, the larger the star. The larger stars burn their fuel up and produce supernovae very quickly, while the smaller stars continue to evolve for many billions of years. Modelling of the nuclear reactions which occur in stars gives information about the shape of the spectrum at different stages of the star's evolution, but modelling the initial collapse of the clouds and the distribution of mass of stars produced is difficult - our understanding of the mechanisms that control the collapse, and therefore the conversion of gas to stars, is poor. Although an assembled star can be modelled by assuming hydrostatic equilibrium, the factors which determine the collapse of clouds into stars is not well modelled. Feedback, the process by which stellar winds and supernovae explosions from recent star formation quench or increase activity, could play an important part. The feedback reduces activity by either blowing away the fuel required for stars to form or heating it to such an extent that it cannot be virialized. Feedback could also manifest itself in a positive sense by shocking gas to such an extent that star formation is triggered. The theoretical modelling of these processes and their dependence on both galaxy type and environment is extremely hard. Star formation varies across time and location in each galaxy, and modelling the different sources of positive and negative feedback and estimating their relative strengths leads to models with many parameters which can only be crudely estimated. In order to understand galaxy formation information from observations is essential. A part of this information is given by the rate of conversion of gas to stars, and how that changes with environment and galaxy type. This star formation rate (SFR) is not only of key importance for theories of formation of individual galaxies; the study of the SFR over cosmic time allows us to probe from the initial stages of protogalaxy formation, and investigate the effect of hierarchial formation on the star formation rate.

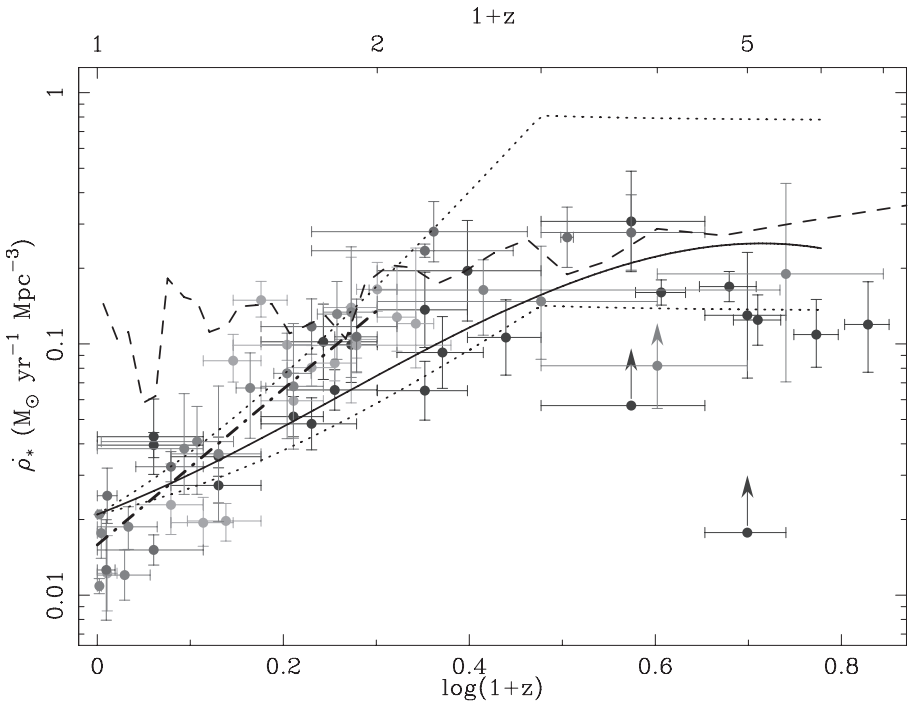


Figure 1.1: Estimates of the SFR by various methods at different redshifts, taken from Hopkins (2004)

1.1.5 Summary of Current Star Formation estimates

The SFR can be measured in two complementary ways. In the traditional approach, the evolution of the cosmic SFR is determined by measurements at different redshifts of the *instantaneous* SFR, and estimates for the Star Formation History (SFH) of galaxies must be estimated by assumption. Methods of measuring the instantaneous SFR are given later, but over a wide range of redshifts many different techniques are required, and although there is some agreement the only clear trend that can be observed between these different measurements is that there was a peak at around a redshift of 1 or higher, and decline since then (Hopkins, 2004), see figure 1.1. The decline at redshifts higher than the peak is debatable - the points are subject to large corrections for dust extinction and integration over the luminosity function. In this thesis I present a different approach - that of interrogating the fossil record. Instead of relying on the instantaneous SFR indicators, this technique analyzes the SFH of individual galaxies, tracing the buildup of their stellar mass by identifying the various stellar populations which make up a galaxy. Since the age of these populations can be determined, the overall SFH allows the calculation of the SFR at many epochs from a single galaxy. Since low redshift galaxies can be used to study high redshift star formation, galaxies from the huge and high quality Sloan Digital Sky Survey (SDSS, Strauss et al. (2002)) survey can be used to probe SF at redshifts previously only accessible to the largest telescopes in the world. The processing of this huge dataset is only possible through the use of MOPED, a data compression algorithm introduced in Heavens et al. (2000) and refined herein.

1.2 Recovering Star Formation Information

An essential probe of formation of structure in the Universe is the Star Formation History (SFH). This is important both on a galaxy by galaxy level, to investigate how galaxies form, and on a grander, cosmic scale to understand how structure has evolved since the big bang. An important way to investigate this build up is by studying the light from distant galaxies. Information gathered from these sources provides a way of probing the interactions between the baryonic matter and the dark matter/energy which makes up $> 95\%$ of the Universe, and cannot be directly observed. In essence, the SFR is of interest as it provides an insight into galaxy formation, tracing of the conversion of gas into stars.

Initial studies of stars within the Milky Way showed a distribution of colour and luminosity which could be modelled by stars with different ages and metallicities. This suggested that the population of our own galaxy was not static - and that it could vary for external galaxies too. For the Milky Way it is possible to resolve a huge number of stars and determine their physical quantities. This *resolved population* technique is also possible for our satellite galaxies, and with space based telescopes has been extended as far as the local group. For the remainder of galaxies the problem is different - information must be extracted from *integrated populations*, where it is impossible to resolve individual stars. It is possible to use the local resolved populations (where exact physical properties are know) to calibrate our interpretation of the behaviour observed in the more distant integrated populations.

Many methods have been used to extract SFR information from galaxies, and in this section I attempt to describe those most commonly used for contemporary analysis. These have evolved from the first attempts to fit stellar colours from evolutionary synthesis models to galaxy colours to methods which can extract information on star formation from emission lines, UV, IR, sub mm and 1.4Ghz radio continuum flux, the 4000Å break and absorption lines. The visible spectrum of a typical galaxy is shown in figure 1.2

1.2.1 The Visible spectrum

Emission Lines

The spectra of certain galaxy types contain massive emission lines. These emission lines are tracers of very recent star formation, as they are created by recombination of ionized gas in nebulae which has been ionized by light with $\lambda < 912$Å. Light at this wavelength is only produced in any quantity by young (< 20 Myr) stars with $M > 10M_{\odot}$. On recombination electrons fall through the energy levels of the atom, emitting light at the wavelengths of H-α, H-β, P-α, P-β, Br-α, Br-γ. Of these, the strongest is the H-α line which can be used to trace these heavy, young stars. The calibration between the strength of the flux and the SFR requires use of stellar models and integration along the IMF. Differences in IMF and models account for a 30% deviation between the calibrations of Kennicutt et al. (1994); Kennicutt (1983)

An issue with the H-α technique is the matter of dust extinction. Dust acts to reduce the height of the emission lines, and without full spectral modelling its effect can only be estimated by examination of the relative heights of emission lines. Although this is fine when a number of emission lines are observed and a model dust screen can be applied, it can lead to large errors in systems which contain large amounts of dust or which have weak emission lines.

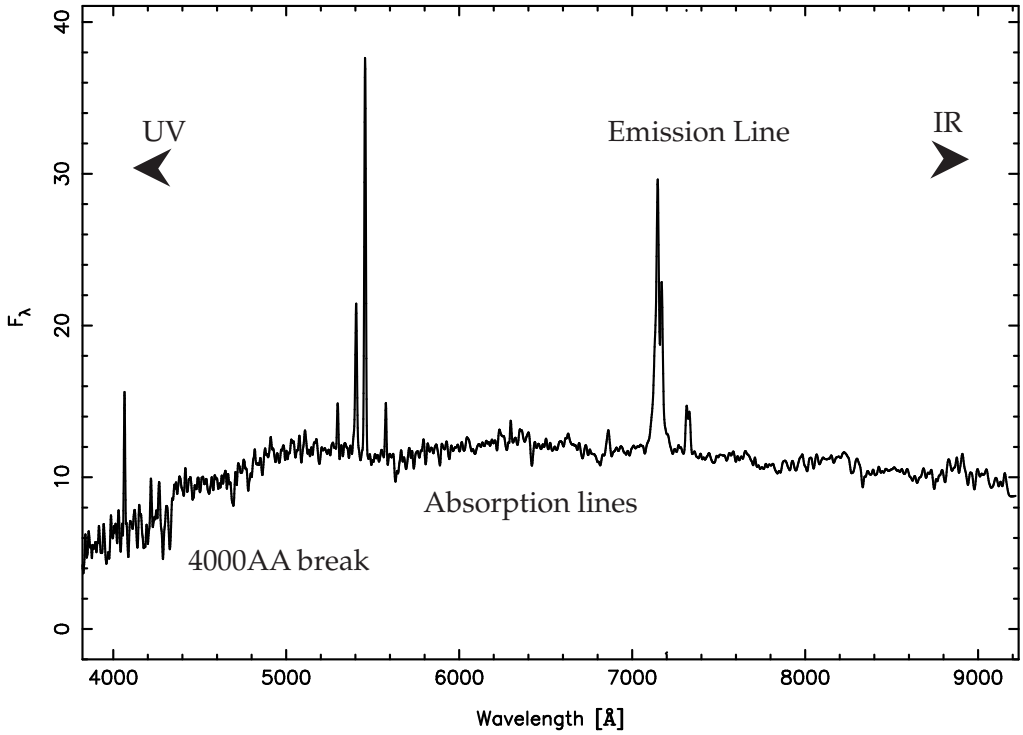


Figure 1.2: Typical features of a galaxy spectrum include emission lines, absorption lines and a 4000Å break(redshifted in this figure, which shows the observed spectrum from a galaxy at $z \sim 1$). The broadband colours of the galaxy can be seen by integrating the continuum.

The H-α line has been used as a reliable probe of SFR for galaxies with $z \gg 2$, although the line moves out of the visible range beyond $z \sim 0.5$. To determine the SFR of galaxies beyond this limit using only visible spectra it should be possible to use other emission lines such as H-β. Unfortunately in the majority of galaxy spectra these lines are weaker than the H-α and accurate calibration is impossible. An alternative emission line is the OII doublet, at 3727Å. This line remains in the visible spectrum until a redshift of 1.6, although calibrations are not quite as accurate as those of H$-\alpha$, with the two established calibrations of Gallagher et al. (1989) and Kennicutt (1992b) varying by a factor of 1.57.

Absorption Lines and the Lick Indices

If higher-resolution spectra are available then it is possible to examine individual absorption lines and link them to chemical abundances of stellar populations in the galaxy under study. The ratio and level of these chemical lines can then be used to fit stellar evolution models to the spectra and deduce the trends of star formation over the galaxies history. This allows perception beyond the instantaneous SFR studied before, and gives insight to the Star Formation History (SFH), or fossil record of stellar buildup in galaxies.

A key problem with these methods is the determination of the continuum level from which the depth of absorption should measured. Burstein et al. (1984); Faber et al. (1985) suggested that instead of attempting to estimate the true continuum, a pseudo-continuum should be established

by analysis of a set region of spectrum either side of the absorption line in question. The pseudo-continuum allows the strength of the feature to be studied independently of dust extinction or spectrophotometric calibration errors. These regions around 21 absorbtion lines were originally determined by the Lick Observatory, and in reference to this are called the Lick Indices. The set of 21 indices measured by the Lick observatory has been extended by Huchra et al. (1996) and Diaz et al. (1989) to around 35. A further limitation of analysis using line strengths is that it neglects the information available from the shape of the continuum, which often contains strong clues about the stellar populations of a galaxy.

Various workers have attempted to estimate the metallicities and ages using grids based on the Lick indices. In these sorts of analysis a line index sensitive to metallicity is plotted against one which is sensitive to age, and simple stellar population (SSP) models compared. This method obviously loses massive amounts of information, since only a few features can be considered. It also does not allow easy disentanglement of different stellar components and strong assumptions about the galactic SFR must be made before SFRs can be recovered. Modern techniques consider Principal Component Analysis (PCA, See later) although such studies are in their infancy (Strader & Brodie, 2004; Kong & Cheng, 2001). There is some potential to break the degeneracy between age and gas metallicity using some high resolution Lick Indices, although this work has not yet come to fruition.

The 4000Å Break

A prominent feature in the spectra of many older galaxies is a sharp discontinuity at 4000Å. This feature can have two sources: one is related to old populations of stars; the other to the overall metallicity of the gas which formed the star. In the first case the superposition of old populations at different black body temperatures forms the continuum: If there are no hot blue stars, the continuum falls sharply at this point. In the second case the break is composed of many absorption lines in a tight wavelength range, which at combine to form the observed break. The feature is known as the 4000Å break. The break was formalized as an index by Bruzual (1983). He found that the index was best estimated in a similar way to the Lick indices, by estimating the pseudo-continuum level either side of the break. This formal determination of the break is known as D(4000), and is used to determine the presence of an aged stellar population - but is obviously heavily affected by the presence of metals. By comparing the break in the spectra of various different stellar types he was able to plot a smooth curve of increasing D(4000) and spectral type. In isolation this is not a particularly useful tool in SFR analysis (Dressler & Shectman, 1987), but can be very usefully combined with other indicators to restrict the metallicity degeneracy (e.g. Kauffmann et al., 2003, see next section).

1.2.2 Advanced Spectral Analysis Techniques

The individual measurements mentioned above are subject to large errors in calibration, and more modern approaches attempt to minimize these errors by considering some collection of spectral features. Charlot & Longhetti (2001) suggests the use of a combination of various line ratios and emission lines are able to constrain SFR far better, and their method has been used to extract SF information from a large number of galaxies in different surveys (Charlot et al., 2002; Brinchmann et al., 2004). Kauffmann et al. (2003) claim that by using a combination of

the 4000Å break and the H-δ line (which, when combined, are thought to be insensitive to dust correction and metallicity) it is possible to determine the intermediate SSP age, and allow some insight into SFH. In fact, by using a collection of indices to obtain information about the SFH all these methods apply a data compression to the spectrum. The compression is determined by the areas which we think are good indicators of the various components of the SFH. While there is no question that this form of compression yields interesting results, it is possible that there are more subtle combinations of features which are able to yield far more information that just crude combinations - this is the aim of both the PCA and MOPED techniques.

Principal Component Analysis of Galaxy Spectra

The Principal Component Analysis (PCA, Murtagh & Heck (1987)) approach allows exploration of the correlations between pixels in a dataset. In the case of galaxy spectra, the data set comprises the N flux measurements. PCA constructs a set of N independent, orthogonal linear combinations of the indices. The combinations, called principal components, are ordered in decreasing contribution to the variance of the entire dataset. This is done algebraically by finding the eigenvectors and eigenvalues of the covariance matrix which is obtained from the variation of flux measurements across a sample of galaxy spectra[1] - the eigenvectors can then be ordered by their eigenvalues. Taking the data as a cloud of points in N-dimensional space, the amount of information contained in the first principal components depends on the correlation of the points. In the uncorrelated case, the points should cover the space widely. In this case the information on the variances will be spread over all the principal components. If some subset of the parameters are correlated, the spread of points will follow this correlation and not be uniformly distributed. In this case the initial principal components will contain far more information about the dataset that those that follow, and may correlate with trends in quantities which are associated with the galaxy spectrum in some complex way rather than with individual spectral features. The key problem here is determining the link between the principal components and physical parameters of the galaxy. Although some work has been done on this issue, at present it is forced to compare principal components of spectral models to principal components of large samples of galaxy spectra (Ronen et al., 1999; Madgwick et al., 2002).

1.2.3 UV Continuum Flux

The light in the ultraviolet continuum between 1250 and 2500 Å is almost completely produced by young stars with masses greater than $5M_{\odot}$. This window falls between the Lyman$-\alpha$ forest region and the part of the continuum where light from older stellar populations begins to dominate. Unfortunately, for galaxies closer than $z = 0.5$, this portion of the spectrum is blocked by the atmosphere and either space-based or high-altitude instruments are required. The rocket missions of Smith & Cornett (1982) provided photometric UV fluxes for 201 galaxies in the Virgo cluster, while data from the Orbiting Astronomical Observatories (OAO-1,2,3) was used by Donas & Deharveng (1984) to estimate the current day SFR. Using instruments mounted on balloons, UV fluxes for many more galaxies could be obtained (Donas et al., 1987), and by the mid 90's the FOCA balloon-born telescope was able to determine the continuum of 254 galaxies in direction of the Coma Cluster, at $z = 0.023$ (Donas et al., 1995).

[1] NB: this definition differs from the covariance matrix used by MOPED in later chapters

The problem of the atmosphere is removed for distant galaxies. For spectra from galaxies with redshift greater than ~ 1.5, the UV continuum is again observable as it has been shifted into the visible. Steidel et al. (1996) used spectrometry from the Keck Telescope to determine the UV flux of galaxies out to $z \sim 3$, and the most recent surveys such as the Subaru/XMM Deep Field (Ouchi et al., 2004) have identified, and measured, the rest frame UV continuum flux from 2600 galaxies within the range $3.5 < z < 5.2$.

Since the UV flux is produced only by short lived stars with $M > 5M_{\odot}$ the flux in the 1250 - 2500 Å region should be directly proportional to the SFR. An extrapolation along the IMF is required in order to derive the SFR of all stars, but although the IMF is reasonably constrained, small changes in slope and shape will affect the recovered SFR dramatically. An addition consideration is the extinction due to dust. This extinction has been found to be non-uniform (Calzetti et al., 1994), which could lead to large errors when calculating the SFR. Differences in stellar evolution models mean that even when a common IMF and reference wavelength are employed estimates of the conversion factor required varies by a factor of ~ 2. Despite this, the UV continuum flux method for recovering SFR information is applicable over a wide range of redshifts, and so far is the most successful probe of SF in high redshift galaxies.

1.2.4 IR Continuum Flux

When starlight is incident on a dust grain, that grain will heat up. These grains then re-emit the light in the Far IR (FIR) region of the spectrum, most strongly in the 10-100 μm range. Adsorption by grains is determined by the relative sizes of the grains and the wavelength of the light - and is strongly peaked in UV wavelengths. If the dust is assumed to be optically thick, then all UV light will be re-emitted in the FIR. Since the UV light is a very strong tracer of young, massive stars (as described in the previous section), the FIR continuum can be used as a SFR measure. The FIR fluxes of $\sim 30,000$ galaxies were observed by the the IRAS survey in the mid 80's, and the European Large-Area ISO Survey (ELAIS) survey (Oliver et al., 2000) extended the redshift range of these methods to the depth of the Hubble Deep Field (HDF). Information of this breadth should be able to yield a vast amount of SFR information but unfortunately, determining an accurate calibration to convert these fluxes to a SFR is difficult. In addition to the IMF problems mentioned in the previous section, assumptions must be made about the star formation timescale. This timescale can be determined by either the timescale of the starburst or by the lifetime of dust cloud, but choosing the second means that at some point the dust can no longer be thought of as optically thick, and the assumption that all UV light is re-emitted by the dust fails. Different timescales yield different calibrations, but assuming the same fraction of light is remitted most lie within $\pm 30\%$ of that suggested by Kennicutt (1998). The exact fraction of light that is re-emitted is well defined for dusty star burst regions in galaxies - it is almost 1 - but for more normal galaxies the distribution of dust is unknown. The distribution varies not only with galaxy type, but also with location in the galaxy. This ambiguity is the major failing of the FIR continuum method, as in all likelihood a different calibration is required for each galaxy. In the special case of dusty circumnuclear starbursts, where the dust can be assumed to be optically thick, the FIR continuum provides an excellent probe of SFR.

1.2.5 Sub-mm Continuum Flux

For higher redshift sources, the light which has been remitted into the FIR region is observed in the sub-mm region. Recent instruments such as SCUBA have allowed observations of the dust re-emitted flux out to $2 < z < 5$. The redshifting of the peak of the re-emitted spectrum into the observing window counteracts the extra distance of higher redshift galaxies and sources can remain relativly bright out to high redshift, which makes this technique very effective at detecting high-z ($z \geq 2$) sources.

1.2.6 Radio Flux

The radio flux of the non-AGN component of a galaxy comes from three sources. At low frequencies ($\nu < 30$ Ghz), synchrotron emission from relativistic electrons dominates. At frequencies over ~ 200 Ghz, the emission is dominated by re-emission from dust. Between these two limits the dominant source of radio is free-free emission from photoionization in H II regions. Both the free-free and the synchrotron emission give some idea of recent starformation. Photoionization of the gas is caused by massive young bright stars. The movement of ionized (free) electrons near to (free) protons gives the thermal free-free emission, tracing the presence of stars emitting light capable of photoionization. The synchrotron component originates from the massive shocks caused by the supernovae (SN) explosions of young, massive stars. Choosing frequency range low enough to avoid contamination from the dust re-emission allows differentiation between these two components of the flux by their slope. For synchrotron emission component the spectrum goes as $\nu^{-0.8}$, while for the free-free component the spectrum is much shallower, with $\nu^{-0.1}$. This allows the standard model for galaxy radio emission (Condon, 1992) to fix the fraction of the flux due to free-free emission at around $\nu \sim 1.4$ Ghz to be 0.1. To properly determine this fraction requires characterization of the radio spectrum, entailing several measures of flux at different frequencies. Instead, since both components are closely correlated and essentially measure of the same thing (recent SFR), it has become common to apply a simple calibration to the flux at 1.4Ghz. This method allows the formation of the heavier stars to be observed.

Again, extrapolation along the IMF is required to produce overall SFR from this high stellar mass information. There is still discussion over the proper calibration of the 1.4 Ghz flux (Condon, 1992; Bell, 2003), but estimates agree to within a factor of ~ 2.

1.2.7 The Current State of the Cosmic Star Formation Rate

Although the various star formation indicators have been calibrated against each other (Rosa-González et al., 2002), there is still considerable scatter between the various measures of the cosmic star formation rate (Figure 1.3). No single method can be applied over the range of redshifts studied, and although the decline from distant times to the present day is certain, the location of the peak, if it is a peak, is not certain. With mild adjustments to the dust and mass function corrections made to these estimates (the corrections can be as large as a factor of 20), it would be possible to have a SFR which is constant earlier than the peak, or rises or falls.

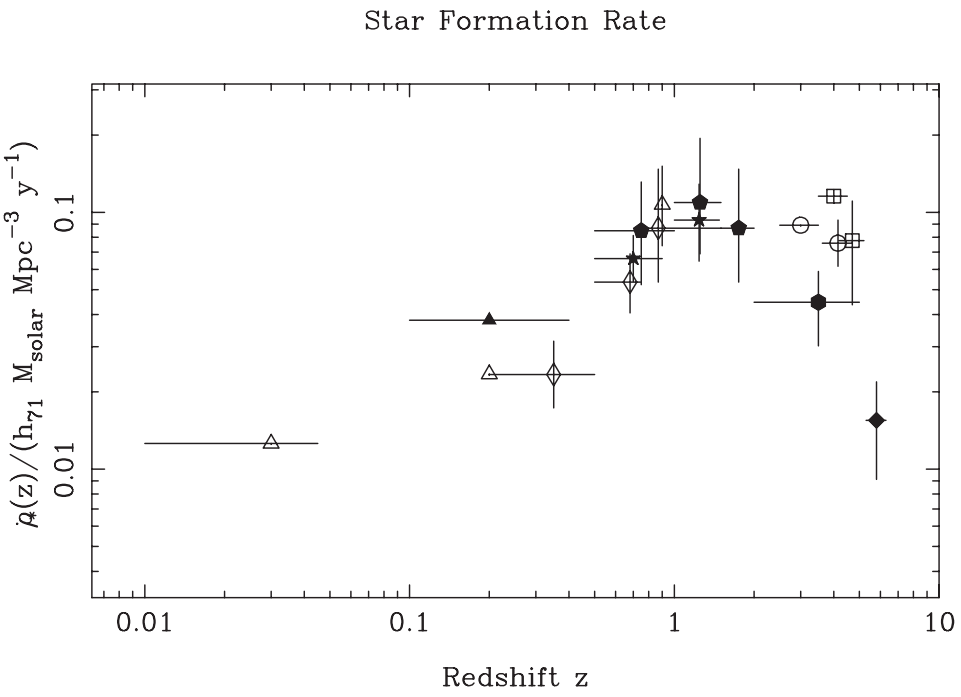


Figure 1.3: The current state of the SFR measured over the lifetime of the Universe, using instantaneous measurements of the star formation rate at various redshifts; $H\alpha$ measurements are open triangles at $z \simeq 0.03$ (Gallego et al., 1995), $z \simeq 0.2$(Tresse & Maddox, 1998), $z \simeq 0.9$ (Glazebrook et al., 1999); UV from Subaru (Ouchi et al., 2004, open squares) , GOODS (Stanway et al., 2003, filled diamond), HST etc (Steidel et al., 1996, open circles), CFRS (Lilly et al., 1996, open diamonds), HDF (Connolly et al., 1997, filled pentagons), galaxies (Cowie et al., 1999, stars), galaxies (Sullivan et al., 2000, filled triangle). The filled hexagon at $z = 3.5$ represents a new estimate of the star-formation density provided by sub-mm galaxies in the redshift range $2 < z < 5$ provided by J. Dunlop for Heavens et al. (2004)

1.3 Modelling of Galaxy Spectra

Both instantaneous SFR indicators and the MOPED algorithm depend heavily on the accurate modelling of the light output of galaxies. In order to construct a normal galaxy spectrum, the effects of the stars and dust in the galaxy must be modelled. In this section I present a summary of the ingredients for stellar population and dust models.

1.3.1 Modelling Stellar Populations

The main component of the optical light in a typical galaxy is from stars. The stars will be of various ages and they will have been formed from gas with different metallicity[2], and accounting for them individually is an impossible task. Instead, when considering their integrated spectrum, it is typical to talk about populations. A simple stellar population is coeval and formed from gas with uniform metallicity. More generally, simple stellar populations (modelled by a single burst of star formation) provide the components of any stellar population. Although they do not provide the whole story of star formation in a galaxy, merging several populations together can

[2]From here on, the metallicity of the gas which formed the star will be referred to as simply "the metallicity"

give excellent agreement with observed galaxy colours. To obtain spectra for these populations of stars, and those with different metallicities in the gas which went to form them, we require complex models. These models consist of stellar isochrones, the initial mass functions (IMF) of the stars, stellar fluxes and absorption line strengths. The following briefly explains and links these components of the model together. A more in depth guide to stellar populations can be found in Binney & Merrifield (1998)

Isochrones

A complete analysis of stellar structure for a non-rotating, spherically symmetric star requires the solution of coupled differential equations. The different components of the system of equations consider conservation of mass, pressure, energy conservation, energy transport and composition change over the lifetime of the star. Since the evolution of the star is a slow process, at all stages the star is very close to hydrostatic and thermodynamical equilibrium and numerical methods can be used to solve the system.

The solutions yield a description of the luminosity and colour (temperature) of a star at a given age which formed from gas with a certain metallicity. Tracing the route that different stars take over their lifetimes gives *isochrones* on a diagram of luminosity against temperature. An isochrone represents the locus of stars in such a diagram with the same age and metallicity. Figure 1.4, adapted from Worthey (1994), shows the changes in an population when either the age or the metallicity changes. When these parameters change simultaneously (as in real populations) it is very difficult to disentangle them. This problem is known as the age-metallicity degeneracy.

The Initial Mass Function

The Initial Mass Function (IMF), $\xi(M)$, determines the relative numbers of stars of different masses in a population which has just been created and is defined as the number of stars per unit volume per unit interval of logarithmic mass. A complete review of the IMF is given in Kroupa (2002). The most commonly used IMF, and that used in this work, is that proposed by Salpeter (1955), a power law of the form

$$\xi(M) \propto M^{-1.35} \tag{1.1}$$

for $0.1M_{\odot} < M < 100M_{\odot}$. More recent studies have shown that in fact the IMF is better fit by a number of power laws, or by including cuts at the high and low mass ends. For example, Scalo (1986) suggests that

$$\xi(M) \propto \begin{cases} M^{-2.45} & \text{for } M > 10M_{\odot}, \\ M^{-3.27} & \text{for } M_{\odot} < M < 10M_{\odot}, \\ M^{-1.83} & \text{for } M < M_{\odot} \end{cases} \tag{1.2}$$

The IMF allows average quantities to be obtained for a population of stars if they are assumed to be coeval.

Stellar Fluxes

The isochrone curves give a luminosity and effective temperature for each point along the tracks of the stars. If we could resolve individual populations it would be possible to extract age directly,

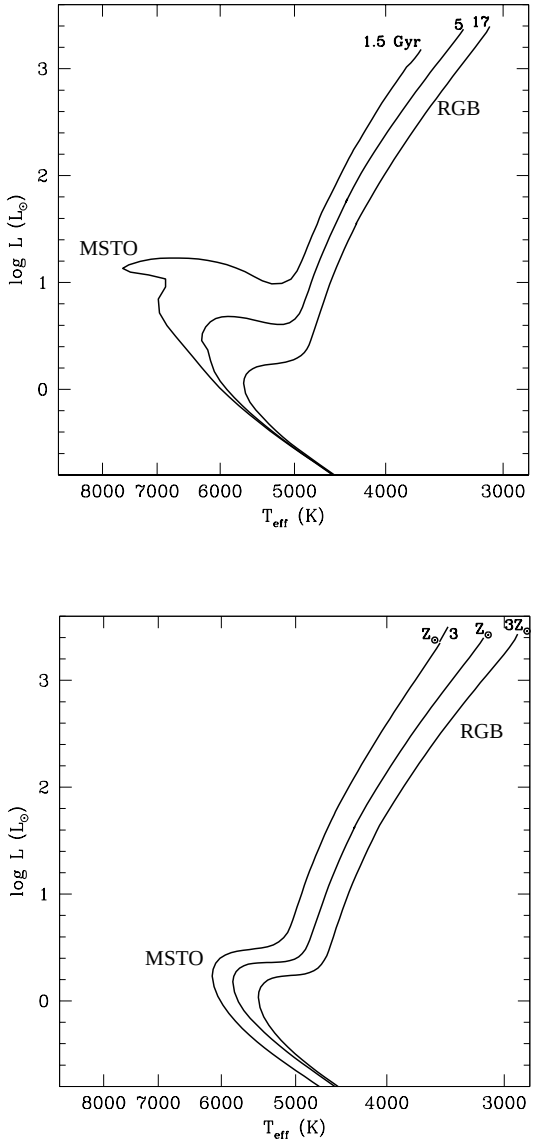


Figure 1.4: Isochrones for (above) Changing age at solar metallicity and (below) changing metallicity at an age of 12 Gyr. Figure adapted from that in a lecture by Scott C. Trager, based on data from Worthey (1994). The Main Sequence Turn Off (MSTO) and Red Giant Branch (RGB) are labelled

but in the case of integrated spectra it is necessary to fit stellar spectra to the flux values to deduce positions. Although it is possible to do this operation empirically, most modelling recipes employ theoretical fluxes calculated from stellar atmosphere models. Of atmospheres, by far the most commonly used are those of Kurucz (1979) which go into the SPEED models used here. Recent stellar modelling codes (Bruzual & Charlot, 2003) have reversed the trend, and now utilize the huge number of stellar spectra obtained for this calibration. It is unclear at this time which is the more accurate approach for making stellar models.

Absorption Line Strengths

As mentioned above, it is impossible to completely break the age-metallicity degeneracy using luminosity/temperature isochrones when individual stars cannot be resolved and accurately placed on an isochrone. The strengths of various spectral lines can be used to pin down the metallicity and identify the location of individual stars which make the lines on the isochrones, breaking the degeneracy. Rabin (1982) showed that the H-Balmer lines can probe the main sequence turn off point (MSTO, figure 1.4), and Worthey (1994) showed that metal lines such as magnesium and iron characterize the level of the red giant branch (RGB, figure 1.4).

Modelling Emission Lines

Galaxy spectra often contain emission lines. These tend to be created by hot gas, either in star forming nebulae or active galactic nuclei accretion disks. Although they are a good indicator of current star formation, they require additional modelling, and give very little information about the older star formation history of a galaxy. They are not used in this study.

1.3.2 Modelling Dust Extinction

Dust between the stellar populations and the observer cause some extinction of light. Dust grains are heated by starlight and re-emit the light at a different wavelength, typically outside of the optical range. The extinction effect is more pronounced at the UV/blue end of the spectrum, where the wavelength of the light is closest to the size of the dust grain and can be modelled by a screen which affects all populations equally. Although there are more complex dust models available (for example Charlot & Fall (2000)), we have chosen to use a one parameter dust screen model, as a compromise between parameterization complexity and processing overhead. In reality the dust will be mixed in with the stars, and different populations may not have identical dust screening.

One Parameter Dust Screen Model

In this model, the dust is assumed to be uniformly distributed in the form of a screen in front of the stellar material under study. The attenuation of the spectrum at a wavelength, λ, is given by

$$I_o(\lambda) = I_e(\lambda) \times \exp\left[-\tau(\lambda)\right] = I_e(\lambda) \times 10^{-0.4A_\lambda} \tag{1.3}$$

where I represents the intensity of the light observed and emitted, τ is the optical depth of the dust screen, and $A_\lambda = m_{\lambda,o} - m_{\lambda,e}$, the difference in magnitudes between the observed and

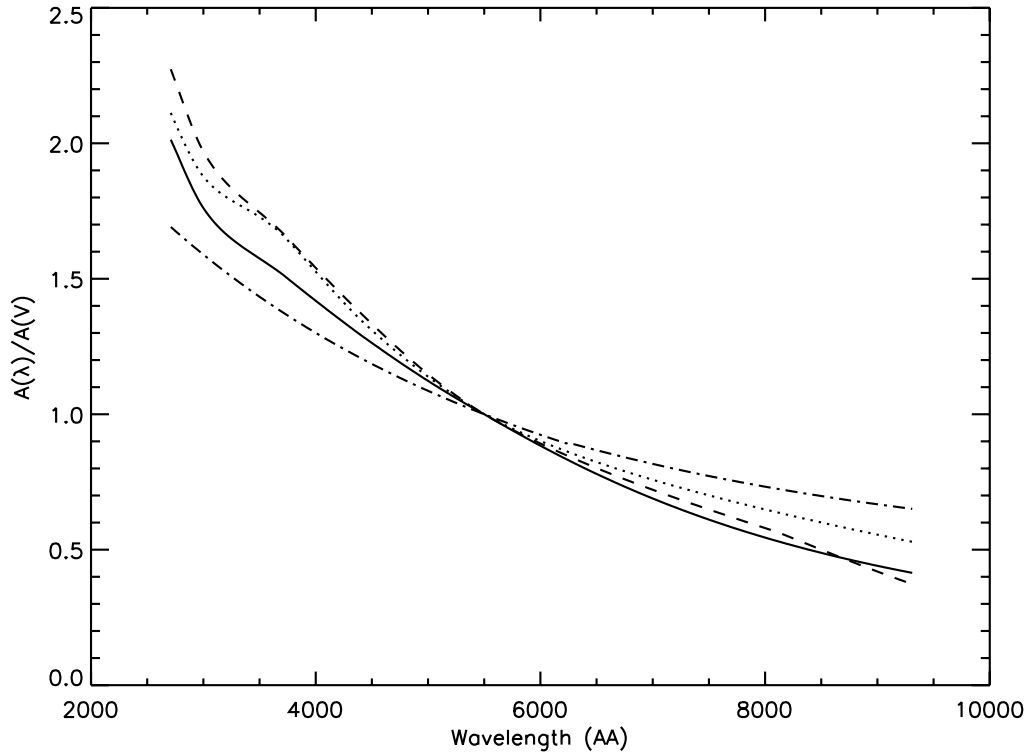


Figure 1.5: Curves giving normalised extinction, $A(\lambda)/A(V)$, for Calzetti (1997) reddining law (dash-dot) and the observed curve for the LMC (solid), LMC wing (dotted) and SMC (dashed) from Gordon et al. (2003)

emitted light. A general description of the amount of dust is the colour excess,

$$E(B-V) \quad = \quad A_B - A_V \tag{1.4}$$

The component of the optical depth which depends on the geometry of interactions with the dust particles (and which is then dependent only on wavelength) is defined as

$$k(\lambda) = \frac{A_\lambda}{E(B-V)} \tag{1.5}$$

which infers that that $k(B) - k(V) = 1$. This part of the optical depth which is wavelength independent can be quantified by the particular value of $k(\lambda = V)$, known as $R(V)$:

$$R(V) = \frac{A_V}{E(B-V)} \tag{1.6}$$

We use an observed extinction curve to give $k(\lambda)$, and modify $E(B-V)$ as the fitting parameter which is varied to allow the amount of dust extinction to change.

Two-Parameter Model

The one parameter model considers a slab of obscuring material in front of the galaxy which affects all populations equally. Charlot & Fall (2000) propose a more complex two parameter model which aims to include paramaterization only of direct relevance to the integrated properties of the galaxy. Young stars with high UV emission will ionize the HII regions around themselves. These ionized regions will have different transmission to that of the ambient ISM of the galaxy. As the stars evolve they will no longer produce the large amounts of UV required to ionize, and they will either disperse or move away from the clouds. By assuming a characteristic lifetime for the ionized regions, and the level of extinction by the ISM, the model allows a more physically accurate description of the dust extinction than a simple slab model.

Different Dust Curves

The choice of dust screen is a complex one, as there is as yet no universally applicable screen. Calzetti (1997) gives a curve which was determined by observations of starburst galaxies, but which departs from the accepted extinction in the LMC and SMC satellite galaxies (Gordon et al., 2003). The differences in these dust screens are shown in figure 1.5, and although there are differences in the normalization all three have very similar slopes in the optical range. MOPED has been tested with all three screens, and although the recovered $E(B-V)$ parameter changes, the shape of the recovered star formation history of the spectrum is robust.

1.4 The Sloan Digital Sky Survey

Previous methods of extracting information have focused on specific spectral features or broad band measures. What happens if instead of dealing with only parts of the spectrum we look at the whole spectrum across a wide range of wavelength? This question forms the main focus of this thesis, and I will start by explaining the data source for such an analysis.

The Sloan Digital Sky Survey (SDSS) is the most ambitious astronomical survey every undertaken. Eventually the imaging survey will cover almost a quarter of the night sky, with 95% completeness to a limiting r band magnitude of ~ 22.2 including some 100 million celestial objects (figure 1.7). Sources with r band magnitudes brighter than 17.77 are targeted for spectroscopic followup observations with the SDSS spectrographs, mounted on the same telescope. The resulting library of spectrophotometrically calibrated spectra will include around a million galaxy spectra, including some 100,000 quasars and a huge number of stars. The redshift distribution of the survey will allow mapping of the cosmos over a volume four times larger than has been covered before. Figure 1.6 shows the distribution of galaxies for 78275 galaxies observed before June 2003; the final survey will contain approximately ten times this number of galaxies.

1.4.1 The Observatory and Telescopes

Observations which contribute to the SDSS are taken at the Apache Point Observatory (APO), in Sunspot, New Mexico. The observatory is located at an altitude of 2800m: at this height the atmosphere contains little water vapor and few contaminants to degrade celestial images. In addition, the observatory is far away from any cities so enjoys some of the darkest skies available in

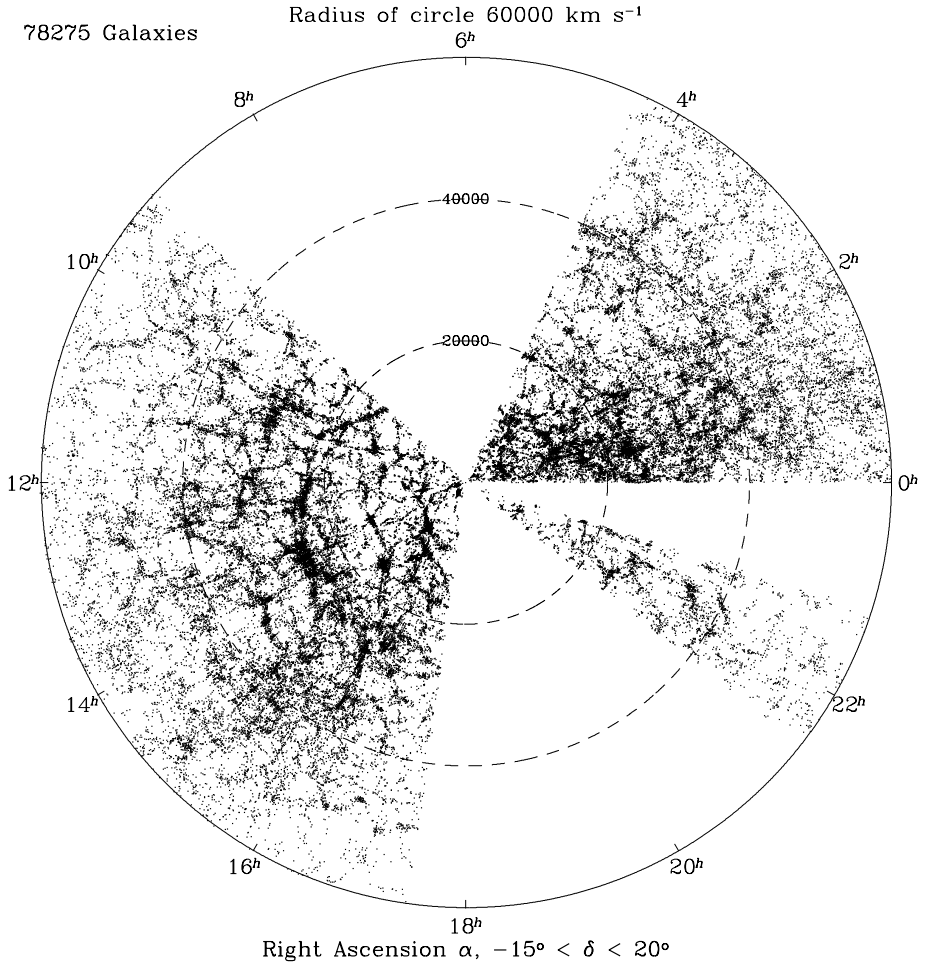


Figure 1.6: A slice through the Universe. The SDSS determines the redshift of objects from their spectra. Plotting this distance radially against RA for galaxies with $-15^\circ < \delta < 20^\circ$ gives a representation of the distribution of galaxies in the Universe. Plot taken from www.sdss.org

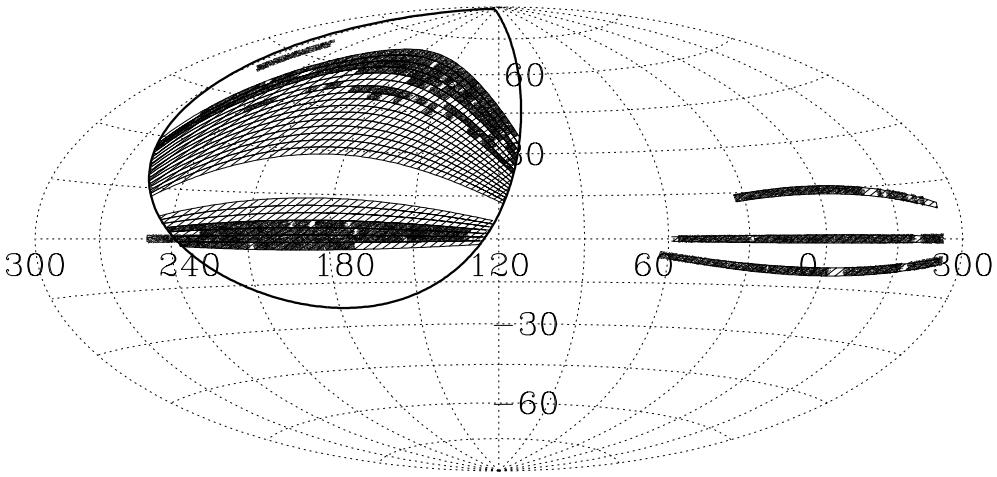


Figure 1.7: Projected SDSS Coverage. The bold line shows the original survey ellipse in the North Galactic Cap. Filled cells show spectrographic plates already observed and contained in the DR2, and open cells show the anticipated coverage with current confirmed funding. Plot taken from www.sdss.org

the USA. Together these factors allow excellent observing at the resolution required for continuous, large scale surveying. The 2.5m survey telescope is specially designed to have a wide field of view to optimize survey efficiency. The camera system used to create the imaging survey comprises 30 CCD chips, each 2″ square, mounted 2″ apart (Gunn et al., 1998). The SITe/Tektronix 2048x2048 pixel CCDs are arranged into five rows, each of which observes behind a different filter. In this way images can be obtained simultaneously in the u, g, i, r and z bands. The telescope operates in a drift scan mode, moving along great circles on the sky so that that images of objects move along the columns of the CCDs at the same rate the CCDs are being read. In addition to the photometric CCDs there are 24 CCDs placed before and after the photometric series to record astrometric data. In total the camera can produce over 200 gigabytes of data in a single night which is stored on magnetic tapes for analysis.

For spectroscopic follow up fibre are attached to the telescope via an aluminium plate in the focal plane which is machined to match the location of sources identified in the field by the imaging survey. The SDSS uses two spectrographs, with half the fibres analyzed in each. The individual spectrographs have separate blue and red channels separated by a dichroic filter, and the flux through each channel is measured by a 2048x2048 CCD. The configuration of the spectrographs allow the spectra of 640 objects to be observed at once, and on a good night the telescope can process between six and nine plates.

1.4.2 Data Processing

Traditionally, observational data is reduced image by image, with user interaction. For large surveys such as the SDSS, this process must be automated into a *pipeline*. The pipeline takes the raw imaging data and applies the usual bias subtraction, flat fielding and photometry corrections, but also selects the targets for follow up spectroscopy. For the spectroscopic data the same process occurs but the wave length calibration of the instrument must be calculated, and contamination

from any non extraterrestrial sources eliminated. The task of processing more than 200 Gb per night is immense, and rather than process the data on site it is sent to Feynman Computing Center at Fermilab on magnetic tapes. The data is passed through a number of different pipelines, which are continually evolving as understanding of the subtleties of the the observing process increases.

1.4.3 Data Releases

Although the SDSS will eventually cover a quarter of the sky, there is a wealth of information that can be gathered from the survey even at a small fraction of its final size. Although the Sloan collaborators are allowed access to the data as it is processed, a condition of the funding states that the data must also be released to the public domain within a reasonable time. So far there have been three data releases, the Early Data Release (EDR, Stoughton et al. (2002)) and Data Release 1 and 2 (DR1, Abazajian et al. (2003), DR2, Abazajian et al. (2004)). The table below shows the relative size of the current releases, and the number of galaxy spectra they contain.

Data Release	Coverage (square degrees)	Number of Galaxy Spectra
EDR	462	40,000
DR1	1360	134,000
DR2	2627	260,490

The Early Data Release gave sufficient amounts of data to be able to test techniques and give an appreciation of what might be possible with the larger data sets. Although there were several known problems in the EDR, with some caveats scientific analysis was possible. Improvements in the pipelines and the analysis they were based on led to significant corrections for the release of DR1, which included all the spectra from the EDR region reanalyzed with the new pipelines. Further, much more subtle, improvements have been applied to DR2 which again includes the reanalyzed spectra of both the EDR and DR1. The improvements are ongoing, and it is planned that the third data release, in October 2004, will contain further refinements to the data which has previously been released.

1.4.4 SDSS Galaxy Spectra

The SDSS spectrograph (York et al., 2000) observes at a resolution of 2Å until the average signal to noise is greater than 4 per pixel at g=20.2. This typically takes 3 exposures of 15 minutes each in good conditions. The resolution and quality of the spectra allow the spectroscopic pipeline to identify and measure the strength and width of up to 35 absorption lines. The positions of the lines allow redshift velocities to be estimated to 30 km/sec rms accuracy for the main galaxy sample. By combining these spectral features with the shape of the continuum and information from the photometric survey, the spectra can be categorized into different celestial object types. SDSS has 6 automated classes in the first instance, galaxies, high and low redshift quasars, stars, M stars and later, sky spectra. Approximately 1% of spectra cannot be confined to these classes and are flagged as unknown. This classification may correspond to spectra where there were instrumentation problems or possibly severe contamination from skylines.

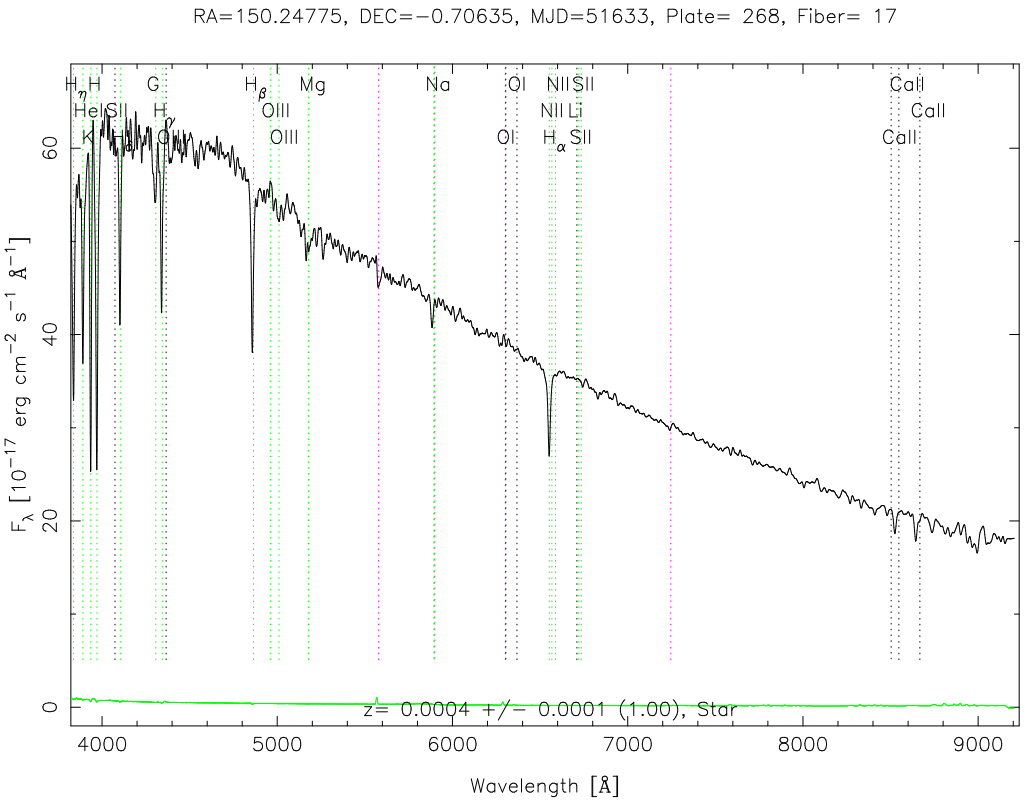


Figure 1.8: Spectrum of a star in the SDSS. The star is identified by the date of observation (MJD), Plate number (Plate) and Fibre (Fiber). The locations of lines measured by the survey are over plotted along with their identification.

Noise Characteristics

As stated earlier, the average signal to noise of each 2Å pixel is greater than 4. Unfortunately, this noise is not distributed evenly over the full spectral range. In particular, the red end of the spectrum is susceptible to contamination by sky lines caused by telluric absorption in the Earth's atmosphere. Since this atmospheric noise is at a redshift of zero it affects observations at different redshifts at different wavelengths. Spectra produced by the SDSS include the noise per pixel. Figure 1.8 shows the quality obtained by the SDSS for a typical stars and figures 1.9 and 1.10 show example spectra of old and emission line galaxies.

Spectrophotometric Calibration

The spectroscopic survey is carried out when the seeing conditions are predicted to be non-photometric (in other conditions the photometric imaging survey has priority over observing time). Since the spectra are obtained from three-arcsecond fibres special techniques must be used to spectrophotometrically calibrate the data (York et al., 2000). On each plate, 16 F8 subdwarfs are selected as standard stars, 8 in each spectrograph. These types of star have very uniform spectra, and are assumed to be comparable to a primary standard star of the SDSS, BD+17 4708. This allows the flux calibration of the spectra. Since each plate is observed at least 3 times, under different seeing conditions, and then combined, differences between exposures must be removed.

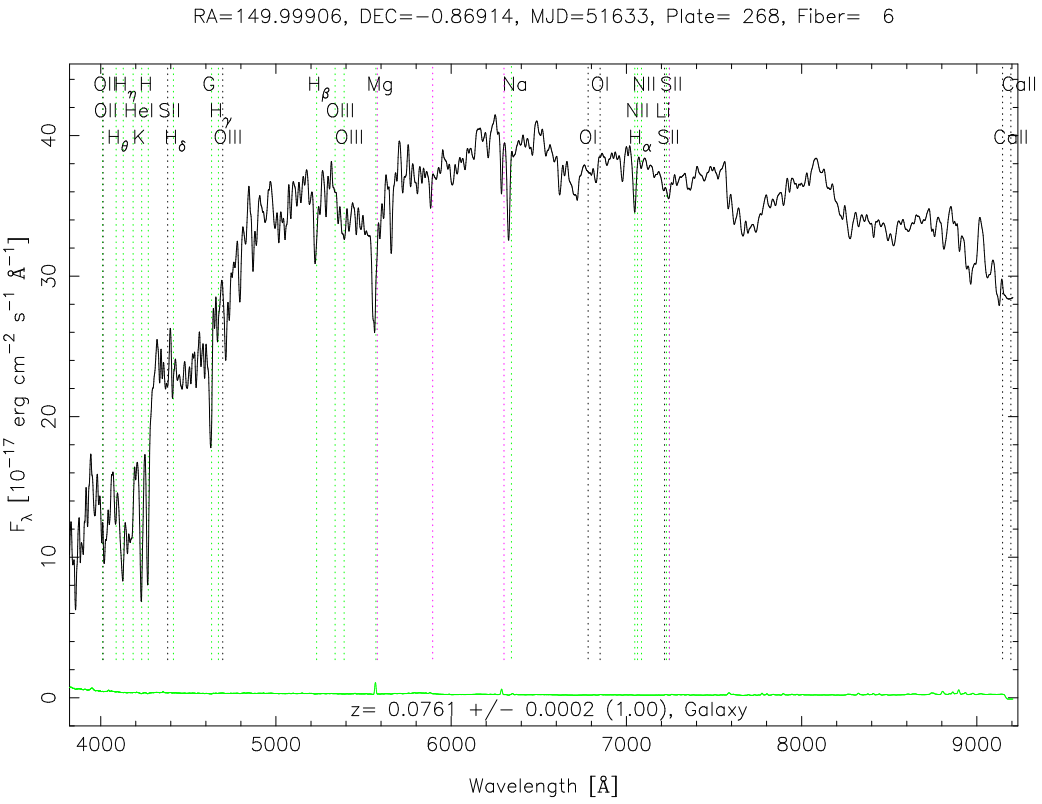


Figure 1.9: Spectrum of a galaxy in the SDSS. The 4000Å break is clear, shifted to a wavelength of $\sim$ 4300Å.

To calculate these differences short, 4 minute *smear* exposures are taken. For each of the 640 fibres the spectroscopy pipeline fits a polynomial to the ratio of the smear exposure and each science exposure. These polynomials are then used to calibrate the science exposures so that their low order spectra match those of the smear exposure. The exposure with the highest signal to noise is used to determine the flux calibration. The eigenspectrum of the 8 flux standards for each spectrograph are computed using principal component analysis, and compared to a model F8 subdwarf. The difference between these two give the flux correction required for each spectrograph.

Comparable Visible Light Surveys

Although no single survey offers the breadth and depth of data products offered by SDSS, it is certainly not the first survey to be carried out in the visible range. The most notable predecessor in the spectroscopic field is the 2dF Galaxy Redshift Survey (2dF GRS, Colless et al. (2001)). Their final data release contained some 221414 reliable redshifts for galaxies, but since the aim of the survey was the extraction of redshifts rather than exact interpretation of line strengths, spectrophotometric calibration of the spectra was not carried out. Previous to the 2dF GRS, Kennicutt (1992a) obtained high quality, spectrophotometrically calibrated spectra for 55 nearby galaxies spanning a range of morphological types. Current state of the art spectrographic surveys on the Keck, Gemini and Very Large Telescopes go much deeper than the SDSS, but none will match its coverage.

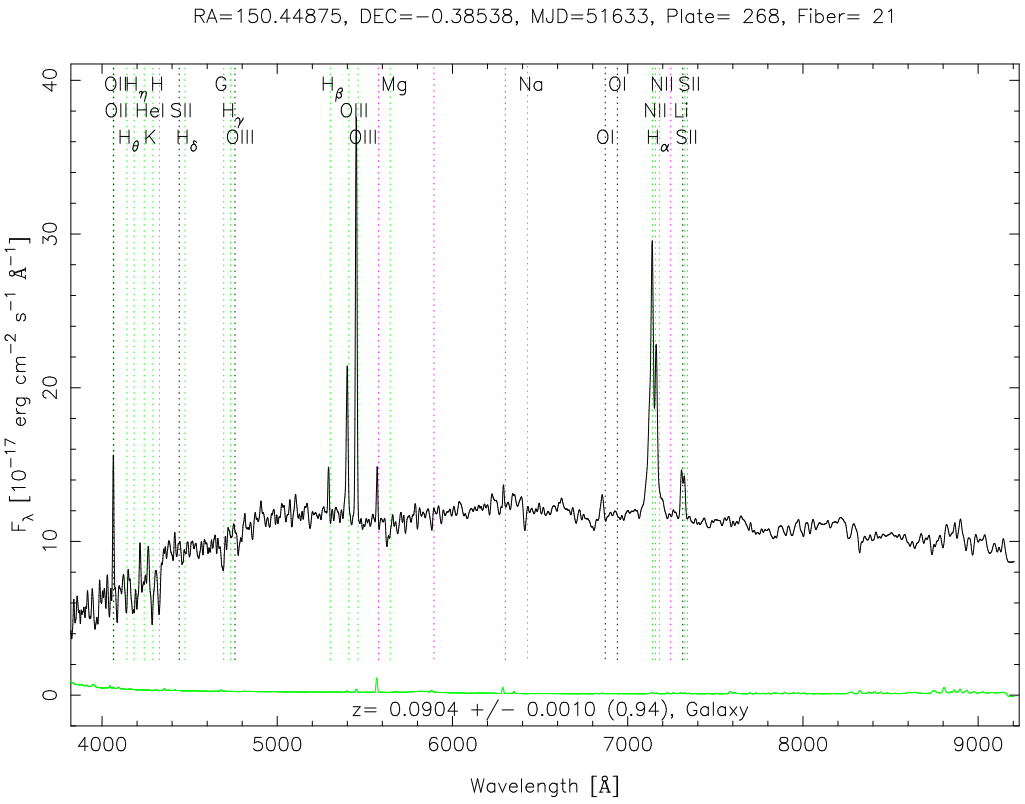


Figure 1.10: This galaxy shows strong emission lines rising to $\sim 3.5\times$ the continuum level. The broadening of the emission lines indicates the presence of an active galactic nucleus.

1.5 A Toolkit for Observational Cosmology

This section provides the basic tools required for the study of the cosmic star formation rate. It is not intended to be a full study of cosmology - there are many excellent texts on this topic already - but to explain the cosmological terms and methods employed in the rest of this text, and their derivation from the Robertson Walker Metric (RWM) and Friedmann Equation. A fuller treatment of the underlying physics can be found in Peacock (1999).

1.5.1 The Scale Factor

Hubble is credited with the first observations of a linear relationship between an object's distance, d, and recessional velocity, v. Computing cepheid distances to 24 galaxies which had redshifted spectra he suggested that

$$v = Hd \tag{1.7}$$

Hubble's estimate of H was 500 km s^{-1} Mpc^{-1} due to errors in his method for measuring distance. The latest calculations of this value put it at 71 km s^{-1} Mpc^{-1}. If the cosmological principle holds, and the Universe is indeed isotropic and homogeneous, the only solution to this equation has a universe which is expanding linearly. If the expansion is linear then mathematically all

position vectors at a time t are just scaled from their values at a reference time t_0.

$$\mathbf{x}(t) = R(t)\mathbf{x}(t_0) \tag{1.8}$$

The derivative with respect to time is

$$\dot{\mathbf{x}}(t) = \dot{R(t)}\mathbf{x}(t_0) = \frac{\dot{R(t)}}{R(t)}\mathbf{x}(t) \tag{1.9}$$

Comparing this expression to the Hubble law, equation 1.7, we can see that Hubble's constant, H, is not necessarily constant in time, in fact it is

$$H(t) = \frac{\dot{R}(t)}{R(t)} \tag{1.10}$$

1.5.2 The Robertson Walker Metric

If the cosmological principle holds, then the overall structure of time and space must be the same everywhere and in all directions. In special relativity, the time interval as experienced by an observer following a path between two events is

$$c^2 d\tau^2 = c^2 dt^2 - [dx^2 + dy^2 + dz^2]. \tag{1.11}$$

A more natural co-ordinate system for cosmology is that of spherical polar co-ordinates. The special relativity metric expressed in this system is

$$c^2 d\tau^2 = c^2 dt^2 - \left[dr^2 + r^2[d\theta^2 + \sin^2\theta d\phi^2]\right] \tag{1.12}$$

This allowed Robertson and Walker to independently derive the following metric for intervals in the Universe:

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t)[dr^2 + S_k^2(r)(d\theta^2 + \sin^2\theta d\phi^2)] \tag{1.13}$$

where $S_k(r)$ is an angular distance measure. $S_k(r) \neq r$ infers curvature of space, and for closed, flat and open universes respectively, $k = 1, 0, -1$. Note that the $R(t)$ term indicates that r,θ and ϕ are comoving coordinates.

$$S_k(r) = \begin{cases} \sin r & (k=1) \\ \sinh r & (k=-1) \\ r & (k=0) \end{cases} \tag{1.14}$$

For the remainder of this chapter I will only consider the current concordant cosmology. This suggests that the Universe is flat: $k = 0$ and the density is equal to the critical density. This allows the RWM to be expressed as

$$\boxed{c^2 d\tau^2 = c^2 dt^2 - R^2(t)[dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2]} \tag{1.15}$$

The concordant cosmological model suggests that $\Omega_{\rm m} = 0.27$, $\Omega_{\rm v} = 0.73$, $H_0 \equiv 100h$ km s^{-1}Mpc^{-1},

$h = 0.71$ (Spergel et al., 2003, references contained therein). A further implication of these results is that for times later than approximately 10^4 years after the big bang, matter dominates radiation in the Universe. For this reason the calculations below ignore the contribution of $\Omega_{\rm R}$, the contribution to the density of radiation.

1.5.3 Redshift

Consider the RWM for a photon travelling from a distant galaxy to an observer. Since the photon travels on a geodesic, $d\theta = d\phi = 0$. Since the photon travels at the speed of light the interval $ds^2 = 0$. The RWM becomes

$$\begin{aligned} 0 &= c^2dt^2 - R^2(t)dr^2 && (1.16)\\ \Rightarrow c^2dt^2 &= R^2(t)dr^2 && (1.17)\\ \Rightarrow \int_{t_{em}}^{t_{obs}} \frac{cdt}{R(t)} &= -\int_{r_{em}}^{0} dr && (1.18)\\ & && (1.19) \end{aligned}$$

If the galaxy does not move (in comoving terms), and the observer is stationary, it also true that

$$\Rightarrow \int_{t_{em}+\Delta t_{em}}^{t_{obs}+\Delta t_{obs}} \frac{cdt}{R(t)} = -\int_{r_{em}}^{0} dr \qquad (1.20)$$

Since the right hand sides of these two equations are the same they can be subtracted to give

$$\begin{aligned} 0 &= \int_{t_{em}}^{t_{em}+\Delta t_{em}} \frac{cdt}{R(t)} - \int_{t_{obs}}^{t_{obs}+\Delta t_{obs}} \frac{cdt}{R(t)} && (1.21)\\ & && (1.22) \end{aligned}$$

If we consider thin strips of the interval, such as the period of an optical photon, $\Delta t_{em} \sim 10^{-15}$ s, $R(t)$ can be assumed to be constant. In this case,

$$\begin{aligned} 0 &\approx \frac{c\Delta t_{em}}{R(t_{em})} - \frac{c\Delta t_{obs}}{R(t_{obs})} && (1.23)\\ \Rightarrow \frac{\Delta t_{obs}}{\Delta t_{em}} &= \frac{\nu_{em}}{\nu_{obs}} = \frac{R(t_{obs})}{R(t_{em})} && (1.24) \end{aligned}$$

The ratio of the frequencies is a key measurable quantity from spectra. The ratio is expressed in terms of redshift, z, defined as

$$\frac{\nu_{em}}{\nu_{obs}} = 1 + z \qquad (1.25)$$

Since

$$\begin{aligned} \lambda &= \frac{c}{\nu} && (1.26)\\ \Rightarrow z &= \frac{\Delta\lambda}{\lambda} && (1.27)\\ &= \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} && (1.28) \end{aligned}$$

So in summary,

$$\boxed{1+z=\frac{\lambda_{obs}}{\lambda_{em}}=\frac{\nu_{em}}{\nu_{obs}}=\frac{R(t_{obs})}{R(t_{em})}} \tag{1.29}$$

1.5.4 The Hubble Constant

Earlier we saw that the Hubble constant is not necessarily constant. In fact, the Hubble *constant* refers to the value of the scale factor at the present epoch, $t = t_0$,

$$H_0=\frac{\dot{R}(t_0)}{R(t_0)} \tag{1.30}$$

To determine the behaviour of $H(z)$ we must invoke the Friedmann equation, which expresses the conservation of energy;

$$\frac{1}{2}\dot{R}^2-\frac{\frac{4}{3}\pi G\rho R^3}{R} = \text{constant} \tag{1.31}$$

Consider the equation at two different times,

$$\left(\frac{dR}{dt}\right)^2-\frac{8\pi}{3}G\rho R^2 = \left(\frac{dR}{dt}\right)_0^2-\frac{8\pi}{3}G\rho_0 R_0^2 \tag{1.32}$$

$$\left(\frac{dR}{dt}\right)^2-\frac{8\pi}{3}GR^2(\rho_m+\rho_\Lambda) = \left(\frac{dR}{dt}\right)_0^2-\frac{8\pi}{3}G\rho_0 R_0^2 \tag{1.33}$$

$$\tag{1.34}$$

Since $\rho_m = \rho_{m_0}(1+z)^3$, and ρ_Λ is unchanged with the volume of the Universe,

$$\left(\frac{dR}{dt}\right)^2-\frac{8\pi}{3}GR^2(\rho_{m_0}(1+z)^3+\rho_\Lambda) = \left(\frac{dR}{dt}\right)_0^2-\frac{8\pi}{3}GR_0^2(\rho_{m_0}+\rho_{\Lambda_0}) \tag{1.35}$$

The critical density, for a flat universe, is

$$\rho_{crit}=\frac{3H_0^2}{8\pi G} \tag{1.36}$$

We use Ω to determine the fraction of the critical density, and Ω_Λ and Ω_m to be the contributions from the vacuum and matter respectively.

$$\Omega_{m_0}=\frac{\rho_{m_0}}{\rho_{crit}}=\frac{8\pi G\rho_{m_0}}{3H_0^2} \tag{1.37}$$

etc.

$$\left(\frac{dR}{dt}\right)_0^2=H_0^2R_0^2 \tag{1.38}$$

$$\left(\frac{dR}{dt}\right)^2 - [\Omega_{m_0}(1+z)^3 + \Omega_{\Lambda_0}]H_0^2R^2 = \left(\frac{dR}{dt}\right)^2 - R_0^2(\Omega_{m_0} + \Omega_{\Lambda_0})H_0^2 \quad (1.39)$$

$$\left(\frac{1}{R}\frac{dR}{dt}\right)^2 - [\Omega_{m_0}(1+z)^3 + \Omega_{\Lambda_0}]H_0^2R^2 = H_0^2\frac{R_0^2}{R^2} - H_0^2\frac{R_0^2}{R^2}\Omega_0 \quad (1.40)$$

$$(1.41)$$

$$H^2(z) = \left(\frac{1}{R}\frac{dR}{dt}\right)^2 = H_0^2\left[\Omega_{m_0}(1+z)^3 + \Omega_{\Lambda_0}\right] \quad (1.42)$$

Which, again for a flat Universe, gives us

$$\boxed{H^2(z) = H_0^2\left[\Omega_{m_0}(1+z)^3 + \Omega_{\Lambda_0}\right]} \quad (1.43)$$

1.5.5 Time and Redshift

From equation 1.43, we can see that the universe expanded faster in the past (large z) than it does today. Computation of the time-redshift relation requires integration of the differential relationship

$$dt = \frac{-dz}{(1+z)H(z)} \quad (1.44)$$

$$= \frac{-1}{(1+z)H_0}\left[\Omega_{m_0}(1+z)^3 + \Omega_{\Lambda_0}\right]^{-\frac{1}{2}}dz \quad (1.45)$$

So to determine the look back time to a given galaxy at redshift z, the following integration must be carried out (numerically):

$$\boxed{t = \int_0^{z'}\frac{1}{(1+z')H_0}\left[\Omega_{m_0}(1+z')^3 + \Omega_{\Lambda_0}\right]^{-\frac{1}{2}}dz'} \quad (1.46)$$

1.5.6 Distance Measures

Comoving Distance

In the context of an expanding Universe, distance can be ambigous. Instead, it is necessary to define distances in context. The base of most of these measures is the comoving distance, D_{com} this is a geodesic distance that remains constant over the expansion of the Universe. From equation

1.16,

$$r = \int \frac{cdt}{R(t)} \tag{1.47}$$
$$= \int \frac{cdR}{R(dR/dt)} \tag{1.48}$$
$$= \int \frac{cdR}{R^2(\frac{1}{R}\frac{dR}{dt})} \tag{1.49}$$
$$= \int \frac{dR}{R^2 H(z)} \tag{1.50}$$
$$\tag{1.51}$$

$$R(t) = \frac{R_0}{1+z} \tag{1.52}$$
$$dR = \frac{-R_0}{(1+z)^2}dz \tag{1.53}$$

Hence

$$R_0 r = \int_0^z \frac{cdz' R_0}{(1+z')^2 \left(\frac{R_o}{1+z'}\right)^2 H_0 \left[\Omega_{\Lambda_0} + \Omega_{m_0}(1+z')^3\right]^{\frac{1}{2}}} \tag{1.54}$$

and the D_{com}, the comoving distance, is

$$\boxed{D_{com} = R_0 r = R_0 \int_0^z \frac{cdz'}{H_0 \left[\Omega_{\Lambda_0} + \Omega_{m_0}(1+z')^3\right]^{\frac{1}{2}}}} \tag{1.55}$$

Luminosity Distance

If our aim is to compare flux and luminosity, we can define the luminosity distance, D_L, as

$$S = \frac{L}{4\pi D_L^2} \tag{1.56}$$
$$D_L \equiv \sqrt{\frac{L}{4\pi S}} \tag{1.57}$$

In the context of the comoving distance, the flux will be

$$S = \frac{L}{4\pi r^2 R_0^2 (1+z)^2} \tag{1.58}$$

where the $(1+z)$ terms come from the reduced arrival rate and the redshift of the photons (the flux is measured in units of energy/time). Hence

$$\boxed{D_L = R_0 r(1+z) = (1+z)\frac{c}{H_0}\int_0^z \frac{dz'}{\left[\Omega_{m_0}(1+z')^3 + \Omega_{\Lambda_0}\right]^{\frac{1}{2}}}} \tag{1.59}$$

Angular Diameter Distance

The angular diameter distance D_A is defined as the ratio of an object's physical transverse size to its angular size (in radians). If we imagine a circle at some distance r then the proper distance subtended by the angle $d\psi$ is

$$\begin{aligned} D_A d\psi &\equiv R(t_{em}) r d\psi & (1.60)\\ &= \frac{R_0 r}{1+z} d\psi & (1.61)\\ D_A &= \frac{D_{com}}{1+z} & (1.62)\\ &= \frac{D_L}{(1+z)^2}. & (1.63) \end{aligned}$$

1.5.7 Comoving Volumes

The comoving volume is that enclosed by two redshift boundaries covering of a solid angle Ω[3] on the sky.

$$\begin{aligned} dV_{com} &= R_0^3 r^2) \Delta\Omega dr & (1.64)\\ &= \Delta\Omega R_0^3 r^2 \frac{dr}{dz} dz & (1.65) \end{aligned}$$

from 1.54,

$$\begin{aligned} \frac{dr}{dz} &= -\frac{c}{H_0 R_0 \left(\frac{R_o}{1+z}\right)^2 H_0 \left[\Omega_{\Lambda_0} + \Omega_{m_0}(1+z)^3\right]^{\frac{1}{2}}} & (1.66)\\ \Rightarrow V_{com} &= \frac{c}{H_0} \Delta\Omega \int_{z_1}^{z_2} \frac{[R_0 r]^2}{\left[\Omega_{\Lambda_0} + \Omega_{m_0}(1+z')^3\right]^{\frac{1}{2}}} dz & (1.67) \end{aligned}$$

which gives the comoving volume between two redshifts z_1 and z_2 as

$$\boxed{V_{com} = \int_{z_1}^{z_2} \frac{c D_A^2}{H_0 [\Omega_{\Lambda_0} + \Omega_{m_0}(1+z')^3]^{\frac{1}{2}}} dz'} \quad (1.68)$$

[3]NB: This is Ω the solid angle, not Ω the ration between the density the critical density!

CHAPTER 2

Data Compression and the MOPED algorithm

The Massively Optimized Parameter Estimation and Data Compression (MOPED) algorithm allows rapid analysis of large datasets through a compression method which is, in most cases, lossless. This chapter contains a summary of the technique developed in Tegmark et al. (1997); Heavens et al. (2000); Reichardt et al. (2001) and also explains the Interactive Data Language (IDL) implementation of the code, originally written by Raul Jimenez, Christian Reichardt and Alan Heavens. Although not completely rewritten from scratch, the code has been very significantly modified by myself. It is now capable of running a variety of models on data with multiple wavelength ranges for calculations at a range of redshifts with trivial modifications to setup files. It is also faster by a factor of ~ 5 due to precalculation optimization and a Markov Chain Monte Carlo (MCMC) hypersurface minimization technique. This section covers the MOPED algorithm and the incorporation of the algorithm into a system of programs which can interpret the spectra of galaxies. I also present some examples of SFH recovered by MOPED for individual galaxies.

2.1 The MOPED Algorithm

2.1.1 An Introduction to Data Compression

Fundamentally, the ideal goal of data compression techniques for parameter estimation is to recover information from a compressed dataset that has the same quality as the information which could be recovered from the original, full dataset. MOPED assumes that the data, written as a vector, consists of signal, μ, and noise, $\mathbf{n}$, components

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{n}. \tag{2.1}$$

If we assume that the noise causes the data to be randomly distributed around the signal with $\langle \mathbf{n} \rangle = \mathbf{0}$, it follows that,

$$\langle \mathbf{x} \rangle = \boldsymbol{\mu}. \tag{2.2}$$

If noise is assumed to be Gaussian, it can be completely described by the covariance matrix, $\mathbf{C}$, with components $C_{ij} = \langle n_i n_j \rangle$. For the remainder of this section, the indices i, j refer to data pixels and α, β to parameters. I first consider general linear compression of the data vector $\mathbf{x}$, then focus on the particular compression used by MOPED. This compression is designed to retain as much information (ideally all) about the galaxy parameters (eg. SFH, metallicity, dust) as possible.

Linear Combinations of Data

Let $\mathbf{y}$ be a vector of linear combinations of the data, produced by multiplying the data vector $\mathbf{x}$ by the matrix of weighting vectors $\mathbf{B}$. Using this nomenclature we can describe the data, its mean and its variance:

$$\begin{aligned}
\mathbf{y}_i &= \mathbf{B}_{ij}\mathbf{x}_j && (2.3)\\
\langle \mathbf{y}_i \rangle &= \mathbf{B}_{ij}\langle \mathbf{x}_j \rangle, \langle \mathbf{y} \rangle = \mathbf{B}\boldsymbol{\mu} && (2.4)\\
\langle (\mathbf{y}_i - \bar{\mathbf{y}_i})(\mathbf{y}_j - \bar{\mathbf{y}_j}) \rangle &= \langle (\mathbf{B}_{il}\mathbf{x}_l - \mathbf{B}_{il}\boldsymbol{\mu}_l)(\mathbf{B}_{jk}\mathbf{x}_k - \mathbf{B}_{jk}\boldsymbol{\mu}_k) \rangle && (2.5)\\
&= \mathbf{B}_{il}\mathbf{B}_{jk}[\langle \mathbf{x}_l\mathbf{x}_k \rangle + \boldsymbol{\mu}_l\boldsymbol{\mu}_k - \langle \mathbf{x}_l \rangle \boldsymbol{\mu}_k - \boldsymbol{\mu}_l \langle \mathbf{x}_k \rangle] && (2.6)\\
&= \mathbf{B}_{il}\mathbf{B}_{jk}[\langle \mathbf{x}_l\mathbf{x}_k \rangle + \boldsymbol{\mu}_l\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\boldsymbol{\mu}_k] && (2.7)\\
&= \mathbf{B}_{il}\mathbf{B}_{jk}\mathbf{C}_{lk} = \mathbf{B}_{il}\mathbf{C}_{lk}\mathbf{B}^t_{kj} && (2.8)
\end{aligned}$$

$\mathbf{B}_{il}\mathbf{C}_{lk}\mathbf{B}^t_{kj}$ will appear many times in the following description, and from here onwards will be written $\mathbf{B}\mathbf{C}\mathbf{B}^t$. $\mathbf{C}$ is the covariance matrix of the data $\mathbf{x}$:

$$\mathbf{C}_{lk} = \langle (\mathbf{x}_l - \boldsymbol{\mu}_l)(\mathbf{x}_k - \boldsymbol{\mu}_k) \rangle = \langle \mathbf{x}_l\mathbf{x}_k \rangle - \boldsymbol{\mu}_l\boldsymbol{\mu}_k. \tag{2.9}$$

These definitions are valid for any weighting vector. In the special case of $\mathbf{B}$ being square and non-singular, it is possible to reconstruct the data completely from the $\mathbf{y}$ vector since $\mathbf{B}$ is invertible - although it is the same length as $\mathbf{x}$ so there is zero compression. Although reducing the length of $\mathbf{y}$ will reduce the size of the dataset, it will not necessarily lose information about para-

meters which may be extracted from those data. Since we are more interested in the information that can be extracted from the data than the data itself, we can use the Fisher matrix to determine how much *information* we lose when we use some combinations of the data determined by $\mathbf{B}$.

2.1.2 The Fisher Matrix

If we consider the likelihood distribution of a set of parameters θ to be estimated from a set of data $\mathbf{x}$ before and after compression, we would like the distributions around the peak to be identical for the two cases. The Fisher matrix describes the shape of the likelihood surface at its peak. In the simple one dimensional form the peak is assumed to be Gaussian, an inverted parabola in log space. The Fisher matrix describes the curvature of the maximum - in essence how well the parameter can be determined from the data available. By comparing the Fisher matrix before and after compression, we can examine the loss of information about parameters determined from the data. The Fisher matrix is defined to be

$$\mathbf{F}_{\alpha\beta} \equiv -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle . \tag{2.10}$$

The likelihood $\mathcal{L}$ above is the probability of data $\mathbf{x}_i$ given a set of model parameters $\boldsymbol{\theta}$. It is necessary to use Bayes' theorem to show that this is proportional to the probability of the parameters given the data.

$$\mathrm{P}(\theta_\alpha|\mathbf{x}_i) = \frac{\mathrm{P}(\theta_\alpha, \mathbf{x}_i)}{\mathrm{P}(\mathbf{x}_i)}, \tag{2.11}$$

where $\mathrm{P}(\theta_\alpha, \mathbf{x}_i)$ is the joint distribution of $\theta, \mathbf{x}$. This can be written as

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\mathrm{P}(\boldsymbol{\theta}, \mathbf{x})}{\mathrm{P}(\mathbf{x})} \tag{2.12}$$

$$= \frac{\mathrm{P}(\mathbf{x}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{x})}. \tag{2.13}$$

In this case, $\mathrm{P}(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood, $\mathrm{P}(\boldsymbol{\theta})$ the prior and $\mathrm{P}(\mathbf{x})$ the evidence. If we assume that the prior is uniform and note that the evidence is not dependent on $\boldsymbol{\theta}$, then

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{x}) \propto \mathrm{P}(\mathbf{x}|\boldsymbol{\theta}). \tag{2.14}$$

Hence the probability of the parameters is, with uniform priors, proportional to the likelihood, or

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) = P(\mathbf{x}|\boldsymbol{\theta}). \tag{2.15}$$

If we assume the noise is Gaussian distributed, then for 1 datum ($\mathbf{x}_1$),

$$\mathcal{L} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}(x_1 - \mu_1)\mathbf{C}_{11}^{-1}(x_1 - \mu_1)\right]. \tag{2.16}$$

Since $\mathbf{C}_{11}$ has only one element, we can simplify to

$$\mathcal{L} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}(x_1-\mu_1)^2/\sigma_1^2\right]. \tag{2.17}$$

To generalize to n independent data $(\mathbf{x}_1, \ldots \mathbf{x}_n)$ we must multiply the probabilities. In the general case,

$$\mathcal{L} = \frac{1}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C})}} \times \exp\left[-\frac{1}{2}\sum_{i,j}(x_i-\mu_i)\mathbf{C}_{ij}^{-1}(x_j-\mu_j)\right]. \tag{2.18}$$

Which simplifies in the uncorrelated case to

$$\mathcal{L} = \frac{1}{\left(\sqrt{2\pi}\right)^n \sigma_1 \ldots \sigma_n} \times \exp\left[-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^2/\sigma_1^2 - \ldots - \frac{1}{2}(x_n-\mu_n)^2/\sigma_n^2\right], \tag{2.19}$$

since

$$\mathbf{C}_i = \langle (x_i-\mu_i)(x_j-\mu_j)\rangle = \sigma_i^2\delta_{ij} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{pmatrix} \tag{2.20}$$

$$\Rightarrow \det \mathbf{C} = \sigma_1^2\sigma_2^2\ldots\sigma_n^2 \tag{2.21}$$

and equation 2.18 still holds if data are correlated, in which case C is has non-zero off-diagonal terms.

Marginal and Conditional Errors

When recovering parameters and their errors from a likelihood surface, it is usual to state either the marginal or conditional error. The marginal error is the extent of the peak if all points are projected onto the axis of the parameter in question and gives a representation of the full spread of the peak. The conditional error is subtly different in that it assumes that all other parameters are fixed, and that the range of error on the parameter in question is only the width of the peak at the value of the other parameters which determine the surface. In the case of uncorrelated errors these errors are the same, as illustrated in figure 2.1

When the errors on the parameters are correlated (ie the parameters have some covariance) then the two errors are not the same. The errors are illustrated for a two dimensional likelihood surface considering two parameters θ_a and θ_b in figure 2.2. The marginal error on the distribution is obviously larger than the conditional in this case.

In the context of the Fisher matrix, we can describe the conditional error by approximating $\mathcal{L}$ by a gaussian (as a function of θ_α at fixed $\theta_\beta, \beta \neq \alpha$),

$$\langle \ln \mathcal{L}\rangle \simeq \ln\left[\mathcal{L}_0 \exp -\frac{(\theta_\alpha - \hat{\theta}_\alpha)^2}{2\sigma_\alpha^2}\right], \tag{2.22}$$

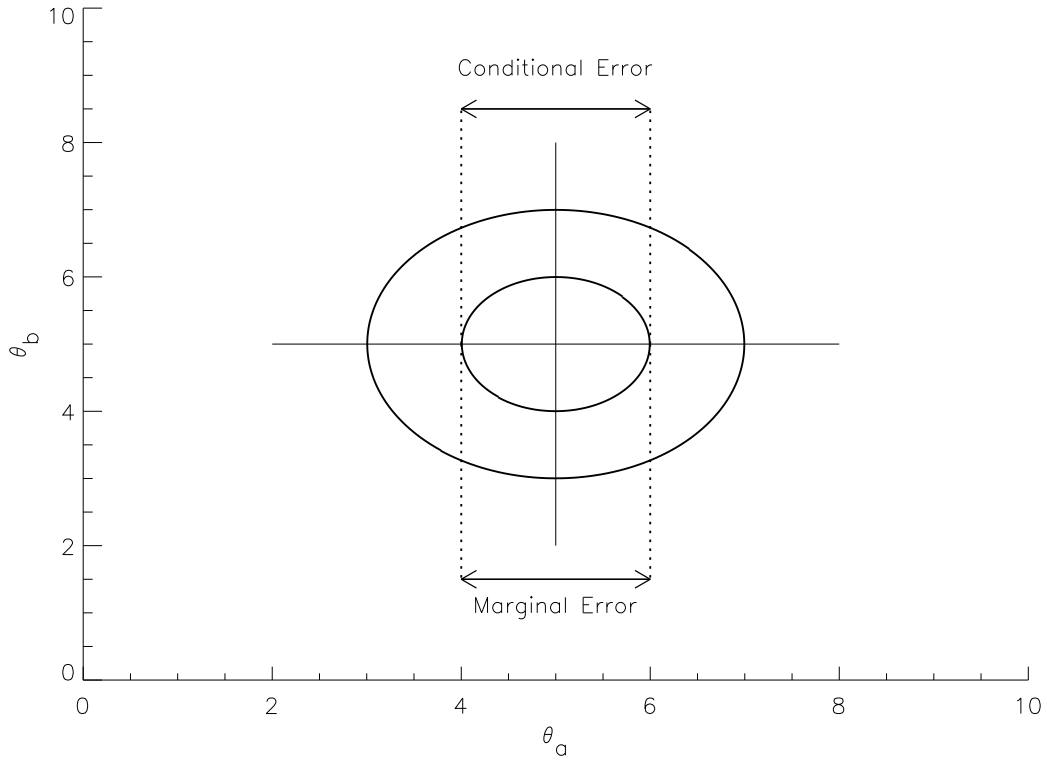


Figure 2.1: Marginal and conditional errors on a two dimensional likelihood surface peak considering two parameters θ_a and θ_b where errors are uncorrelated.

where $\mathcal{L}_0$ is the likelihood at the peak and $\hat{\theta}_\alpha$ is the parameter at the peak. From the Fisher matrix,

$$\mathbf{F}_{\alpha\alpha} = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_\alpha^2} \right\rangle = \frac{\partial^2}{\partial \theta_\alpha^2}\left[\frac{(\theta_\alpha - \hat{\theta}_\alpha)^2}{2\sigma_\alpha^2}\right] = \frac{1}{\sigma_\alpha^2} \tag{2.23}$$

$$\sigma_\alpha = \frac{1}{(\mathbf{F}_{\alpha\alpha})^{\frac{1}{2}}}. \tag{2.24}$$

The solution is a little more difficult when considering marginal errors. The marginal error concerns the distribution of the parameter under study over the distribution of all other points,

$$\mathrm{P}(\boldsymbol{\theta}_\alpha) = \int_{\text{excl. } \theta_\alpha} d\theta_1 \ldots d\theta_n \mathrm{P}(\boldsymbol{\theta}|\mathbf{x}) \tag{2.25}$$

$$\propto \int_{\text{excl. } \theta_\alpha} d\theta_1 \ldots d\theta_n \mathcal{L}(\mathbf{x}|\boldsymbol{\theta}). \tag{2.26}$$

$$\tag{2.27}$$

Integrating the distributions gives

$$\sigma_\alpha^{\text{marginal}} = [(\mathbf{F}^{-1})_{\alpha\alpha}]^{\frac{1}{2}}. \tag{2.28}$$

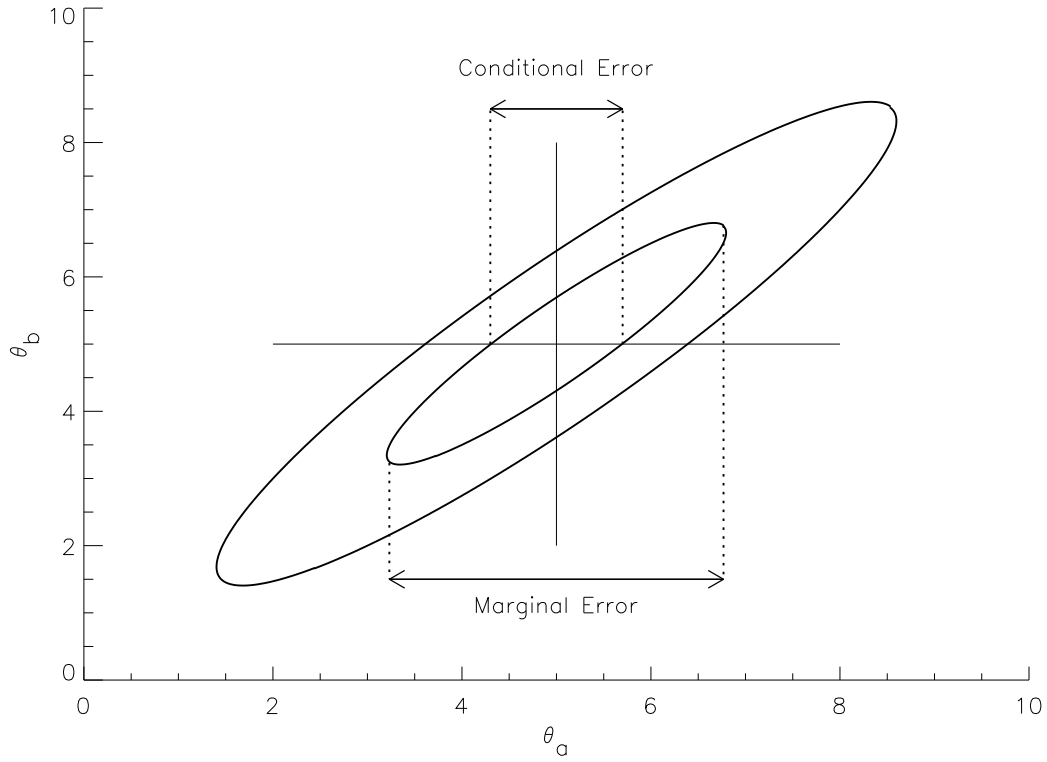


Figure 2.2: Marginal and conditional errors on a two dimensional likelihood surface peak considering two parameters θ_a and θ_b where the errors are correlated.

Obviously, if there is any chance of the estimates being correlated, we must use the marginal errors. What is important is that these marginal errors are still determined from the Fisher matrix.

Comparing information in compressed and uncompressed data

In this section, I show how to choose the first weighting vector of the data. This is constructed to minimize the conditional error on one parameter. It will turn out that the algorithm is considerably more powerful than this modest constraint suggests.

Equation 15 of Tegmark et al. (1997) gives the Fisher matrix for Gaussian distributed data in terms of what is measurable:

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[\mathbf{A}_\alpha \mathbf{A}_\beta + \mathbf{C}^{-1}\mathbf{M}_{\alpha\beta}\right]. \tag{2.29}$$

In this equation, $\mathbf{M}_{\alpha\beta} \equiv 2\boldsymbol{\mu}_{,/alpha}\boldsymbol{\mu}_{,/beta}$ and $\mathbf{A}_\alpha \equiv \mathbf{C}^{-1}\mathbf{C}_{,\alpha}$, where the comma indicates the derivative with respect to $\boldsymbol{\theta}_\alpha$.We now assume that $\mathbf{C}$ is independent of θ_α - this is the main restriction on MOPED. The Fisher matrix can be written

$$\mathbf{F}_{\alpha\beta} = \mathrm{Tr}\left[\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}\right]. \tag{2.30}$$

Where Tr indicates Trace, the sum of the diagonal terms of the matrix. We now have two equa-

tions which describe the Fisher matrix - one for the full dataset and one for the compressed dataset.

$$\begin{aligned} \mathbf{F}_{\alpha\beta} &= \mathrm{Tr}\left[\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}\right] && -\mathrm{Data} = \mathbf{x} \\ &= \mathrm{Tr}\left[(\mathbf{B}\mathbf{C}\mathbf{B}^{\mathrm{t}})^{-1}(\mathbf{B}\boldsymbol{\mu}_{,\alpha})^{\mathrm{t}}(\mathbf{B}\boldsymbol{\mu}_{,\beta})\right] && -\mathrm{Data} = \mathbf{y}. \end{aligned} \tag{2.31}$$

Considering the case where $\mathbf{B}$ is square and $\mathbf{B}^{-1}$ exists, we would expect the compressed Fisher element to be exactly the same as the uncompressed. This is readily shown:

$$\begin{aligned} \mathbf{F}_{\alpha\beta} &= \mathrm{Tr}\left[(\mathbf{B}\mathbf{C}\mathbf{B}^{\mathrm{t}})^{-1}(\mathbf{B}\boldsymbol{\mu}_{,\alpha})(\mathbf{B}\boldsymbol{\mu}_{,\beta})\right] && (2.32)\\ &= \mathrm{Tr}\left[(\mathbf{B}\boldsymbol{\mu}_{,\alpha})^{\mathrm{t}}(\mathbf{B}\mathbf{C}\mathbf{B}^{\mathrm{t}})^{-1}\mathbf{B}\boldsymbol{\mu}_{,\beta}\right] && (2.33)\\ &= \mathrm{Tr}\left[\boldsymbol{\mu}^{\mathrm{t}}_{,\alpha}\mathbf{B}^{\mathrm{t}}[(\mathbf{B}^{\mathrm{t}})^{-1}\mathbf{C}^{-1}\mathbf{B}^{-1}]\mathbf{B}\boldsymbol{\mu}_{,\beta}\right] && (2.34)\\ &= \mathrm{Tr}\left[\boldsymbol{\mu}^{\mathrm{t}}_{,\alpha}\mathbf{C}^{-1}\boldsymbol{\mu}_{,\beta}\right] && (2.35)\\ &= \mathrm{Tr}\left[\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}_{,\beta}\right]. && (2.36) \end{aligned}$$

In general 2.31 gives the marginal errors for any linear combination of the data, including when $\mathbf{B}$ is not square and the data are compressed.

2.1.3 MOPED Compression

If the matrix $\mathbf{B}$ consists of only one row we can write it as $\mathbf{b}$, a vector. In this case $\mathbf{B}\mathbf{C}\mathbf{B}^t$ is $\mathbf{b}\mathbf{C}\mathbf{b}^t$, the product of a vector, a square matrix and a vector - a number. Putting this into the equation above,

$$\mathbf{F}_{\alpha\beta} = \left[\frac{(\mathbf{b}.\boldsymbol{\mu}^t_{,\alpha})(\mathbf{b}.\boldsymbol{\mu}^t_{,\beta})}{\mathbf{b}\mathbf{C}\mathbf{b}^t}\right]. \tag{2.37}$$

Earlier the conditional error on θ_α was determined to be $\mathbf{F}^{-\frac{1}{2}}_{\alpha\alpha}$. To minimize this error we need to maximize $F_{\alpha\alpha}$;

$$\mathbf{F}_{\alpha\alpha} = \left[\frac{(\mathbf{b}.\boldsymbol{\mu}_{,\alpha})^2}{(\mathbf{b}\mathbf{C}\mathbf{b}^t)}\right]. \tag{2.38}$$

Maximizing $F_{\alpha\alpha}$ subject to the condition that $\mathbf{b}\mathbf{C}\mathbf{b}^t = \mathrm{constant} = 1$, can be achieved with a Lagrange multiplier Λ:

$$\frac{\partial}{\partial \mathbf{b}_i}\left([\mathbf{b}_j(\boldsymbol{\mu}_{,\alpha})_j]^2 - \lambda \mathbf{b}_j\mathbf{C}_{jl}\mathbf{b}_l\right) = 0 \tag{2.39}$$

$$2[\mathbf{b}_j(\boldsymbol{\mu}_{,\alpha})_j] - \lambda\mathbf{C}_{il}\mathbf{b}_l - \lambda\mathbf{b}_j\mathbf{C}_{ji} = 0. \tag{2.40}$$

since $\mathbf{C}$ is symmetric, $\mathbf{C}_{ij} = \mathbf{C}_{ji}$.

$$\Rightarrow \mathbf{b}_j \underbrace{(\boldsymbol{\mu}_{,\alpha})^t_j (\boldsymbol{\mu}_{,\alpha})_i}_{\mathbf{M}_{ij}} = \lambda \mathbf{C}_{il}\mathbf{b}_l \tag{2.41}$$

$$\mathbf{M}_{ij}\mathbf{b}_j = \lambda \mathbf{C}_{il}\mathbf{b}_l \tag{2.42}$$

$$\mathbf{M}\mathbf{b} = \lambda \mathbf{C}\mathbf{b}. \tag{2.43}$$

This is a generalized eigenvalue problem. Since $\mathbf{C}$ is by definition a square and symmetric matrix, it can be diagonalized. The resultant matrix would contain the variances of the rotated set, which must all be positive (the eigenvalues along the diagonal are the variances of the now orthogonal data). Since the eigenvalues do not change under rotation, the matrix is positive definite (all terms must be greater than zero). Such problems can be solved using the "Cholesky decomposition", $\mathbf{C} = \mathbf{L}\mathbf{L}^t$, where $\mathbf{L}$ is a lower triangular matrix.

$$\mathbf{M}\mathbf{b} = \overbrace{\lambda \mathbf{L}\mathbf{L}^t}^{\mathbf{C}}\mathbf{b} \tag{2.44}$$

$$\mathbf{L}^{-1}\mathbf{M}\mathbf{b} = \lambda \mathbf{L}^t\mathbf{b} \tag{2.45}$$

$$\mathbf{L}^{-1}\mathbf{M}((\mathbf{L}^t)^{-1}\mathbf{L}^t)\mathbf{b} = \lambda \underbrace{\mathbf{L}^t\mathbf{b}}_{\mathbf{f}} \tag{2.46}$$

$$\underbrace{\left[\mathbf{L}^{-1}\mathbf{M}(\mathbf{L}^t)^{-1}\right]}_{\mathbf{N}} = \lambda \mathbf{f} \tag{2.47}$$

$$\mathbf{N}\mathbf{f} = \lambda \mathbf{f}. \tag{2.48}$$

This is an ordinary eigenvalue problem, and can be expanded to

$$(\mathbf{L}^{-1}\boldsymbol{\mu}_{,\alpha})(\boldsymbol{\mu}_{,\alpha}\underbrace{(\mathbf{L}^t)^{-1})\mathbf{L}^t}_{\mathbf{I}}\mathbf{b} = \lambda \mathbf{L}^t\mathbf{b} \tag{2.49}$$

$$(\quad = \quad \text{scalar})$$

$$\Rightarrow \mathbf{L}^t\mathbf{b} \propto \mathbf{L}^{-1}\boldsymbol{\mu}_{,\alpha} \tag{2.50}$$

and setting λ so that $\mathbf{b}\mathbf{C}\mathbf{b}^t = 1$ as required:

$$\mathbf{b} = \text{const.}(\mathbf{L}^{\mathrm{t}})^{-1}\mathbf{L}^{-1}\boldsymbol{\mu}_{,\alpha} \tag{2.51}$$

$$\mathbf{b}_1 = \text{const.}\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha} \tag{2.52}$$

$$\tag{2.53}$$

which, when normalised, gives the solution

$$\mathbf{b}_1 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}^t_{,1}\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}}. \tag{2.54}$$

Putting this solution into our original compression equation **??**, we get a compressed data set which consists of only one number, y_1, such that

$$y_1 = \mathbf{b}_1.\mathbf{x}. \tag{2.55}$$

In addition, the solution appears to work as expected in the simple uncorrelated case, giving high weight to those data which are most sensitive to the parameters (large changes in $\boldsymbol{\mu}$) and low weight to those that are noisy (small C^{-1} elements). To determine how good the compression is, we need to compare the Fisher matrix elements for the complete and compressed data sets. Substituting $\mathbf{b}_1$ into equation 2.37 gives

$$\mathbf{F}_{11} = \boldsymbol{\mu}_{,1}^t \mathbf{C}^{-1} \boldsymbol{\mu}_{,1} \tag{2.56}$$

which, given $\mathbf{C}$ is independent of θ, is identical to the Fisher matrix element when using the full data set. This means that the compression from the whole dataset to a single number is in fact lossless – y_1 loses no information about θ_1.

Caveat

It is important to realize that the MOPED algorithm allows information to be recovered from an optimally compressed dataset - it is perhaps most accurately described as an information compression method. Its purpose is not to recover the initial data, but to recover parameters which depend on that initial data. By comparing the Fisher matrix of the data before and after compression, it is shown that the method can be lossless in terms of *information* although this is not the same as lossless in terms of *data*.

Two parameters

The method outlined above shows how one parameter can be extracted from a dataset. Although this is useful, a more common application would be the calculation of several parameters simultaneously. In order to perform such an analysis a second number, y_2, is used to encapsulate the information from the data about the second parameter, θ_2. In a similar manner to y_1,

$$y_2 \equiv \mathbf{b}_2^t \mathbf{x}. \tag{2.57}$$

By construction, y_2 is chosen to be uncorrelated with y_1 – this condition is satisfied if $\mathbf{b}_2^t \mathbf{C} \mathbf{b}_1 = 0$. It must also contain as much information as possible about θ_2. As explained in Heavens et al. (2000), this requires two Lagrange multipliers ($\mathbf{b}_2$ is normalized by setting $\mathbf{b}_2^t \mathbf{C} \mathbf{b}_2 = 1$). This gives

$$\mathbf{b}_2 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,2} - (\boldsymbol{\mu}_{,2}^t \mathbf{b}_1)\mathbf{b}_1}{\sqrt{\boldsymbol{\mu}_{,2}^t \mathbf{C}^{-1} \boldsymbol{\mu}_{,2} - (\boldsymbol{\mu}_{,2}^t \mathbf{b}_1)^2}}. \tag{2.58}$$

It is possible to generalize this solution to M parameters - this requires M orthogonal vectors $\mathbf{b}_m, m = 1, \ldots M$, with each y_m containing as much information as possible about its corresponding θ_M not already contained in y_q, $q < m$. This is essentially the Gram-Schmidt orthogonalisation modified by using $\mathbf{C}$ as the curved metric tensor to give the general solution:

$$\mathbf{b}_m = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1} (\boldsymbol{\mu}_{,m}^t \mathbf{b}_q)\mathbf{b}_q}{\sqrt{\boldsymbol{\mu}_{,m}^t \mathbf{C}^{-1} \boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1} (\boldsymbol{\mu}_{,m}^t \mathbf{b}_q)^2}}. \tag{2.59}$$

The remarkable features of this solution are that

- The data compression can be massive
- There are as many y values as parameters
- The Fisher matrix is unchanged by the data compression

given the constraints that

- The errors on the spectrum are Gaussian
- $\mathbf{C}$ is independent of $\boldsymbol{\theta}$

A proof that this method is lossless for many parameters is given in the appendix of Heavens et al. (2000). It is important to note that the b-vectors are not unique: they will depend on the order of the parameters. It was shown in Reichardt et al. (2001) that this did not influence recovered parameters.

2.1.4 Applying MOPED to Galaxy Spectra

MOPED is ideally suited for determination of stellar populations from galaxy spectra. Stellar spectra are well understood, and by combining several models together and compensating for dust a galaxy spectrum can be formed. In a traditional χ^2 analysis, each combination of spectra examined would require a calculation proportional to the number of pixels in the spectrum. If MOPED is used the likelihood for a combination could be determined from the compressed data set without losing any information about the parameters (in this case the parameters are the relative fractions of the various populations and their metallicities). Instead of calculations involving, say, 4000 flux points, MOPED would allow calculations involving only the number of parameters - typically 20 or so. This yields a speed up of ~ 200 times when estimating the likelihood given a set of parameters.

Crucial to this is the question of how well the algorithm behaves when dealing with real data. Noise on galaxy spectra has several sources - at least one of which, the photon counting noise, is parameter dependent. In this case, $\mathbf{C}_{ii} \propto \boldsymbol{\mu}_i$. Extensive testing in (Heavens et al., 2000; Reichardt et al., 2001) has shown that in fact the this method still recovers parameters exceptionally well, with degradation of errors by $\sim 1\%$ being typical.

2.1.5 The Fiducial Model

To construct the weighting vectors some fiducial model is required. It is important to note that even if the fiducial model is wrong $\mathbf{y}$ values are still unbiased (although not optimal, and not lossless) - and the likelihood analysis for each y_m is still correct. If desired, one could iterate, replacing the fiducial model with the best fitting parameters until the solutions converged. (Reichardt et al., 2001) showed that this is unnecessary - in fact the data compression is still very nearly lossless and extremely powerful when applied to spectral recovery. This means that the calculation of the b-vectors only has to occur once for a given galaxy. In fact, Panter et al. (2003) shows that a small set of b-vectors can even be used to cover all galaxies over a range of redshifts if the data is rebinned.

2.1.6 Parametrization of Spectra

MOPED gives a method for estimating a set of parameters from data. To operate the compression, a set of parameters must be defined that define the shape of the spectrum and the model's dependence on these parameters measured. In the past, the SFH of galaxies has been modelled by an exponential decay with a single parameter - for more complex models one or two bursts of formation are allowed. In fact, it would be better not to put any such constraints on star formation, particularly considering that each galaxy may have (as a result of mergers) several distinctly different aged populations. Star formation takes place in giant molecular clouds, which have a lifetime of around 10^7 years. Splitting the history of the Universe into the lifetimes of these clouds give a natural unit of time for star formation analysis, but unfortunately it would require several thousand of such units to map the age of galaxies formed 13 billion years ago. A different approach considers the change in the models over time. The problem is constrained by the time taken to perform the analysis, so the first step is to pick a number of wider bins which, taken together, will map the history of star formation in the subject galaxy. The boundaries between the different bins are then determined by considering bursts of star formation at the beginning and end of each period (at a fixed metallicity) and set the boundaries such that the fractional difference in the final spectrum is the same for each bin. In the case of a galaxy whose age is 14 Gyr and a desired resolution of 8 bins, the optimal bin boundaries will be at 0.03, 0.23, 0.42, 0.66, 0.92, 2.56, 6.33 and 14 Gyr (considering fixed solar metallicity). These boundaries are almost equally spaced in logarithmic space - for the case above the bin boundaries would fall at 0.02, 0.05, 0.13, 0.33, 0.84, 2.15, 5.49 and 14.0 Gyr. It is easier to calculate, and later analyze, these equally spaced logarithmic bins, so that is the parametrization chosen for MOPED. It is quite possible that the metallicity of the gas which goes to make up these differently aged populations varies, and for simplicity we give each bin a further parameter which determines its metallicity. A further complexity to the parametrization to deal with post-merger galaxies which contain gas which has followed dramatically different enrichment processes would be to have several populations with the same age but independent metallicities. Although this case is not covered in this thesis it will be the focus of future work.

In order to probe high redshift star formation we have split the final bin into two at a redshift of $z \sim 2$. This may introduce a degeneracy between the two split bins, but we are confident (see later trials) that over a sufficient number of galaxies there is no bias between the bins.

As explained before, the dust content of a galaxy can be modelled by a screen with a set absorption curve. By fixing the curve, a natural way to parameterize the amount of dust in the galaxy is to modify E(B-V). We therefore choose E(B-V) as our dust parameter.

Although it was shown in Reichardt et al. (2001) that the choice of fiducial model does not influence the recovery of the correct SFH, the fiducial spectrum used in this work has a SFH which follows the traditional assumption for an elliptical galaxy, with SF decreasing exponentially with time. The metallicity increases from solar to twice solar from the youngest bin to the eldest, and the noise is assumed to have uniform amplitude and be uncorrelated [1]. This choice of fiducial model does not bias the results and a poor choice simply results in worse than optimal error bars. In particular, it is not crucial to get the noise model correct. Note that the parameter recovery does not depend on the amplitude of the white noise.

[1] Noise with these properties is known as *white noise*

2.2 Interactive Data Language MOPED Implementation

I have shown that the MOPED algorithm has great potential to tackle problems that were previously intractable. In order to be useful the algorithm must be embedded into a program, and the implementation of the MOPED algorithm used for this thesis has been coded in the Interactive Data Language (IDL) programming environment. IDL is a proprietary high level interpreted language, distributed by Research Systems International. Although languages such as C++ and Fortran are perhaps quicker for raw number crunching, IDL was chosen for its ease of use and heavily vectorized and matrix-optimized routines.

MOPED was initially coded for analysis of individual galaxy spectra on a one-by-one basis. It was adapted into a very simple batch form in order to process the galaxies in the Kennicutt Sample (Kennicutt (1992b); Reichardt et al. (2001)), but significant changes, outlined below, were necessary for the analysis of the SDSS.

2.2.1 MOPED Core

The programs which make up the core of MOPED can be broken down into two stages: *pre-data*, where the environment is configured and the model dependent but data independent b vectors are calculated, and *post-data*, where analysis is carried out on data to estimate the star formation fractions and metallicities of the various stellar populations in the galaxy.

Pre-data

The pre-data phase configures the MOPED environment to the size of the spectra that will be used, and builds the b-vectors from stellar models and the dust extinction curve. Most of the configuration settings are contained in the program `Change_Setup.pro`. This sets such parameters as the number of bins, the location of external files, which models to use, where they are located etc. It also describes the wavelength range of the data. This data is assumed to be at the restframe of the galaxy in question. For most applications, the boundaries of the bins are scaled to be equally spaced in the log. `logbins.pro` calculates the boundaries of the bins required to spread them equally in log time between the big bang and the present day.

Once the environment has been configured, the b-vectors must be calculated. The b-vectors require spectra at the resolution of the input data for each age bin for which star formation is to be calculated. `prep_spectra.pro` takes the spectra from the modelling files, where they are stored with irregular wavelength spacing, and rebins them to 20 Å resolution over the wavelength range required. It also removes regions which may suffer contamination from emission lines and normalizes the models such that the flux at 5500 Å is 1.0. In order to increase the processing speed later on, spectra for a grid of metallicity and dust values are precalculated by `comp_yave.pro`. When the models have been read in for each grid point, `prep_compress.pro` precomputes the partial derivatives of the various models which will be later used to calculate the b-vectors. An essential component of the galaxy model is correction for dust extinction. The program `dust_corr.pro` is used to apply a dust extinction curve to the model spectra. The

model is contained in a data file chosen in the configuration stage. Finally, `comp_evector.pro` Calculates the MOPED b-vectors from the fiducial model, set to a typical evolved galaxy.

Post-data

Once the environment has been configured for the particular wavelength range of the spectrum under study, individual spectra can be imported and the relative strengths of the stellar components calculated. The program `input_dataNM.pro` takes the spectrum from whatever format it has been stored in (in the case of SDSS, the Flexible Image Transfer System (FITS) format) and converts it to the IDL internal format. In order to be processed the spectrum must be shifted to the rest frame of the galaxy. To determine the redshift, the program accesses the header of the FITS file and reads in the SDSS calculated redshift. After shifting, the spectrum is rebinned to 20 Å resolution. The program extracts the regions taken out by `prep_spectra.pro` and normalizes the spectrum to the same 5500 Å flux value. This allows MOPED to determine the relative fractions of the spectrum. The normalization value is stored, as it will later be used to calculate the stellar mass of the galaxy. The spectrum is now in a condition suitable for the MOPED algorithm to be applied.

Given 50 initial random start points, `runminspectraNM.pro` uses the conjugate-gradient method to search the likelihood hypersurface for local maxima. The search uses the `SpectraMinCJG.pro` program assesses the likelihood of a given set of parameters using the MOPED compressed dataset as outlined above. This is the step where MOPED out performs any other form of lossless compression - instead of estimating the likelihood from a full χ^2 analysis with $t \propto N_{pixel}$, it is able to calculate likelihoods $t \propto N_{parameters}$. Although the solutions derived by the conjugate gradient search may give good solutions in low dimensional problems and with few parameters, for a full multidimensional search a Markov chain technique is required. The chain is managed by `mcmc.pro` which finds the global likelihood peak given the best guess of the conjugate gradient as a starting point. More details of this program and the Markov Chain Monte Carlo (MCMC) technique are contained later in chapter. By this stage the parameters for each galaxy have been determined. If a visual check of the spectra is required then `make_specNM.pro` can be used to produce the best synthetic spectral fit to the spectrum, reconstructed from the parameters extracted by MOPED. The programs' interactions are summarized in figure 2.3.

2.2.2 Markov Chain Hypersurface exploration

MOPED allows rapid calculation of the likelihood of a given set of parameters. Given that the number of parameters will be usually large (typically more than 10 and in the case of MOPED 20-25), computation of a grid to explore the likelihood surface is impractical – simply using 10 grid points per dimension would require 10^{25} evaluations. If the likelihood surface is in any way complex or experiences sharp changes then it is possible that a the best solution may be missed.

MOPED originally used a Conjugate Gradient (CG) method to locate the best solution. The CG method starts at a guess value and then moves in a direction determined by the local gradient to fall into the best local solution. Using several start points distributed over the parameter space allows multiple solutions to be explored and is likely to be successful for a reasonably simple surface. If the shape of the likelihood surface local to the solution can be approximated to a multivariate Gaussian, the Fisher matrix can be used to compute errors. It is not certain that this

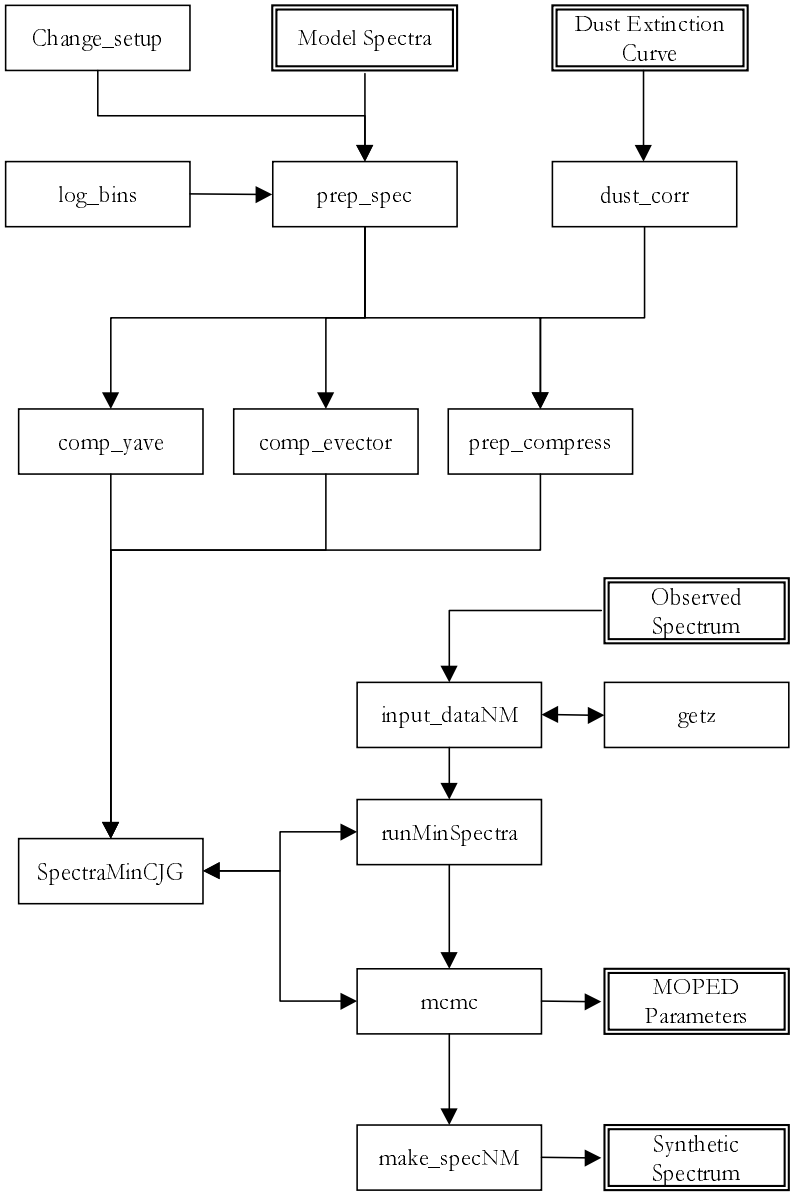


Figure 2.3: Simplified flowchart of interaction of the programs which make up the MOPED system. Single box indicates program, double external inputs/outputs.

is a valid assumption for increased dimension problems, when the surfaces may become both degenerate and complex with small scale structure. A method which can quickly, automatically and accurately explore multiple solutions of a hypersurface is required.

In this general case, an efficient method to sample the likelihood surface is through the Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953; Hastings, 1970)[2]. In essence the Markov chain algorithm is very simple: a chain of likelihood values in the parameter space is created in the following way. At each point, a random step is made in parameter space, and a random number between 0 and 1 is drawn. This number is essentially compared with the ratio of the likelihood values between the current step and the previous one, although there are some slight modifications, especially near the boundaries of the parameter space. If the value of the likelihood ratio is bigger than 1 or the random number, then the current step is accepted and added to the chain. If, however, it is smaller than the random number, then the point is rejected and not added to the chain: instead the last point is repeated. Asymptotically, the distribution of points in the chain samples the likelihood surface in an unbiased way Gilks et al. (1996).

For MOPED we used a uniform prior within certain bounds to produce the step. If knowledge about the shape of the surface is known a priori, then the random stepping can be done more efficiently by using this a priori information. The Fisher matrix can sometimes be useful for this, but if the topology of the likelihood hyper-surface is not known, it may be inaccurate. Marginal errors are then trivial to compute by looking at the distribution of all points for a single parameter. The method is very fast and efficient but the challenge is to adjust the step size of the jump so the likelihood surface around the maximum is explored with the minimum number of steps.

There are some rules to decide the size of the jump a priori (see Gilks et al., 1996), but we find that the most efficient time step can be found by exploration of a few thousand chains for a few galaxy spectra.

As an example of Markov chain convergence is given in figure 2.4. The top-left panel shows a typical spectrum of a galaxy with an old stellar population. The top-right panel displays the recovered star formation history with errors computed using the MCMC. The middle-left panel shows the distribution of points in the plane of parameter values recovered from the MCMC for only 30,000 points in the chain, showing only those within reduced $\Delta\chi^2 = 1$ of the minimum χ^2 point. The middle-right panel shows the same but this time for a chain with 300,000 points. The chain with a small number of points shows a rugged pattern, with a few unexplored areas, but the chain with 300,000 points has covered the region of the likelihood space that is most favoured quite well. More specifically, it is clear that the chain with a small number of points has not come back to the starting point of the chain a few times. This feature is required to establish convergence, and the left hand chain is said to be not well mixed, although in this case it provides a good estimate of the errors. On the contrary, the right hand chain with a small step has oscillated a few times around the starting point.

In the above chains the time step was chosen by trial and error. In principle, the step size can be chosen optimally a priori. For one parameter, for example, the step size should be such that the rejection rate of points is about 60%, leading to a non-negligible chance of the chain exploring regions in the likelihood surface that are more than 3σ away from the best solution. Obviously, for more parameters the rejection rate will increase significantly, since there are many more ways the jump can explore an unlikely region of the parameter space. Conversely, high acceptance rates are indicative of too small a jump step (see below).

[2] An excellent account of Markov chains techniques can be found in Gilks et al. (1996)

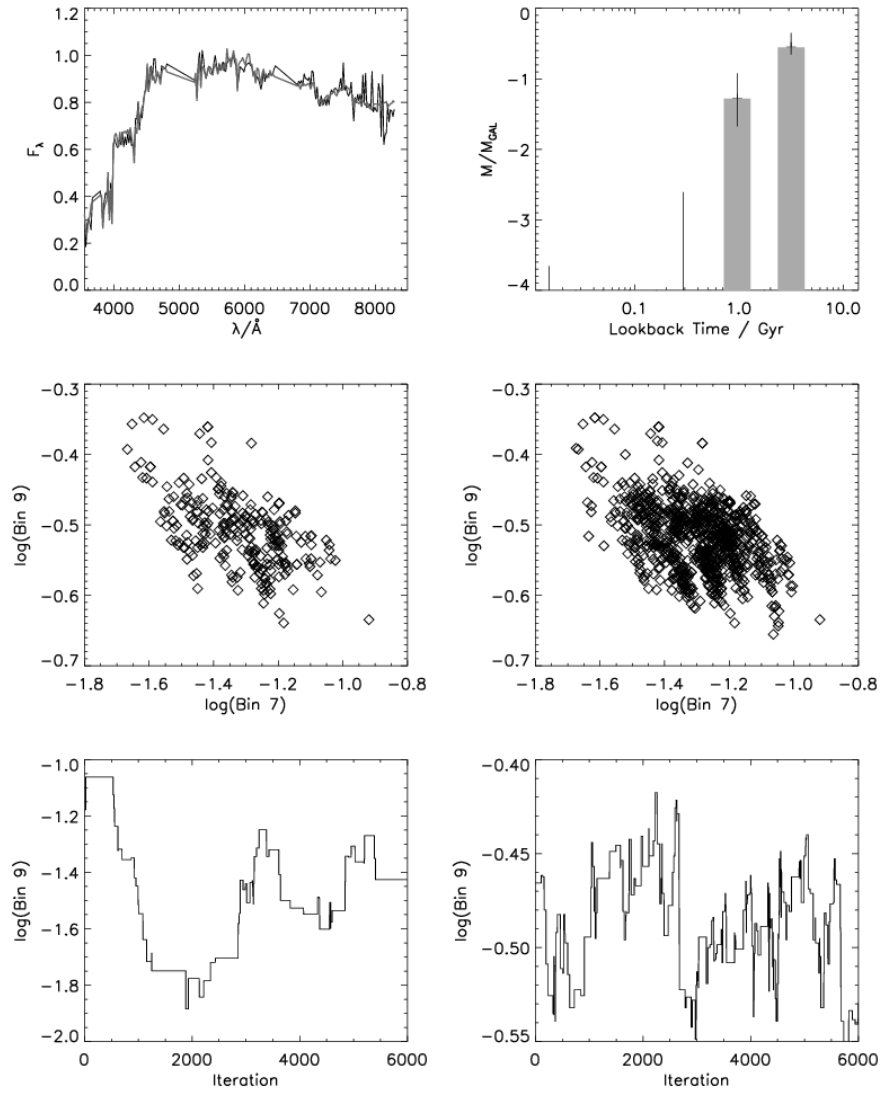


Figure 2.4: Some details of the application of MOPED to an individual SDSS galaxy. The top left panel shows the galaxy spectrum (black) and the best fit Jimenez model (grey). The corresponding star formation fractions are shown in the top right panel. The middle panels show the MCMC points with highest likelihood values, in projection onto parameter plane 7 and 9, which are the bins showing significant star formation, for chains of 30,000 (left) and 300,000 (right). The lower panels show the effects of different step sizes on the estimation of parameter 9.

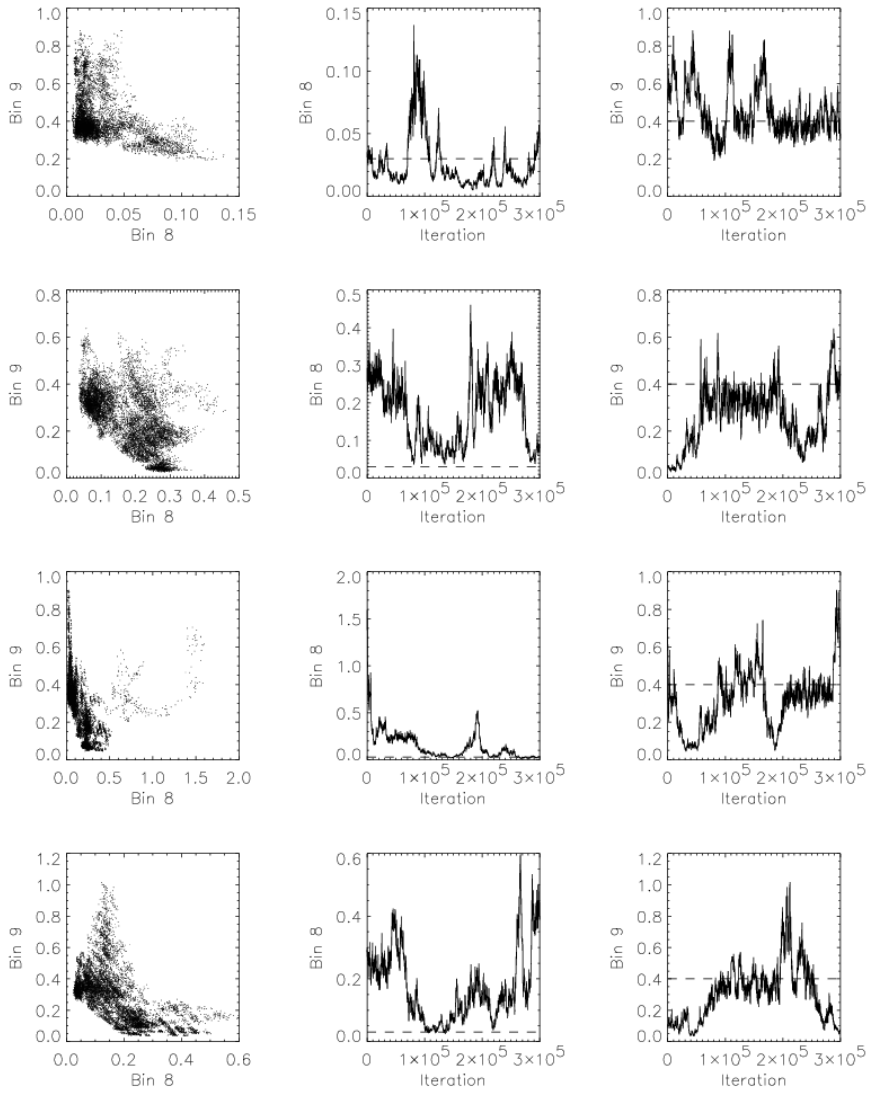


Figure 2.5: Convergence test for MCMC. Four chains of length 300,000 steps have been started at different points in parameter space for the galaxy shown in Fig. 2.4. The far left panel shows the values of the star formation in bins 8 and 9 for each chain. The right-hand panels show the values of the star formation parameters explored by the chains. Note that the excursion to the right in the third row is in fact the starting point of the chain. The long-dashed horizontal lines show the best-fitting solution.

The two bottom panels of figure 2.4 show the values of one parameter (star formation for the bin with most star formation) for part of two chains. The left-bottom panel shows a chain with a step that is too big. Note how the chain remains at same value of the parameter for many steps. The path of the parameter value for the change shows a clear "staircase" pattern, and the chain is inefficient. On the other hand, the right-bottom panel shows a chain with a much better step size. In this case the chain does not dwell for long on a single value of the parameter.

Our experiments show that significant improvements in chain convergence can be obtained by using a nonuniform jump size. The star formation history is divided into two sections, the first covering mass fractions from 10^{-7} (essentially zero) to 10^{-4}, the second from 10^{-4} to 10. In the first region we have 100 logarithmically-spaced points, in the second a thousand. The maximum jump we allow is 20 steps. Similarly, there are 64 values of metallicity in a grid, equally spaced in the log between $0.01 < Z/Z_{\odot} < 5$ and the optimal jump size is 5 grid points. Dust is computed on a linear grid, with 64 elements, between 0.02 and 1.28 and optimal jump size of 3 grid elements. The chain is well mixed for 300,000 steps and the acceptance rate is typically about 2% in 25 dimensions.

A more robust way to estimate the convergence of the chain is the following: start 4 or more chains from widely-separated points in the parameter space and check when the variance for all parameters within the chain and between chains are indistinguishable; at this point the chains have converged. The point is well illustrated in Fig. 2.5. The left four panels show the distribution of points in a projection of the likelihood hyper-surface for two adjacent bins for a random galaxy with only significant star formation in bins 8 and 9. The second and third column of panels show the values of bins 8 and 9 for the four chains and 300000 jumps. The dashed line is the best-fitting value for the parameter. A few features become apparent by visual inspection. When the chains start from values different from the best fit, it takes 50-100,000 steps for them to converge. After this, the chain remains on the good valley solution for some time, before undergoing a random excursion away from the best solution, before returning at some point. This returning behaviour is required for convergence.

If we use the above convergence criterion, we see that chains converge at different points in path. For example, the chain on the top panel for bin 9 only converges after 200,000 points. While the second chain from the top, does so after only 50,000 steps for bin 9. This illustrates how important it is to run chains for a long enough time and estimate convergence.

An alternative approach is to run only one chain from one point in the parameter space[3] and let it run for long enough as to explore most of the likelihood space. A good test to check convergence then is to monitor the likelihood values as a function of the step. The chain should return to close to the maximum likelihood solution. In the present paper we follow this criterion and in all cases we find chains have converged after 300,000 steps, as illustrated in the example of Fig. 2.5.

It is worth emphasizing that the chain needs to be able to explore the likelihood hyper-surface well in the vicinity of the peak in order to be sure errors are derived properly. Also, there is some danger in using local approximations to the likelihood surface (the Fisher matrix) to compute the length of each jump. Imagine a flat valley in the likelihood surface with a narrow region within where the likelihood is better. A first jump may bring you into the valley, then the Fisher matrix will indicate that the hyper-surface is locally very flat which will systematically provide a very large jump and therefore the better value will be systematically missed.

[3]Using the conjugate gradient method it is easy to choose this point as the one closest to the best solution

Covariances between bins

The left panels in Fig. 2.5 illustrate how covariances appear between adjacent bins. The far left panel shows the values of the whole chain, whereas the middle-left panel shows only those points within reduced $\Delta\chi^2 = 1$ of the best-fitting solution. Note how the paths for bin values follow a common pattern: star formation in one bin is traded by star formation on the adjacent bin (with different metallicity) - the sum of bins 8 and 9 is well constrained in this example. We also see from the right-hand panels that the chains spend longer periods close to the best-fitting parameter where the parameter has more star formation (parameter 9) than less (parameter 8).

2.2.3 Emission Line Removal

Emission lines in galaxy spectra have two main sources - hot, diffuse gas around stars and AGN. Since there are not yet detailed models for these components of the spectrum, and they play no part in synthetic stellar models, it is essential to remove them from the data. To perform this removal we simply mask out the emission line contaminated regions from the data and the synthetic stellar spectra which are used to calculate the b-vectors (It is necessary to perform this operation before calculating to maintain orthogonality in the vectors). The three main candidates for emission line contamination are the H-α, H-β and [OII] doublet lines. These have central wavelengths of 6562.8 Å 4861.342 Å and 3726.0, 3728.8 Å respectively. The exact regions to be removed were determined by visual examination of around a thousand Sloan EDR galaxies as 3700 - 3760, 4840 - 5200, and 6500 - 6800 Å.

Although these regions exclude emission lines, they also exclude some signal which could be used for MOPED analysis. In order to check that the removed signal does not interfere with the spectral fitting, we have compared the b-vectors (see Reichardt et al. (2001)) generated with emission line data to those generated with that section removed (Figures 2.6 and 2.7). Testing with noisy synthetic galaxy spectra created with the Jimenez et al. (2004a) models has shown that removing the emission line regions does not significantly effect MOPED's ability to recover star formation history.

2.2.4 Accounting for Dust

The initial version of MOPED, and that used to analyze the SDSS EDR, used the Calzetti (1997) estimate for the attenuation of light through interstellar dust grains. The Calzetti law was calculated for starburst galaxies, and is not considered to be appropriate in the general case. For the DR1 analysis we chose instead to use the Large and Small Magellenic Cloud dust attenuation curves of Gordon et al. (2003). To allow model choice it was necessary to modify `dustcorr.pro`, the program which calculates the dust correction to be applied to the synthetic spectrum. The simple, one parameter dust screen used by MOPED is not the most sophisticated model available, but it has the advantage that it only occupies one parameter in the hypersurface that must be explored. Changing the dust model between the starburst shape suggested by Calzetti (1997) and the LMC and SMC curves of Gordon et al. (2003) resulted in very little difference in recovered parameters, although the average of the wavelength independent part of the dust correction (the free parameter in the model) shifted. This was later found to be due to a normalization error. Changing the dust screen model used does not appear to change the determination of the SFH of

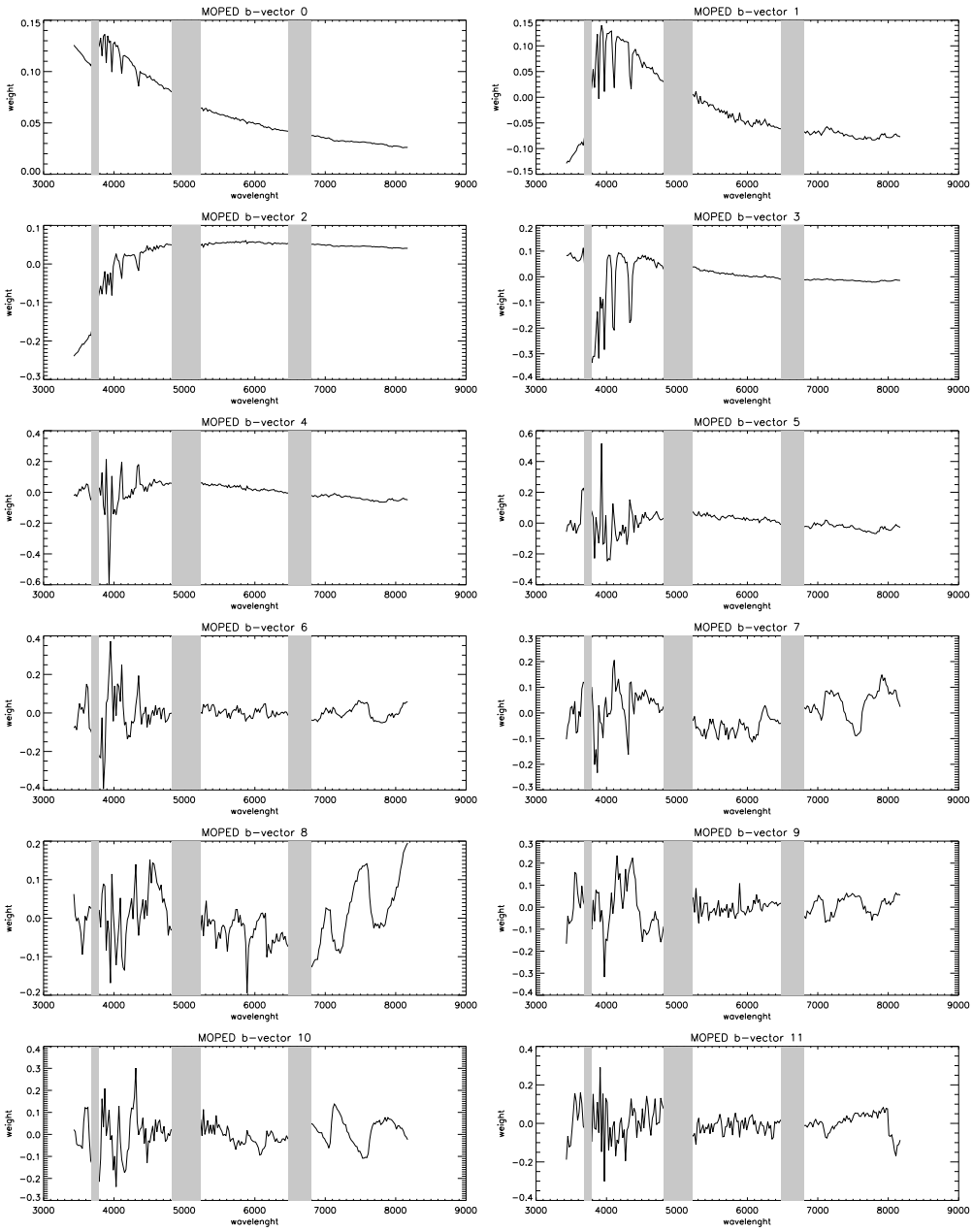


Figure 2.6: b-vectors 0 – 11. The b-vectors show the weighting of pixels in the data vector for each parameter (in this case b-vectors 0 to 10 are the SF Bins from youngest to oldest and b-vector 11 is the youngest metallicity bin). The grey areas are those removed from the spectra.)

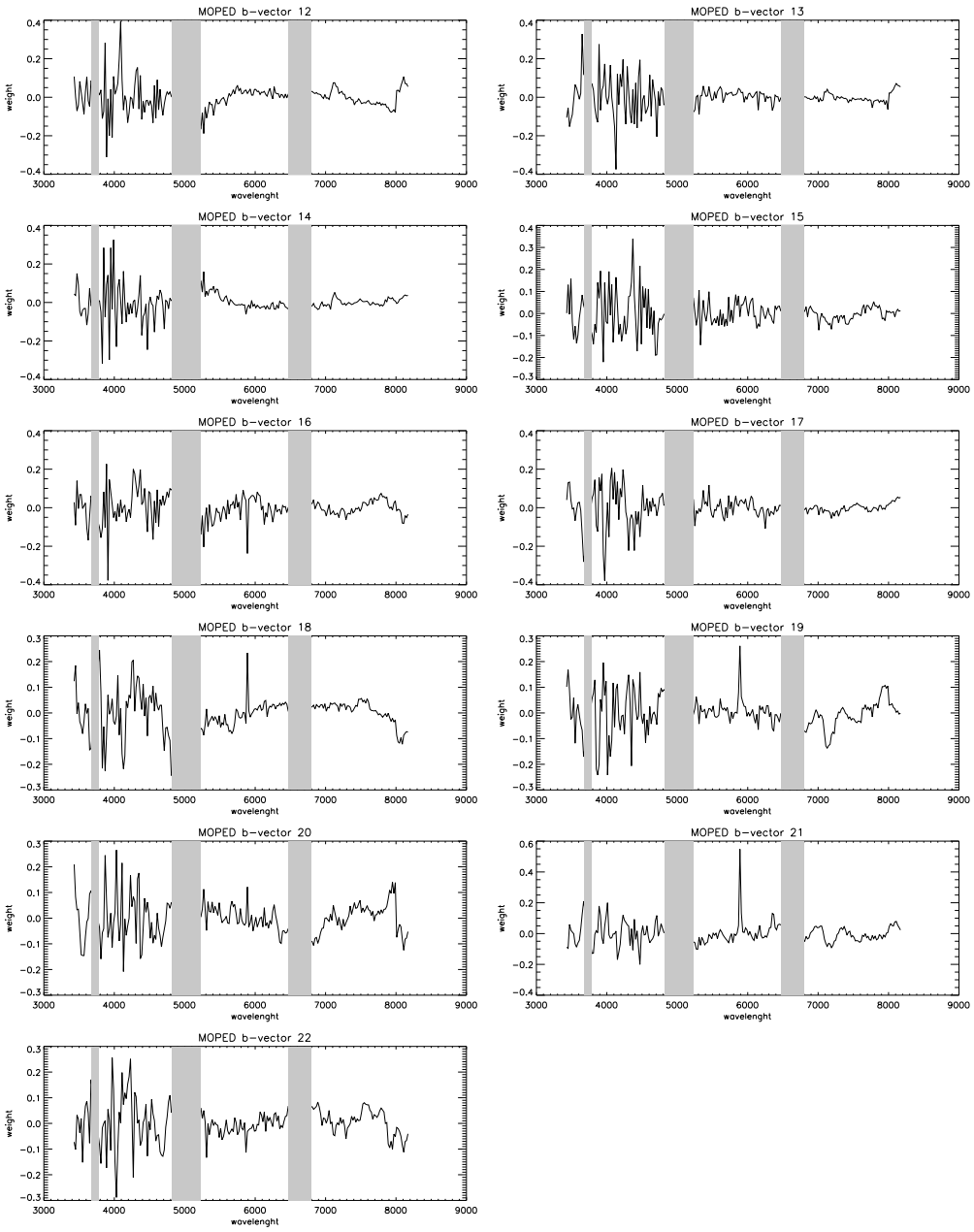


Figure 2.7: b-vectors 12 – 22. The b-vectors show the weighting of pixels in the data vector for each parameter (In this case b-vectors 12 to 21 are the ten oldest metallicity bins and 22 is the dust parameter). The grey areas are those removed from the spectra.

a galaxy. We chose the curve of the LMC as it is the one thought to be most universally applicable in galaxies which are not undergoing a burst of star formation.

2.2.5 Determining the Stellar Mass from MOPED results

For a given galaxy spectrum and set of models, MOPED will return the fractions of each model represented in the spectrum, and hence the mass *fractions*. Although this information is useful, it is considerably more useful when converted into an absolute mass - allowing direct measurement of star formation rates for example.

The extra information required over the fractions supplied by MOPED are the redshift of the galaxy and its flux. From this the luminosity of the source can be calculated and from that the relative absolute masses of the different aged stellar populations extracted. Adding these fractions together and allowing for losses due to supernovae and winds gives the total stellar mass of each galaxy at the time of observation. These masses are subject to an aperture correction, which can be calculated from the fiber and petrosian magnitudes recorded in the Sloan database.

Converting fractions to absolute masses

The synthetic stellar models give their output in $I_\lambda/M_\odot/L_\odot/$Å, and the data from the Sloan spectra is given in units of 10^{-17} erg/cm^2/s/Å. Since we wish to know the stellar mass equivalence of each of the MOPED parameters, it is required to convert the units of the spectrum to those used by the spectral models. In this way the mass fractions given by MOPED can give absolute masses in terms of the masses of the synthetic models. In the units of the models, the Sloan fluxes are measured in $10^{-20}\mathrm{W\ m^{-2}\ \text{Å}^{-1}}$

If mass in bin i = m_i then power output per Å is

$$P_\lambda = \sum I_{\lambda,i} m_i L_\odot, \tag{2.60}$$

so Flux is

$$f_\lambda = \frac{\sum I_{\lambda,i} m_i L_\odot}{4\pi D_L^2} \tag{2.61}$$

and normalised flux:

$$G_\lambda = \frac{F_\lambda}{F_{5500}} = \frac{10^{22} L_\odot}{4\pi D_L^2 F_{5500}} \sum I_{\lambda,i} m_i \tag{2.62}$$

$$= \frac{10^{22} L_\odot}{4\pi D_L^2 F_{5500}} \sum \frac{I_{\lambda,i}}{I_{5500,i}} m_i. \tag{2.63}$$

$$G_\lambda \equiv \sum \frac{I_{\lambda i}}{I_{5500i}} f_i, \tag{2.64}$$

where f_i are the moped bin "fractions":

$$\frac{10^{22} L_\odot}{4\pi D_L^2 F_{5500}} m_i I_{5500,i} = f_i. \tag{2.65}$$

So to get the Galaxy masses from the relative fractions we have

$$m_i = \frac{4\pi D_L^2 F_{5500} f_i}{10^{22} L_\odot I_{5500,i}}, \tag{2.66}$$

where F_{5500} is the normalization factor "norm" from the `input_dataNM.pro` program (stored for each galaxy) and values of I_{5500} from the "tmp" value in the `prep_spectra.pro` program (which is constant for all ages and metallicities to maintain normalization to solar spectrum).

This is not the whole story however - there are two places where factors of (1+z) come into our equation: for spectral bandwidth and luminosity distance. The spectral bandwidth:

$$f_{\nu_{obs}} = \frac{L_{\nu(1+z)}}{4\pi(R_o S_k)^2(1+z)} \tag{2.67}$$

is needed in terms of wavelength, not frequency:

$$f_\lambda = f_\nu \left|\frac{d\nu}{d\lambda}\right| = f_\nu \frac{c}{\lambda^2} \tag{2.68}$$

$$f_\lambda = \frac{L_{\nu(1+z)}}{4\pi(R_o^2 S_k^2)(1+z)} \frac{c}{\lambda^2} \tag{2.69}$$

$$L_{\nu_e} d\nu_e = L_{\lambda_e} d\lambda_e \tag{2.70}$$

$$\frac{c}{\lambda_e^2} L_{\nu_e} = L_{\lambda_e} \tag{2.71}$$

$$f_\lambda = \frac{L_{\lambda/(1+z)}}{4\pi(R_o^2 S_k^2)(1+z)} \frac{{\lambda_e}^2}{\lambda} \tag{2.72}$$

$$\lambda_e = \frac{\lambda}{1+z} \tag{2.73}$$

$$f_\lambda = \frac{L_{\nu(1+z)}}{4\pi(R_o^2 S_k^2)(1+z)^3}. \tag{2.74}$$

The calculation of Luminosity distance also contributes:

$$D_L = R_o S_k (1+z) \tag{2.75}$$

$$f_\lambda = \frac{L_{\lambda/(1+z)}}{4\pi D_L^2 (1+z)} \tag{2.76}$$

$$L_{\lambda/(1+z)} = (1+z) f_\lambda (1+z). \tag{2.77}$$

$$\tag{2.78}$$

So overall an additional factor of (1+z) is required when considering the cosmological corrections for our masses.

Correcting for supernovae / stellar wind losses

MOPED considers the fractions of the total mass ever formed in a galaxy, and the output takes no account of mass that has since disappeared due to stellar winds and supernovae. This recycling fraction, R, can be calculated by analysis of the IMF and stellar models used. We have followed the method of Cole et al. (2001) and used a recycling fraction of 0.28. The mass of the galaxy is therefore

$$M_{\text{present day}} = M_{\text{ever created}} \times (1 - R). \tag{2.79}$$

Aperture Correction

The Sloan spectrograph has a fibre width of 0.2 mm, which corresponds to an angle of 3 arc seconds on the sky York et al. (2000). The flux received by the spectrograph is not necessary all the flux from the source. To compensate for this we scale the masses to the petrosian magnitude, as measured by the photometric survey. This causes a potential problem, in that for large galaxies we may sample an area whose spectrum is not representative of the whole galaxy. Glazebrook et al. (2003) shows that the average colours in the fibres are the same as the average colours from the petrosian magnitudes (except for large galaxies at very small z, of which there are few). This suggests that although scaling the light from the fibre to the photometrically modelled magnitude is liable to fail for individual galaxies, overall it is successful. The correction applied is as follows:

$$M_{\text{aperture corrected}} = M_{\text{uncorrected}} \times 10^{0.4\times(M_p - M_f)}. \tag{2.80}$$

Overall Algorithm

In summary, to go from the fractions returned from MOPED to the stellar masses M in terms of $M_\odot$, the following algorithm is used:

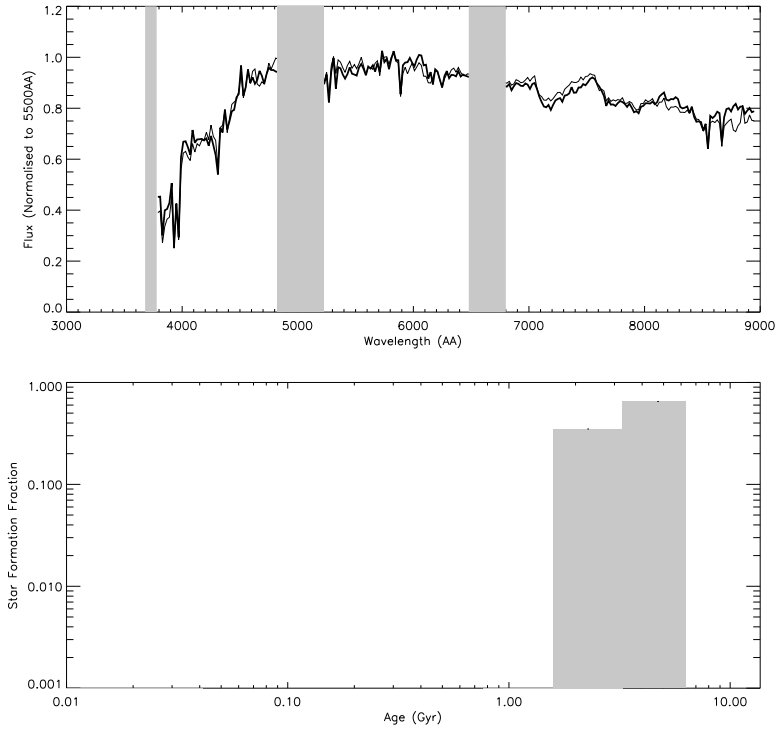
$$M = \sum \frac{4\pi D_L^2 F_{5500} f_i}{10^{22} L_\odot I_{5500,i}} \times (1 + z) \times (1 - R) \times 10^{0.4\times(M_p - M_f)}. \tag{2.81}$$

2.2.6 Choice of Initial Mass Function

The masses recovered by this method are dependent on choice of IMF. We have chosen a Salpeter IMF to allow direct comparison with theory and other predictions. Since we are using a Salpeter IMF with a low mass cut off at 0.1 $M_\odot$ we are insensitive to any mass included in brown dwarfs.

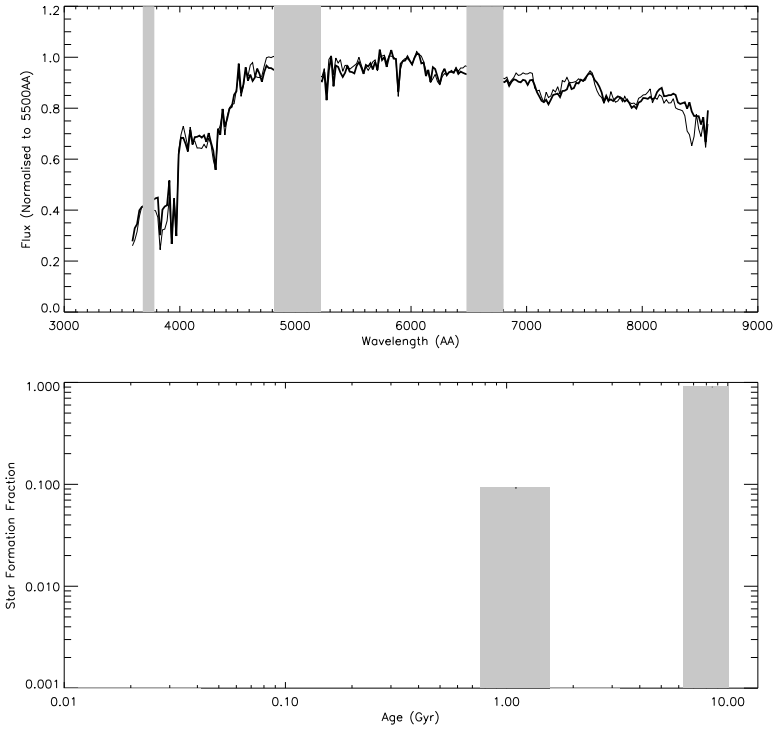
2.3 Examples of Recovered Spectra

MOPED is able to run, and extract spectral parameters for, almost all of the galaxies in the SDSS. There follows a series of figures with information about each included in the figure captions. The upper part of each plot shows the reconstructed spectrum – the thin line is the galaxy spectrum, the thick line the synthetic fit curve generated by MOPED. In each case, the grey areas correspond to regions omitted from the study due to emission line contamination. The lower part of each plot gives a barchart representing the distribution of SF among different aged bins – the SFH of the spectrum. The ages quoted are the lookback times from the present day. The vertical lines on the bars (not always present) reflect the formal errors on the solution. For a non degenerate hypersurface these are very small, but when degeneracies exist between solutions they reflect these degeneracies (often only one or two further solutions).



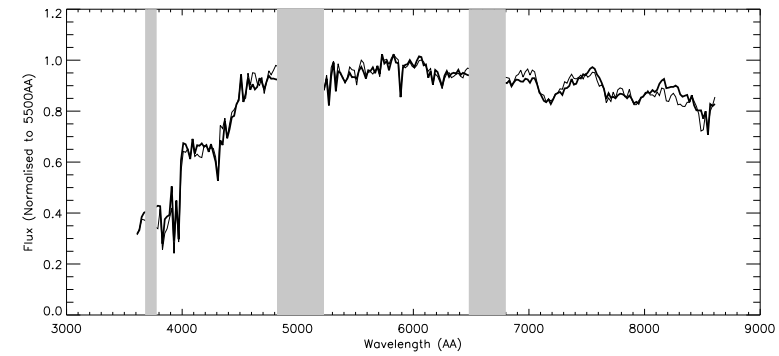
Age:	SFF:	Met:	Dust:
0.0138985	3.0764553e−07	0.10670359	0.020000000
0.0287800	5.2245714e−06	4.5303250	
0.0595954	2.7418926e−05	0.46858786	MJD:
0.123406	2.7418926e−05	0.46858786	51877
0.255539	7.3798963e−06	0.42457087	
0.529150	9.5071512e−05	0.12997527	Plate:
1.09572	8.4732629e−06	0.053492579	0385
2.26894	0.34918089	1.6893571	
4.69835	0.65020523	0.76735264	Fibre:
8.50000	0.00016521546	0.039789733	595
12.0000	0.00027736430	0.15832233	

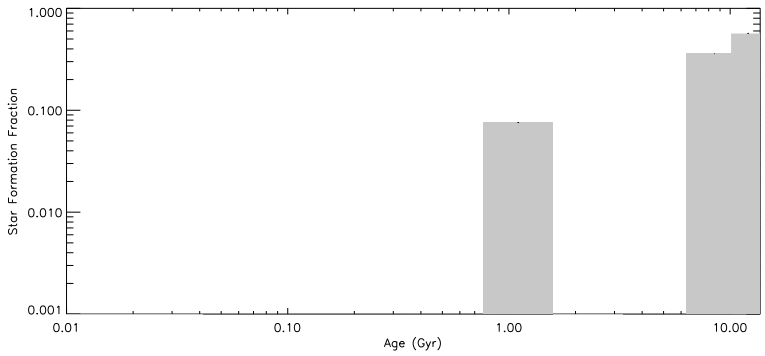
Figure 2.8: This spectrum, of an aged galaxy which has not had a recent star burst, illustrates just how good MOPED can be at recovering the SFH of a galaxy. The reconstructed spectrum fits the majority of the line strengths in the observed spectrum, and there are no degenerate solutions in the hypersurface. The high metallicity of the 0.028 Gyr bin, out of the trusted region of modelling, is explained by the absence of SF in that bin - it is essentially unconstrained.



Age:	SFF:	Met:	Dust:
0.0138985	1.4345623e−05	0.087598666	0.039999999
0.0287800	1.0155922e−05	0.043914908	
0.0595954	0.00024083485	0.34855285	MJD:
0.123406	1.5729662e−06	3.3698202	51818
0.255539	0.00054540529	0.011036741	
0.529150	0.00045890158	0.19285186	Plate:
1.09572	0.092623109	1.6893571	0383
2.26894	0.00051489599	0.25926640	
4.69835	9.4780631e−06	0.032665521	Fibre:
8.50000	0.90514807	1.5306662	557
12.0000	0.00043323091	0.016375854	

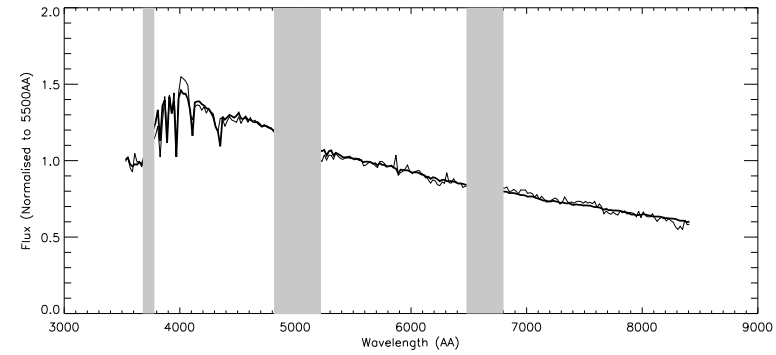
Figure 2.9: This spectrum, again well fitted by the MOPED reconstructed spectrum, shows that roughly 90% of the population of the galaxy was formed around 8.5 Gyr ago, and that 10% was formed 1 Gyr ago. This could reflect either a burst of star formation or a merger between two galaxies which formed many Gyr apart.

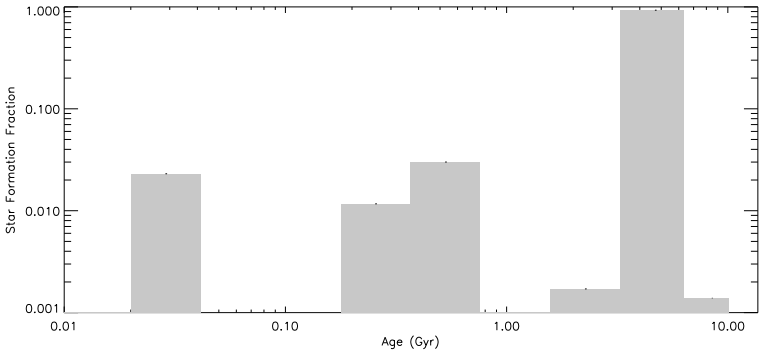




Age:	SFF:	Met:	Dust:
0.0138985	1.5414356e−06	1.5306662	0.059999999
0.0287800	4.1488305e−07	0.12997527	
0.0595954	6.4257784e−05	0.34855285	MJD:
0.123406	0.00017295199	0.087598666	51999
0.255539	0.00017097219	0.15832233	
0.529150	0.00050457472	1.1385657	Plate:
1.09572	0.075496256	1.3868829	0503
2.26894	0.00023330575	1.3868829	
4.69835	2.1773360e−06	0.31581139	Fibre:
8.50000	0.35721147	0.84690696	613
12.0000	0.56614207	1.3868829	

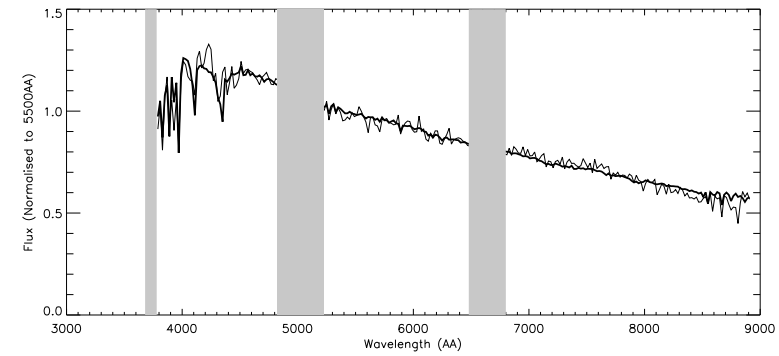
Figure 2.10: A further example of a 1 Gyr burst of star formation, although with much smaller formal errors on the parameters than in the previous spectrum. This reflects the fact that the peak in the hypersurface was strong enough to break the degeneracy between the final two bins.





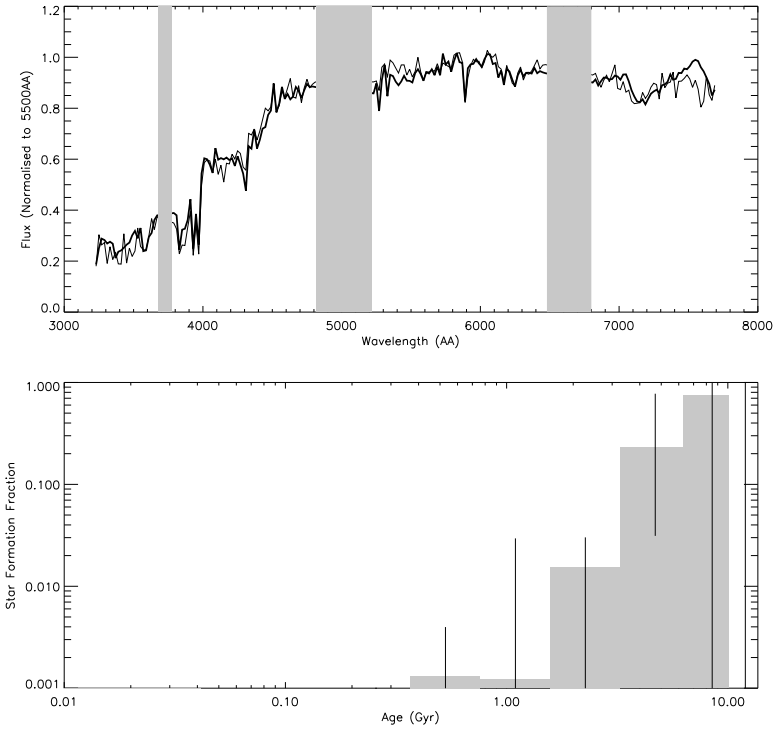
Age:	SFF:	Met:	Dust:
0.0138985	0.00057394532	0.10670359	0.099999994
0.0287800	0.023113757	0.022015388	
0.0595954	1.9447801e−06	1.0316143	MJD:
0.123406	0.00044042354	0.38468863	51788
0.255539	0.011718458	0.11776599	
0.529150	0.030121081	0.071914392	Plate:
1.09572	6.9003433e−07	0.018073605	0386
2.26894	0.0017134471	1.0316143	
4.69835	0.93083043	0.62996066	Fibre:
8.50000	0.0013927416	0.065159103	474
12.0000	9.3082993e−05	0.62996066	

Figure 2.11: This spectrum shows a recent burst of star formation, and the MOPED recovered parameters clearly point to three separate episodes during the SFH where significant star formation occurred.



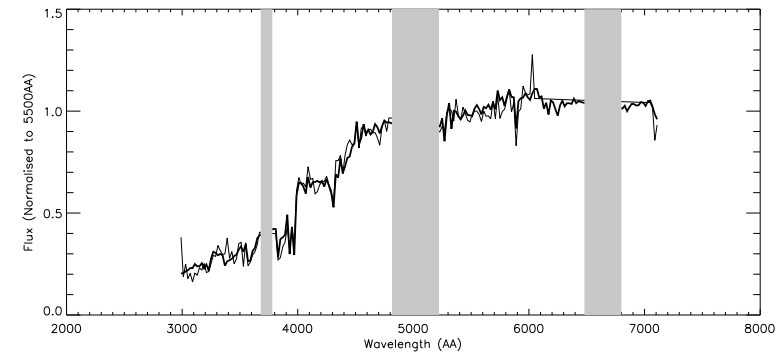
Age:	SFF:	Met:	Dust:
0.0138985	0.0047170976	0.087598666	0.079999998
0.0287800	0.012843253	0.022015388	
0.0595954	0.0018996556	0.036052078	MJD:
0.123406	1.5983609e−05	0.12997527	51788
0.255539	7.4760969e−06	0.57078505	
0.529150	0.0014245408	0.026816859	Plate:
1.09572	0.066630755	0.053492579	0386
2.26894	0.80107813	0.57078505	
4.69835	0.094118466	0.15832233	Fibre:
8.50000	0.00033394507	1.1385657	520
12.0000	0.016930701	0.048467732	

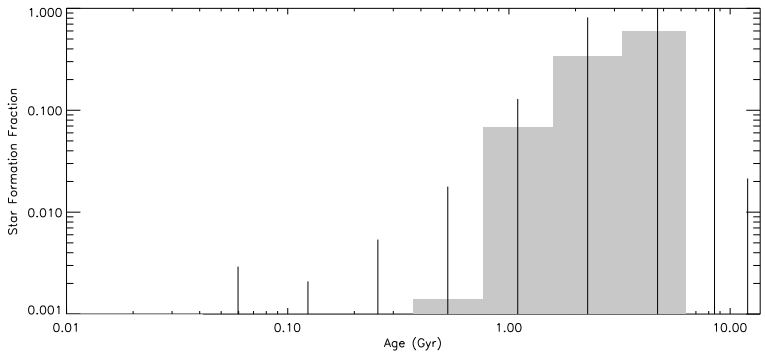
Figure 2.12: This spectrum shows a recent star burst, as does the recovered SFH, but the MCMC errors betray that there are several degenerate solutions which will match the spectrum to a reasonable degree.



Age:	SFF:	Met:	Dust:
0.0138985	4.1660695e−06	0.010000000	0.020000000
0.0287800	6.7565739e−06	0.84690696	
0.0595954	0.00014279947	0.17473629	MJD:
0.123406	5.1253868e−06	0.28614560	51788
0.255539	0.00021863844	4.5303250	
0.529150	0.0013174268	0.018073605	Plate:
1.09572	0.0012154208	0.043914908	0386
2.26894	0.015478421	0.31581139	
4.69835	0.23159363	1.2566051	Fibre:
8.50000	0.74942233	1.5306662	512
12.0000	0.00059528744	1.6893571	

Figure 2.13: A MOPED solution showing the traditional assumption of an exponentially decaying SFH is shown here, with corresponding large errors. The errors on the bar chart indicate that the spectrum is not well constrained, as is apparent from the reconstructed spectrum. There is also a clear degeneracy between the final two bins. This is, however, the best fit spectrum for this particular galaxy.





Age:	SFF:	Met:	Dust:
0.0138985	7.2338288e−07	0.019947364	0.059999999
0.0287800	0.00061569912	4.1047660	
0.0595954	0.00036255031	3.0532773	MJD:
0.123406	0.00020154332	0.19285186	51877
0.255539	0.00018809118	0.024297813	
0.529150	0.0013783784	0.026816859	Plate:
1.09572	0.067510062	4.1047660	0385
2.26894	0.33447867	4.1047660	
4.69835	0.59479621	1.6893571	Fibre:
8.50000	1.0218057e−06	0.053492579	598
12.0000	0.00046705540	0.46858786	

Figure 2.14: There are a small number of galaxies in the SDSS catalogue which, for instrumental reasons, are incomplete. Even with such problems (see the range $6000 - 7000$Å) MOPED is able to determine the SFH and supply a reconstructed spectrum.

CHAPTER 3

MOPED for surveys

Until this point consideration of MOPED has been limited to analysis of individual galaxy spectra to recover the SFH. Although this is interesting for the study of particular galaxies, even more desirable is a robust method for estimating the SFH of *thousands* of galaxies in a survey such as the SDSS. To accomplish this task the I have written a parallel processing MOPED implementation, Royal Observatory Auto-MOPED (ROAM), which allows use of many (non dedicated) computers simultaneously. ROAM relies on the batch processing allowed by the new MOPED pipeline and the use of precalculated sets of b-vectors.

The results of such a study can be combined to yield general information about the evolution of galaxies. The later sections of this chapter cover the programs required for interpretation and combination of the results in a cosmological context. Combining a hundred thousand galaxies is not a trivial task, and I include some tests which were developed to assess the combination process.

3.1 Parallel Computing

3.1.1 Motivation

While the MOPED algorithm allows very fast computation of the components of a galaxy spectrum, the process is far from instantaneous. The average time taken to analyze a spectrum is 2 minutes: to process the ~ 1 million galaxies anticipated in the final SDSS spectroscopy survey would still take around 4 years on a single CPU. For MOPED, results obtained from one spectrum do not depend on the results for any other spectrum, so any subset of the data can be analyzed independently of the remainder. This sort of problem is described as *trivially parallelizable*, which means that no part of the task depends on other parts to have already been processed. By splitting the analysis over a N CPUs the total time taken for the job can be reduced to $1/N$ of that required for 1 CPU.

The time taken to precalculate the MOPED b-vectors is negligible compared to the time taken to process a large dataset: around 7 hours. A further speed gain can be obtained by calculating these b-vectors on one machine and simply copying them across the network to the others, although since this phase is required for all spectra, it does not fall into the trivially parallel category.

This type of computational problem requires High Throughput Computing (HTC). Rather than requiring the massive interconnected memory and CPU resources which are characterized by High Performance Computing (HPC), HTC can be performed using off the shelf components. Instead of looking at the instantaneous computational speed, of far more interest is the average rate at which jobs can be completed. In this case it becomes less important that computers are dedicated to the project, and be used part time as and when they become availiable. While it would be prohibitivly expensive to purchase multiple CPUs for any one project, there are many CPUs in an academic institution which are not used to their full potential. HTC allows the use the "wasted" CPU time from these computers, while at the same time not impairing those computer's ability to work for their owners.

A number of 3rd party solutions to this problem were investigated, but finally a custom environment was created picking the best components of the other systems.

3.1.2 The Condor Project

The Condor Project is a well established HTC environment created by the Computer Science Department of the University of Wisconsin (Litzkow et al., 1988; Thain et al., 2004). The main aim of the project is to provide easy HTC to users without requiring them to actually know what is happening to their job. For instance, if a researcher wishes to run a long simulation on a certain type of computer, that researcher should just submit the job to condor and then return to find the job done. It does not matter that the job was queued for a period, started on computer X, moved to computer Y when the load from the computer X's owner was high, automatically recovered and restarted on Z when Y failed... just that the job was submitted and is now done.

Behind the scenes Condor involves several compromises to allow this process to work. A user specifies the level to which Condor is allowed to use their CPU, and what should trigger Condor to start a job. These constraints can range from a period of inactivity on the keyboard to set times of day or CPU load. To allow tasks to *migrate* from one computer to another *checkpoints*

are required where the task will be able to recover from if it is forced to stop. The checkpoints and migration are implemented by compiling Condor libraries into the code. To allow commercial (binary)[1] code to use Condor, a *vanilla* mode is available. This version of Condor allows a user to submit jobs which are then run on computers when they become free. If the owner of the computer comes back while the job is running, it is cancelled and resubmitted to another computer. Obviously, for short tasks this will usually be successful, but processes requiring a longer time to run will have a larger chance of failing.

Since MOPED is written in commercial and non-compilable software, the only option open is to use the vanilla universe, negating many of the most useful abilities of Condor. It would be possible to rewrite the MOPED algorithm in a non-commercial language, for example C, and then use Condor, but this would involve a huge amount of work. Instead I chose to write a parallel processing environment customized for MOPED. In order to combine this with the precalculation step outlined below it was necessary to develop a MOPED pipeline.

3.1.3 Precalculated MOPED Sets

A full MOPED run using the original programs (Reichardt et al., 2001) takes around 10 minutes and there are around 50 galaxies in the Kennicutt sample. The resulting run of 500 minutes is possible to complete over night, with no pressing requirement to be optimized further. Calculating a new set of b-vectors for each galaxy allowed variations in redshift, observed redshift range and spectral binning. Later analysis showed that in fact the precalculation stage of the process took around 4 times as long as the actual analysis phase. In this case *precalculation* refers to the tasks which depend on the range of the data being supplied, but not the actual data itself (ie. the b-vectors, `yave` grids, rebinned model spectra, partial derivatives and covariance matrix, the pre-data stage identified in the previous chapter).

The SDSS will eventually consist of around 1 million spectra: It would obviously be advantageous to remove the precalculation overhead for these spectra. The Sloan data has some useful characteristics which help this process. Since the spectra are observed using a single spectrometer they have a reasonably well-fixed observed wavelength range. The exact range is dependent on observing run, but all galaxy spectra have information between 3815 and 9195Å in earth frame. MOPED requires the data to be in the emitted wavelength frame to allow fitting with spectral models, and for this to be the case the spectra must be shifted towards the blue end dependent on their redshift. For the range of redshifts in the survey of interest, $0 < z < 0.34$, this shift is up to ∼2300Å at the red end and ∼ 970 Å at the blue.

The Sloan data has a mean resolution of 2Å, far higher than typical spectral models. After blue shifting the data, we rebin the data down to 20Å resolution, increasing the S/N and allowing comparison with stellar models. By choosing the rebinned wavelength values to be on a standard grid, it is possible to use the same wavelength range for a set of galaxies over a short range of redshift. The data spreads over a redshift range of $0.005 < z < 0.34$. To determine the number of precalculated sets required it is necessary to see how many 20Å shifts are needed to cover the length from the lowest redshift minimum shifted wavelength to the highest,

$$N = Ceiling\left(\left(\frac{\lambda_{observed,min}}{1+z_{min}} - \frac{\lambda_{observed,min}}{1+z_{max}}\right)/20\right) + 1 \tag{3.1}$$

[1]Precompiled code where the source is proprietary and therefore cannot be compiled with the Condor libraries

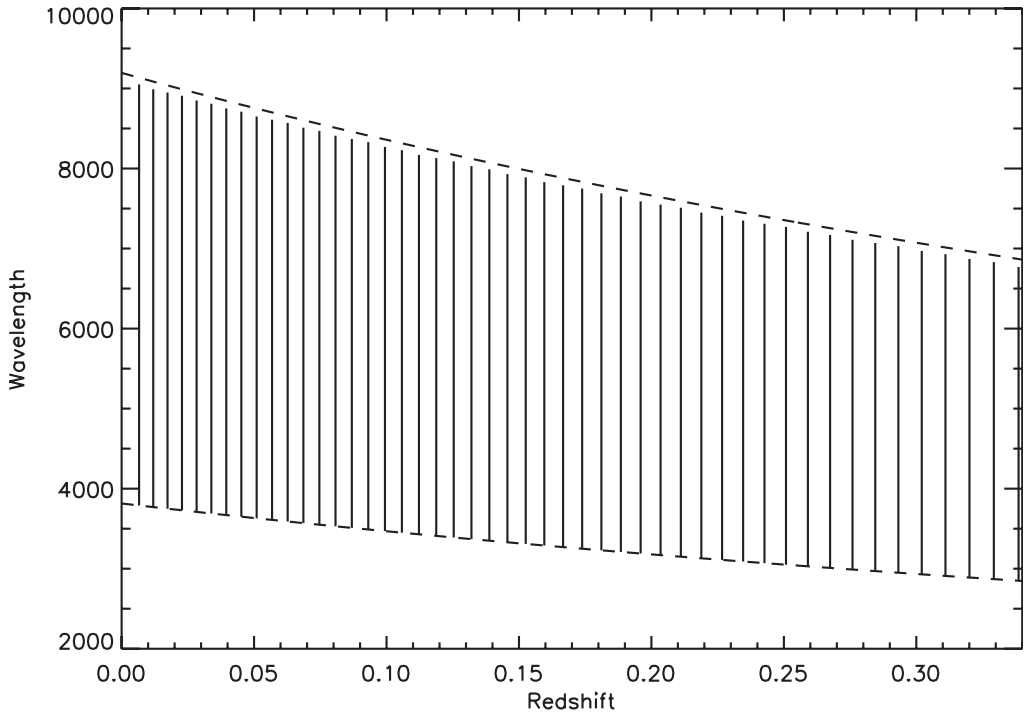


Figure 3.1: Precalculated sets. The figure shows how the range of rest frame wavelengths of galaxies in the SDSS sample changes with redshift (dashed lines), and how precalculated models can be used to cover the whole range when spectra are rebinned to 20 Å (solid lines).

The minimum observed frame wavelengths are used for this calculation. Ceiling refers to the computational function of increasing any decimal to the next highest integer. This method results in degenerate maximum wavelengths for a given minimum wavelength, but a reduced number of precalculated sets and therefore reduced processing time. To use the lower number of sets it is necessary to crop all data and the synthetic spectra used to manufacture each set to the level of the smallest maximum wavelength in each range. Although this may result in one lost data pixel at the red end of the spectrum, this end of the spectrum suffers contamination from skyline noise. Since the minimum spectrum still contains over 150 pixels, this corresponds to a loss of $< 1\%$ of the available data. It also reduces by a factor of two the number of precalculated sets required. Although when considering the amount taken to process the entire SDSS a doubling of precalculation time is negligible, the disk space required to store these precalculated sets is important. A full set totals around 500 Mb of disk space, doubling this number would limit the number of machines available for MOPED to run on. For future runs an ordering of spectra may relieve this storage requirement and allow more precalculated sets to be used. The sets used are illustrated in figure 3.1.

3.1.4 Creating a MOPED Pipeline

In order to run MOPED on the many galaxies of the SDSS it is necessary to create a pipeline from the various MOPED programs which can process any given spectrum without requiring interaction. This pipeline is driven by the program `sdss_handler_mcmc.pro`, and the code is given below:

```
pro sdss_handler_mcmc, sdssfile; , ntrials

;load dust model
 common dust_tab, dust_lookup
 restore, 'dust_model.idl'

 wave=dust_lookup[*,0]
 em_free=where((wave le 3680) or ((wave ge 3780) and (wave le 4820)) $
    or ((wave ge 5220) and (wave le 6480)) or (wave ge 6800))

 dust_lookup=dust_lookup[em_free,*]

;determine number of Conjugate Gradient (CG) and MCMC trials

 ntrials=50     ;CG
 n_mcmc=300,000 ;MCMC

;controls the sequencing of a MOPED run on a SDSS galaxy spectrum

 input_dataNM, sdssfile

 restore, 'Nmetsettings.idl' ;w_min, the precalculated set identifier, stored here

 spawn, 'cp ' + Preppath + 'mean_partial' + strtrim(string(fix(wmin)),2) $
     + '.idl ' + Savepath + 'mean_partial.idl'

 spawn, 'cp ' + Preppath + 'compress' + strtrim(string(fix(wmin)),2) $
     + '.dat ' + Savepath + 'compress.dat'

 spawn, 'cp ' + Preppath + 'yave' + strtrim(string(fix(wmin)),2) $
     + '.dat ' + Savepath + 'yave.dat'

 spawn, = 'cp ' + Preppath + 'bin'+strtrim(string(nbin),2) $
     +'Mspec' + strtrim(string(fix(wmin)),2) + '.idl ' + Savepath $
     + 'bin'+strtrim(string(nbin),2)+'Mspec.idl'

 runminspectraII, sdssfile, ntrials

 restore, FRespath + 'MOPED-' + sdssfile + '.dat'

 str1=best_sfh ;stored in file above

 res=mcmc(str1, N_mcmc, specexp, expwave, sdssfile)

 save, res, norm_val, filename = (Respath + sdssfile+'-errors.idl') ;save local
 save, res, norm_val, filename = (Rempath + sdssfile+ '-errors.idl') ;save remote

END
```

Since there is a precalculated set for each possible (20 Å rebinned) rest frame wavelength, once `InputDataNM.pro` has been used to import and blue shift the spectrum the minimum wavelength of the spectrum (before emission line extraction) can be used to assign the correct set to the spectrum. This minimum is used to copy the correct precalculated set to the MOPED working directory. `runminspectraII.pro` then uses this set to compute solutions from random start points on the MOPED likelihood surface using the conjugate gradient method. The best solution

is stored along with various parameters and any warnings that were generated in a datafile. The pipeline reads this file and passes the best solution to `mcmc.pro`, which then uses the solution to start a Markov Chain to explore the surface. The solution and errors found by the chain are then written to both the local machine and the machine being used to collate results and the pipeline is complete.

3.1.5 Royal Observatory Auto-MOPED

While the Condor project provides a useful environment for distributed computing, it is not ideal for running MOPED as an HTC application. Learning from experience with condor, I determined that for MOPED to work in a HTC environment, that environment should:

1. Automatically reduce the load from MOPED when the owner of the computer required CPU time.
2. Recover without intervention if a computer on which it was running upon was rebooted
3. Require minimal traffic of data over the network after initial set up
4. Have some mechanism to balance load between the various different specification computers available to it.
5. Collate all output to a single area to allow rapid collection of data when the task was finished
6. Be robust such that the failure of any computer should not affect the others.
7. Be scalable when further computers were available.

The Royal Observatory Auto-MOPED (ROAM) environment attempts to satisfy these constraints. Written in IDL's scripting language, it has allowed MOPED to process over half a million jobs on up to 35 computers. Although it was written with MOPED in mind, it has already proved useful in other HTC tasks at the observatory which use IDL[2].

ROAM consists of several programs which run on each computer of the parallel system. Although there is no central computer, the system is reliant on a shared filesystem which has to exist on one computer only. To reduce network load during the analysis, several pieces of information are stored on each computer. The first phase of ROAM runs from one machine, and distributes the precalculated b-vectors (around 0.5 Gb) and a numbered list of spectra to temporary local storage on the other machines. It also distributes its own program to each machine, and resets the environment to be ready to start a new run.

The second phase of ROAM is run individually on each machine, and reads the environmental parameters `current_limit` and `block_size`. These two parameters give the limit of allocated spectra, and the size of each block which will be allocated. It then allocates itself the spectra from `current_limit` to `current_limit + block_size`, and changes `current_limit` to `current_limit + block_size` ready for the next computer to be allocated spectra. It stores the number of the spectrum it is working on, and will recover automatically if rebooted. When the block is finished the process restarts and the computer is allocated another block.

This process can be illustrated by the following program:

[2]Michael Davidson's serendipitous XMM Cluster finding routines, running simultaneously on ~ 40 machines.

```
PRO CJ_MID

 common SDSS_I
 spawn, 'hostname', abc
 name_end=strpos(abc, '.')
 name=strmid(abc, 0, name_end)
 block_size=200

 restore, '~/DR1_GALS.idl'          ;open list of jobs

 for n=0,300 do begin               ;set upper limit of 300 blocks

    restore, '~/S_E_V/'+name+'.idl'     ;read Start and End Values
    if s_e_v[0] eq s_e_v[1] then begin  ;check for end of block

        restore, '~/curr_lim.idl'
        curr_lim2=curr_lim
        curr_lim=curr_lim+block_size
        save, curr_lim, filename='~/curr_lim.idl'

        s_e_v[0]=curr_lim2
        s_e_v[1]=curr_lim

        save, s_e_v, filename='~/S_E_V/'+name+'.idl'
        endif

    for sdss_id_num = s_e_v[0], (s_e_v[1]-1) do cj_pro, names[sdss_id_num], &
         sdss_id_num, name

    command='echo '+ abc + ' completed at ' + systime() + ' now working from '&
        + string(s_e_v) + ' | sendmail bdp@roe.ac.uk'

    spawn, command

    endfor

 command='echo '+ abc + ' completed at ' + systime() + ' now out of
 tasks | sendmail bdp@roe.ac.uk' ;notify if finished

 spawn, command

END
```

The program above sends an email on block completion and job completion.

The shared file space (in this case `~/S_E_V/`) where these environmental parameters are maintained is the weakest link in ROAM. Should this main server fail, no more blocks of spectra are allocated. In fact this server is one of the most important on the site, and is very reliable. If it does fail, each computer attempting to access the `curr_lim.idl` parameter file will just wait until access is possible. Similarly, if the data storage area fails, MOPED will wait until access is available again before being able to continue working.

In order to run on many machines it is necessary that ROAM does not interfere with the machines' owners' work. To accomplish this the LINUX program `nice` was used to give ROAM tasks low priority. `nice` works by assigning each task a level of "niceness" from -19 to 19. The

lowest niceness tasks are allowed the most CPU time: if `nice` is given two identical programs A and B with respective niceness 1 and 5, time will be preferentially allocated to A. CPU time is not the only thing that determines usability of a computer. Some programs may be heavily dependent on memory and disk access, which is not controlled by `nice`. MOPED does not fall into this category, requiring only a small amount of memory and relatively little disk I/O. Allowing ROAM to use a computer with a `nice` value of 5 is virtually unnoticeable to the machine's owner.

A feature of the ROAM environment is that as long as the block size is small, the system will automatically load balance. If one machine is particularly slow, then it will not request as many blocks as a faster machine. The system is scalable to the point where reallocating blocks of spectra is attempted by multiple machines simultaneously. Since this is a very quick operation, the block size can be very small. For simplicity of tracking, I chose a block size of 100, although it could be made smaller.

3.1.6 Improvements to ROAM

ROAM was written relatively quickly to allow MOPED to run on many machines simultaneously. It works very well, and I was reluctant to spend more time tweaking the environment. There are however, a number of ways that it could possibly be improved.

Redshift Ordering of Galaxies

The precalculated sets (see previous chapter) are large and it would be beneficial if not all were required to be stored on each machine. Since the order of processing is not important, it would be advantageous to order the spectra in redshift. This would mean that, for a given block, all spectra would require the same precalculated set. The copying phase which moves the sets takes only a small period of time, but when multiplied by 1,000,000 spectra the few seconds required becomes significant. By maintaining a library of sets on a single machine and copying one set across the network every block (several hours) approximately half a gigabyte would be saved on each of the machines assigned to MOPED.

Improved Reporting

Although overall ROAM is very efficient at getting all spectra processed, it is difficult to analyze the productivity of any one machine. It would be useful to be able to show which machines were under performing. Although the cause of the slow down could be from high load from the machines user, it can also be a symptom of low swap space or insufficient `nice`ing on the part of another remote user. If the owner of the ROAM job is made aware of this, steps can be taken to remedy the situation.

Remove Reliance on Central Server

Although every effort has been made to remove reliance on any one machine in the ROAM cluster, there is one machine which is essential – the central file space server. Redundancy could be

built into the system by having a method which writes the job description data to a separate location. If the primary source of information is compromised, ROAM would automatically read and write from the second.

Astrogrid

As HTC develops, it increasingly requires processing time which is unavailable at individual institutes. The Astrogrid[3] project aims to seamlessly integrate data storage, processing and result reporting through GRID computing techniques. The general idea is an extension of the Condor ideal - it does not matter where the machine is that does the processing, just that it gets done. Although this project is still in its infancy, MOPED has been chosen to be incorporated as a test bed. In collaboration with the Astrogrid team I hope to make a GRID implementation of MOPED.

3.2 Tools for Combining the Star Formation Fractions of Galaxies

MOPED analysis of the SDSS has yielded a huge amount of data. To be able to usefully analyze this information, and its reliability, various statistical techniques have been incorporated and developed. This sort of problem is known as *data mining*: the process of extracting concise information from a large dataset. More data is available to astronomers than ever before, and in the future techniques such as these will be required to perform accurate analysis.

3.2.1 Rebinning to the Earth frame

MOPED determines the rest frame SFH of galaxies. This is interesting if the object of the study concerns the intrinsic properties of individual galaxies, but variations in redshift mean that for a more general analysis it is necessary to shift these results to a common set of time bins. For simplicity the set of time bins chosen are the same as those used in the galaxy analysis. This allows direct comparison of galaxy star formation in terms of cosmic time, and universe-wide conclusions to be drawn

By combining the total mass created in each galaxy with the MOPED Star Formation Fractions (SFF) a mass created in each time bin can be generated. By overlaying the galaxy rest frame time bins with the observed (earth) frame time bins and redistributing the mass between the new time bins the earth frame SFH can be determined.

In order to determine the new mass distribution the fraction of mass in each original bin contributing to each shifted bin must be calculated. Four cases arise from the different relative positions of the Earth Frame (EF) and Galaxy Frame (GF) bins.

1. The EF bin may overlap the bottom edge of the GF bin
2. The EF bin may overlap the top edge of the GF bin
3. The GF bin may lie totally within the EF bin

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

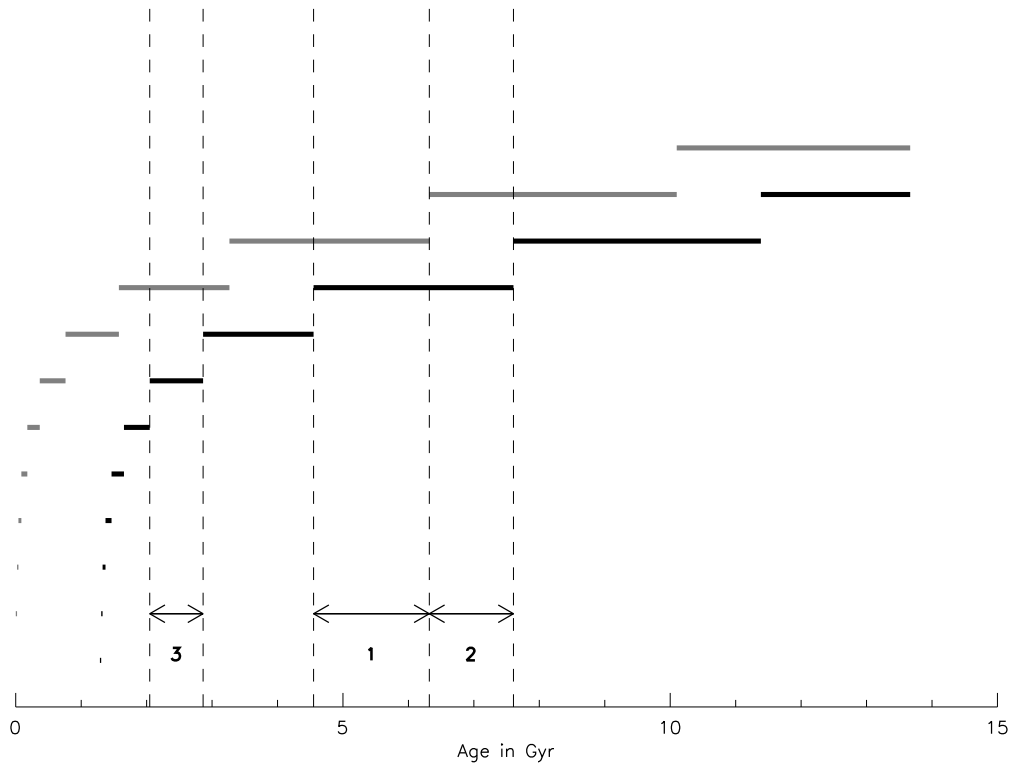


Figure 3.2: Rebinning between rest and observed frame. The grey lines represent the EF time bins which are common to all galaxies. The black lines represent the GF time bins for a galaxy at $z = 0.1$. Note the upper boundary of the oldest bin is always 13.6 Gyr, the age of the Universe, and that no information for the youngest bins in the EF is given by this galaxy as no signals have been received for this period. The annotations refer to the overlap descriptions in the text.

4. The EF bin may lie totally within the GF bin

The two sets of bin boundaries for a galaxy at $z = 0.1$ are shown in figure 3.2. The first three conditions outlined above are labelled, although the fourth does not appear at this redshift.

These different cases are identified by several `if...then...else` loops in the program `redshifter6e_m.pro`. The mass is then redistributed by assuming constant generation of mass across a time bins in linear time. For each bin in the original GF time bins, the fraction lying in each of the new EF bins is calculated - for example for the 9th GF bin in 3.2, just over half the mass will go to the 9th EF bin and the remainder to the 10th EF bin (cases **1** and **2**). The boundaries of the 7th GF bin lie totally within those of the 8th EF bin, so in this case all the mass will be transferred (case **3**).

There is obviously some area where the oldest bin in the GF is shifted beyond the limit of the oldest EF bin. This "overspill" mass is put into the final EF bin. This method ensures conservation of mass through the rebinning process. There is very little difference in a spectrum from either end of the age range covered by the bin (10 - 13Gyr), and these oldest GF bins should be considered as having width to whatever the age of the Universe was at that redshift.

3.2.2 Completeness - $V_{\max}$

In a magnitude limited survey such as the SDSS the range of galaxy types and sizes included in the survey will vary over the redshift range studied. Some mechanism is required to compensate for this change and determine the overall bulk parameters for the sample. The $V_{\max}$ method attempts to correct for this problem in order to enable the average properties of a large sample of galaxies to be calculated without biassing results to those galaxies which are over represented due to their physical parameters.

The $V_{\max}$ method gives each galaxy a weight equal to $1/V_{\max}$, where $V_{\max}$ is the maximum volume of the survey in which the galaxy could be observed. This gives an unbiased estimate of the space density f of the property F of the galaxy under investigation, be it mass, luminosity etc:

$$f = \sum_{\mathbf{galaxies}\ i} \frac{F_i}{V_{max,i}} \tag{3.2}$$

On smaller scales the estimator is affected by source clustering, but the SDSS is deep enough that these variations should not be significant. Any properties which change with redshift and which could determine inclusion in the sample must be calculated for each galaxy over the redshift range of the survey to determine whether or not it would have been included. Our selection criteria are those of Shen et al. (2003). Our main galaxy sample is determined by red apparent magnitude limits of $15.0 \leq m_r \leq 17.77$, and we also place a cut on surface brightness of $\mu_r < 23.0$. The magnitude limits are set by the SDSS target selection criteria, as discussed in Strauss et al. (2002). The target criteria for surface brightness was $\mu_r < 24.5$, although for $\mu_r > 23.0$ galaxies are included only in certain atmospheric condition. In order to remove any bias we have therefore cut our sample at $\mu_r < 23.0$. At low redshifts the Sloan galaxies are subject to shredding - where a nearby large galaxy is split by the target selection algorithm into several smaller sources. To reduce this effect, for our star formation analysis we use a range of $0.005 < z < 0.34$.

In order to calculate the $V_{\max}$ assigned to each galaxy it is necessary to consider the apparent magnitude and surface brightness evolution over the redshift range. In order to compute this we use the same stellar evolution models used in the MOPED analysis to calculate luminosity over the life time of the galaxy due to its recovered star formation history.

Evolving the Luminosity and Surface brightness of Galaxies

The magnitude of a galaxy over its lifetime depends on the luminosity behaviour of the various stellar populations that make it up - the star formation history. As young stars, the populations will have a very high light output, which will reduce as they age. This information is encoded in the galaxy spectrum and recovered by MOPED, which gives the relative fractions of different aged populations. To compute the observed magnitude if the galaxy were to be at higher redshifts, we need to evolve the models over time and track the changes in luminosity. Obviously, as the galaxy is projected to a further redshift the younger fractions do not contribute, as galaxy is being "observed" before these populations were born. Since the spectral energy distribution of the light changes with evolution of the star, it is also necessary to apply the filters used by Sloan to determine the flux included in the r band.

The program used to evolve the luminosity was originally written by Raul Jimenez, and is called `mag_corr.pro`. I adapted the code to run for multiple galaxies and with any time bin

structure. It provides the magnitude corrections required for a grid of redshift values.

$$r_{i,z} = r_{i,z_{obs}} + c_{i,z} \tag{3.3}$$

These magnitude corrections can then be used to calculate corrections which need to be applied to the surface brightness. The surface brightness of the galaxy at each redshift z is calculated from the evolved luminosity correction $c_{i,z}$as

$$\mu_v = r + c_{i,z} + 2.5\log_{10}\left[\pi r_{50}^2 (dl_z/dl_{z_{obs}})^2\right] + 2.5\log_{10}(2) \tag{3.4}$$

where r_{50} is the petrosian half light radius and z_{obs} is the observed redshift of the galaxy. This algorithm assumes that the size of the galaxy does not change over the redshift range - although this assumption is valid for low redshift sources, it will need to be developed if the technique is applied to deeper surveys.

Interpreting the $V_{\max}$ weighting

The MOPED technique gives the relative strengths of the different spectral models. The mass originally created to make these masses is then calculated, and by dividing this by the maximum volume over which the galaxy could be observed gives the star forming density, ρ_i. By adding all the $V_{\max}^{-1}$ weighted star forming densities of galaxies in the sample together the overall star forming density ρ can be found for the region studied.

Determining the SFR

1. Use MOPED to recover Star Formation Fractions (SFF)
2. Calculate the Star Formation Mass Fractions (SFMF)
3. For each time bin, convert SFMF to Star Formation Rate (SFR) by diving by the duration of the time bin
4. Divide by Vmax to give contribution to star forming density (ρ_i) in that time bin
5. Integrate over all galaxies to give ρ for entire sample.

It is clear from figure 3.2 that there are always some bins in the earth frame for which we have partial or no information. We chose to ignore these bins, and have altered the $V_{\max}$ calculations accordingly: those bins are not included in the calculation of ρ. The equation to calculate the star formation density, ρ, in time bin j from n galaxies is:

$$\rho_j = \sum_{i=1}^{n} \frac{(SFMF)_{i,j}}{\Delta t_j} \frac{1}{V_{\max,i,j}}, \tag{3.5}$$

where Δt_j is the width of the bin in look back time.

3.2.3 Bootstrap Error Calculation

The formal statistical errors on individual ρ_i measurements are dependent on the shape and degeneracies of the hypersurface likelihood peak. These errors are calculated by looking at the

portion of the hypersurface which is within a certain height range of the peak. Since the peaks are generally very narrow, this method yields error estimates that are either huge (including the degenerate peak) or tiny (characterizing the global maximum) rather than a range of sensible estimates. The root of this difficulty is that although the stellar models can provide fits to the spectra which look reasonable they are rarely a formally good fit in the statistical sense. There are a few regions of the spectra which the models find difficult to fit. To deal with this we have chosen to use a bootstrap technique which calculates the error using the spread of estimates rather than their formal errors. This technique allows an error to be estimated for some bulk quantity obtained from a large set of data without requiring individual errors for each datum, assuming that the set of data collected properly samples the distribution of the complete set.

$$\rho = \sum \rho_i \tag{3.6}$$

Where there are a large number of ρ_i values, it is expected that their distribution is such that a large enough subset should have the same distribution as the original set. By calculating the bulk parameter from a randomly re-sampled set of ρ_i a further estimate of the parameter is obtained. Repeating this process many times gives multiple estimates of the parameter based on different subsets of the whole dataset (Press et al., 1992). The variance of these estimates should be equal to the variance of the parameter, as long as the distribution of the original set is the same as the real distribution.

The bootstrap calculation is illustrated by the following section of code, taken from `DR1_SFH_vmax_new.pro`. In this example, `sfh_contrib` is a vector of ρ_i values which are used to calculate ρ.

```
estimates=fltarr(n_boot)

for b=0,n_boot-1 do begin   ;need to run n_boot times
     resample=round(bin_count[param]*randomu(seed, bin_count[param]))
     new_sfh_contrib=sfh_contrib[resample]
     estimates[b]=total(new_sfh_contrib)
     end

average_see_p=moment(estimates) ;moment returns the mean,
standard deviation etc of the vector "estimates"

average_see[param]=sqrt(average_see_p[1])
```

A further advantage of the bootstrap error technique is that it requires no modification to calculate errors on $1/V_{max}$ weighted results - the weighted results are used as the sample instead.

3.2.4 Jackknife Errors

A slight modification to the bootstrap error technique yields the Jackknife method, used for the calculation of errors on the mass function data presented later. The sample is cut into X equally sized samples. The method begins by throwing out the first sample of the data, leaving a jackknifed data set of resampled values. The statistical analysis is performed on the reduced sample, giving a measured value of the bin height. Next a new resample is taken, this time throwing out the second measurement, and a new measured value of the parameter is obtained. The process

is repeated for each set in the sample, resulting in a set of parameter values c. The standard deviation of these estimates, σ_s, is calculated. The standard error for the estimate on the complete sample, σ_c, is given by the formula

$$\sigma_c = \sqrt{\sigma_s \frac{(X-1)^2}{X}} \tag{3.7}$$

3.2.5 Database Linking

Although interesting in isolation, the real power of the data extracted from the SDSS spectra by the MOPED algorithm becomes apparent when combining or comparing changes in SFH with various other galaxy properties. To enable this comparison it is necessary to compare our database to the various recovered parameters in the main Sloan database. Since both databases are greater than a Gigabyte in size, some method is required to speed up comparisons. A simple and effective way is to generate an index which links each galaxy included in our sample to its corresponding SDSS record. By combining the date of the observation (`mjd`), plate (`plate_id`) and fibre number (`fiber_id`) a unique Sloan identifier can be generated in the form `mjd-plate_id-fiber_id` (York et al. (2000)). Data from the Sloan survey is received as a large table, and is matched to MOPED records through the program `DR1_find_matches.pro`. The index generated by this program is then used to generate tables of data for direct comparison to the MOPED results. This program performs the matching operation by three sequential searches carried out for each galaxy. For a galaxy in the MOPED sample, the search first selects all sloan records with a matching mjd. In those records, the program searches for matching plate numbers. The final step is to search in the galaxies which have passed the first two stages for the complete match. For comparison, this approach was tested against another program which used all the identifies in one long number. The stepped search was found to be ~ 3 times faster.

3.2.6 Memory Issues

For the current EDR and DR1 analysis, the physical size of the combined database is just small enough to be loaded concurrently on a LINUX workstation with 1Gb of RAM and a similarly sized swap disk. For the future it will be necessary to either perform the postprocessing on a more highly specified machine or split the database of MOPED galaxies into pieces and work on several machines in parallel.

3.3 The Trials of MOPED

Although MOPED has been extensively tested before (see Reichardt et al., 2001, for examples), further tests were conceived in the course of my work on the SDSS. The tests described in this section examine both the underlying technique and the treatment of the results for the MOPED analysis of Sloan spectra. We find that the method is reliable and efficient, even when given very noisy data to work with.

3.3.1 Modelling

Stellar Modelling

At the heart of the MOPED analysis are the stellar models. Stellar modelling is certainly not in its infancy, but between the various groups who present state of the art modelling there is still some discrepancies. In order to test the robustness of our findings with respect to these differences, I reconstructed the star formation history changing the theoretical stellar populations models to those of Bruzual & Charlot (2003). This allows an assessment of the effect of systematic modelling errors in our determination of the SFH, since the Jimenez and Bruzual-Charlot models are based on different stellar interior and atmospheres models. Reassuringly, the shape of the recovered star formation history is hardly changed by using Bruzual-Charlot stellar population models.

Overall Metallicity

Although the stellar modelling codes used are considered to be wholly accurate up to double solar metallicity, beyond this there is some argument over their ability to predict spectra. We tested this by calculating the average metallicity in the bins where star formation is found - the metallicity of the bins with no star formation is not constrained at all, and is not a reflection of any physical parameter of the galaxy. The numbers are extremely reasonable, peaking at $Z_{star}/Z_{\odot} = 1.92$, with the majority near to solar - well within the range supported by the Jimenez et al. (2004a) models. This crude analysis was undertaken without $1/V_{\max}$ weighting, and it is likely that the average would decrease further if this were to be included.

3.3.2 Averaging the Spectra

We argue that although for an individual galaxy, there may be some covariance between the oldest bins, this fluctuation is purely statistical, and by averaging over many spectra the affect can be removed. In order to test this, I created a number of spectra with the same star formation history as that expected from the global SFR. The spectra were distorted by noise, first as a simple random noise then with more complex characteristics designed to mimic the typical SDSS spectra.

Simple S/N Test on Spectra

Figure 3.3 shows the Star Formation History (SFH), recovered from 300 simulated galaxy spectra with the best-fit global SFR from the SDSS. Each galaxy has a different Gaussian noise realisation, applied such that the average S/N is 40, and the results are weighted by likelihood. The parameter recovery is very good, including the intermediate-age population, which indicates that the method is not biasing the results or leading to a broader recovered star formation rate peak. NB low-z bins have few galaxies contributing, so errors are larger.

Complex Noise Model

A more stringent test of the MOPED with SDSS spectra requires a more complex treatment of noise on the spectrum. The differences between these spectra and those used above are:

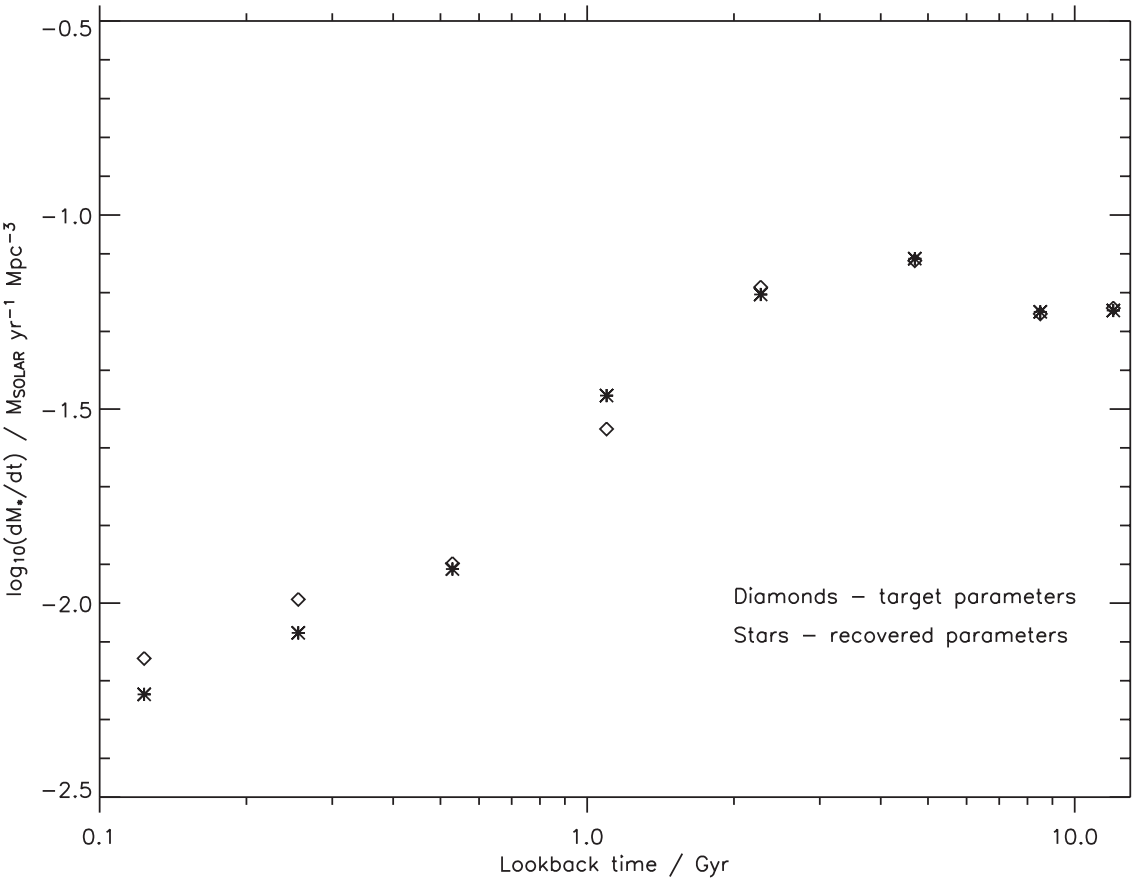


Figure 3.3: Average recovered SFF parameters for a spectrum with simple noise. 300 synthetic galaxies were created with Gaussian noise at a S/N of 40, and their SFF recovered using MOPED.

- Accurate Sloan Noise. Instead of uniform strength Gaussian noise over the spectrum, this test uses noise which is scaled to the shape of the SDSS typical noise curve (Figure 3.4), rising at either end of the spectrum and with some noisy areas which are contaminated by skylines. The normalisation of the noise is carried out such that the average signal to noise over the whole spectrum is 40.

- Blue End Calibration. There is thought (Abazajian et al., 2004) to be calibration problems with the blue end of the spectrum below about 4000Å, which may lead to a random continuum slope error of as much as 10% in some cases. To create this affect in the synthetic spectra a random tilt was applied to the spectrum below 4000Å. The tilt, defined as the maximum deviation from the original level, was different for every spectrum, selected from a Gaussian centred on 1 with a standard deviation of 0.12.

- Emission Line Filling. In some fraction of our sample there will be some contamination from AGN and hot gas emission lines. Although we mask the main regions subject to such contamination from our analysis, we do not mask the area of the H-δ line. This line is traditionally associated with intermediate age stellar populations, and filling from AGN may interfere with our determination of the presence of an intermediate population. To compensate, the 20Å pixel containing the line was filled according to the upper half of a

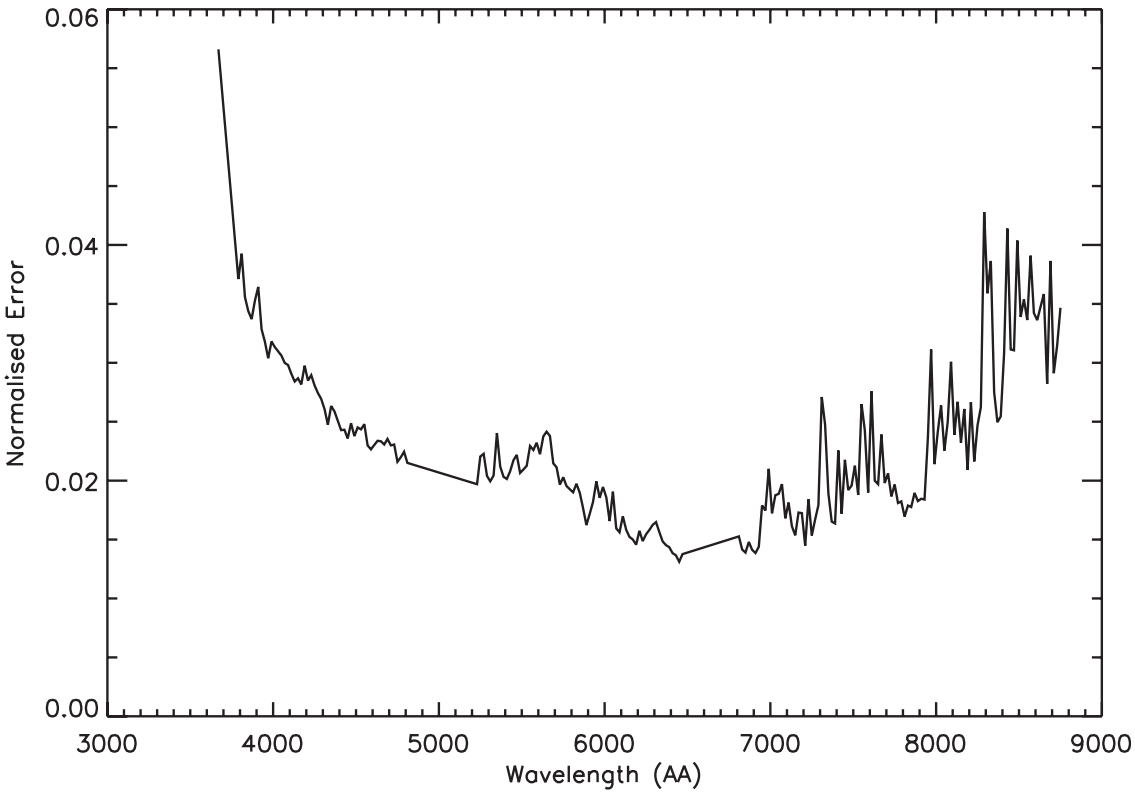


Figure 3.4: Example of SDSS spectrum noise, normalised to the galaxy signal flux at 5500Å. Note that this spectrum has been blue shifted, and that the features will fall at different wavelengths for galaxies at different redshifts. The regions removed from the spectrum to eliminate emission line contamination are shown by lines of constant gradient.

Gaussian centred on the original height with standard deviation the approximate distance between the pixel and a smooth continuum.

Even with this more complex noise MOPED was able to recover the SFH of the mangled spectra remarkably well with only a few hundred spectra, showing that although the adjacent bins of individual galaxies can be covariant in the presence of excess noise, the degeneracies are not biased and can be eliminated by using many spectra.

Recovering Different z Peaks

In order to test whether MOPED biases the peak of SF to a lower redshift I created various spectra with different SFH from our estimate for the global SFR. One of these new spectra is based on the previously accepted cosmic SFR; results are shown in figure 3.5. In all cases given ~ 300 spectra MOPED was able to resolve the SFH accurately, and no tendency to shift towards the recovered global SFH over many trials was observed.

3.3.3 Aperture Bias at Low Redshift

Although we felt that Glazebrook et al. (2003) addressed the issue of aperture bias, a further test examines the star formation fractions of the oldest few bins with redshift. If there is an aperture

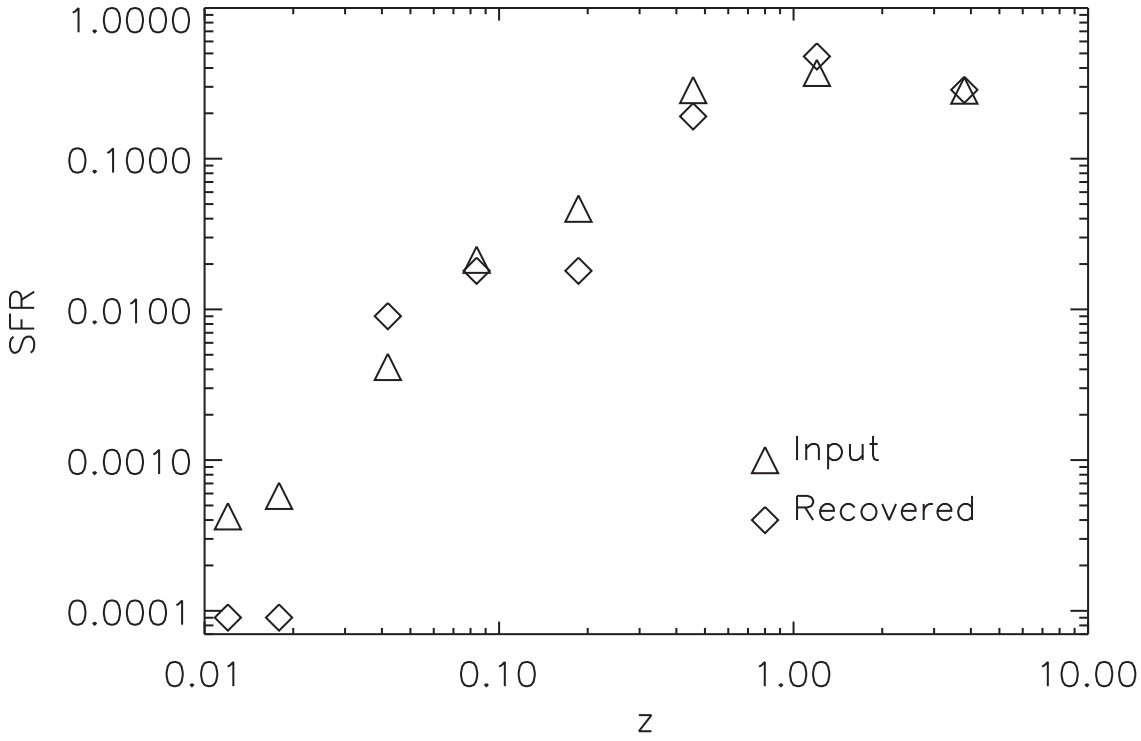


Figure 3.5: Average recovered parameters with complex noisy spectra from 300 galaxies. In this case the input spectrum was created with a SFH which matched that of previous determinations of the cosmic SFR.

bias these bin fractions will vary. In order to minimise other effects we have taken the average of the SF fractions recovered for two sets of similar mass galaxies at differing redshifts. To be able to compare these two directly it is important to shift their SFF to the earth frame, and of course the recent SFH of the most distant set cannot be compared, as we cannot extract SFH of these galaxies after their light has been emitted. As Galaxies reach a higher redshift, their younger bins will not show in the earth frame. Figure 3.6 demonstrates that there is no change in the normalized SFF over the redshift range 0.01 – 0.15 for similar mass systems.

3.3.4 Volume Limited Samples

The V_{max} correction is applied to correct bias introduced by sampling over a range which will include different galaxies at different redshifts. In order to test that the $1/V_{max}$ weighting does not introduce any problems I constructed some volume-limited samples. Figure 3.7 shows such samples for three redshift ranges, $z = 0.05 - 0.07$, $0.08 - 0.1$ and $0.1 - 0.15$. Each sample probes a different luminosity range and therefore different ranges in mass. Because of this, we should expect some significant differences between the samples. Indeed, we see the expected trends: as we move to more distant volume-limited samples (probing more luminous galaxies), the peak of star formation is pushed to earlier times, in agreement with our results for galaxies of larger stellar mass. These are less informative than the $1/V_{max}$-weighted result, as many galaxies are removed, but they show the same basic result, and this gives confidence that the findings are robust.

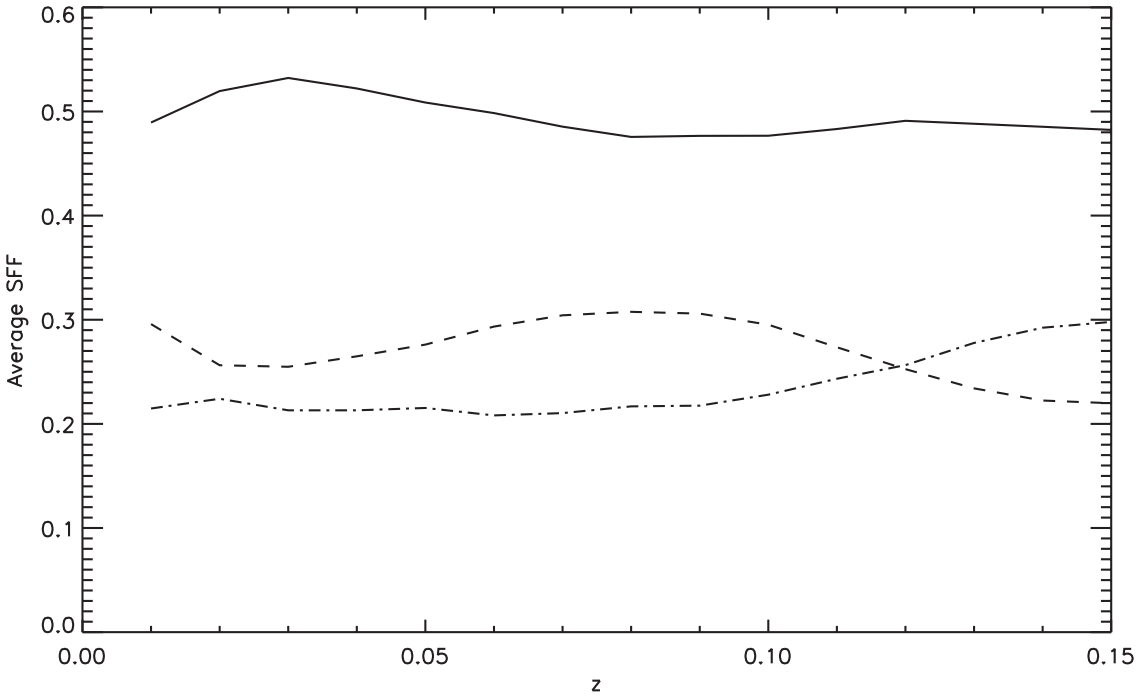


Figure 3.6: The SFF in each of the top three bins remain static between similar sources at different redshifts. The size of galaxy chosen for this test had observed mass $(0.95 - 1.05) \times 10^{11} M_{\odot}$

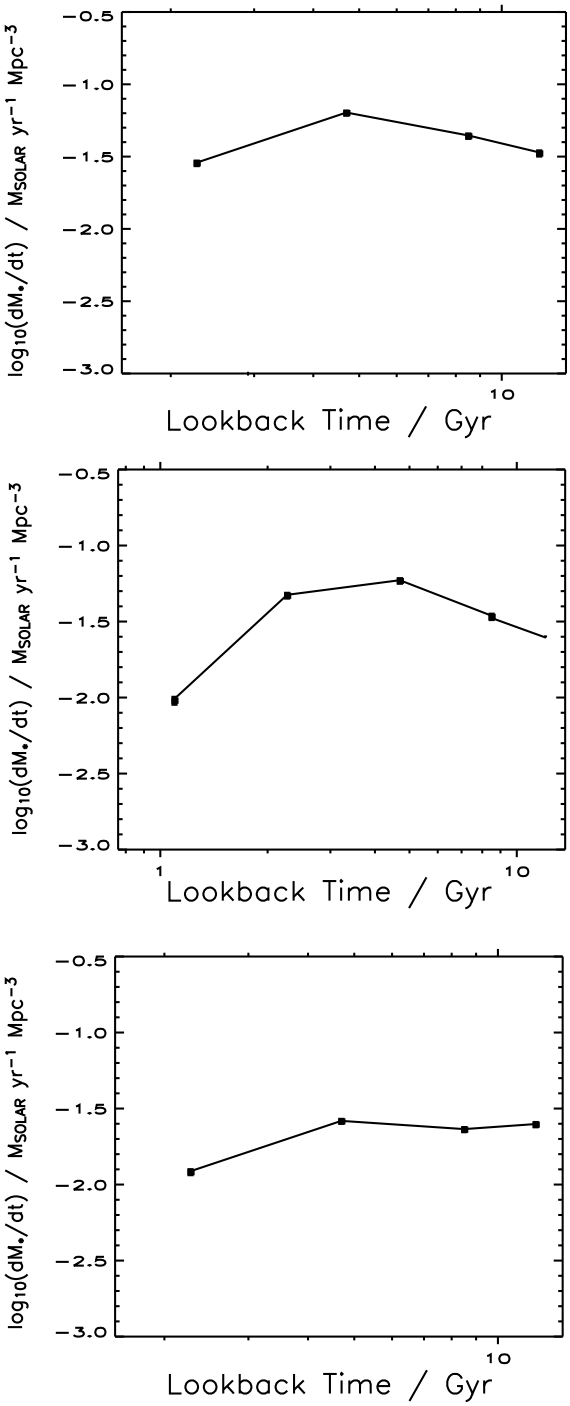


Figure 3.7: SFR in Volume Limited Samples. Using volume limited samples to contruct the SFR for different ranges of mass in order to check SFR-mass relationship. Redshift ranges $0.05 - 0.07$, $0.08 - 0.1$ and $0.1 - 0.15$ are plotted from top to bottom

CHAPTER 4

Results

The MOPED analysis of the DR1 can be exploited in many ways. In this chapter I present the results of the analysis in terms of the cosmic star formation rate, the distribution of galaxies' stellar mass components and a comparison between the build up of baryonic matter and dark matter in galaxies.

4.1 Cosmic Star Formation Rate

The cosmic SFR is of key importance for studies of galaxy birth and evolution. By determining the SFR as a function of cosmic time or redshift, global properties associated with the Universe and the conditions which lead to galaxy formation can be studied. Figure 4.1 shows the co-moving star formation rate determined by MOPED and SDSS DR1, as a function of redshift. For comparison, the SFR estimated from other determinations is also plotted. These estimates are for the most part based on instantaneous star formation rate indicators (UV flux, Hα emission, sub-mm emission etc). The two sets of indicators, calculated in totally different manners, give good general agreement.

The decline of the SFR at low redshift is clear, agreeing well with the local determination of the SFR from H-α surveys. The decline also seems to be flattening, as would be expected at recent times - there is little evidence from local galaxies for a large change in star-forming activity in recent times.

The high-redshift behaviour of the SFR is very similar to that expected from the studies using the instantaneous SFR of high redshift galaxies. The SFR from these galaxies has been estimated in a completely different way from our results, and the various different techniques involve corrections which, with only slight tweaking, agree readily with our determinations. In particular the dust corrections made to the high redshift galaxies are not well known, and the mass function of galaxies is not well determined at such high redshifts. Corrections for each of these typically involves a multiplicative factor of ~ 5. Small adjustments in these two large corrections would give excellent agreement in the high redshift environment. It is possible to estimate the amount of dust obscuration which can feasibly be applied in the analysis of high-redshift samples of Lyman-break galaxies from the difference between the two methods. Jimenez et al. (1999a) proposed that starburst galaxies are either obscured almost completely (like SCUBA sources) or the dust correction is only a factor 6 for moderate starbursts. Comparing the SFR's presented here with that of the Lyman-break galaxies gives support for this factor 6 dust correction with the remaining missing star formation being provided by SCUBA sources, which effectively have extremely large dust corrections.

The agreement between these two independent techniques gives further insights into our understanding of the Universe than just the SFR. The Copernican Principle is supported as the estimates determined locally produce the same star formation rate as that found at large distances. This suggests that we are correct to assume that we are not privileged observers, and that the parameters which determine the cosmology of the Universe do not depend on location. It is interesting to note that our method complements the others in use of the initial mass function. The fossil method infers the number of early-forming high-mass stars, long since gone supernova, from the numbers of early-forming low-mass stars, which are still contributing to the galaxy spectrum. The traditional approach moves along the IMF in the opposite direction, observing either the highest mass stars or their effects, and estimating those that populate the lower mass end. That both answers come up with similar results suggest that, at least in broad terms, the shape of the IMF is correct, and that the Copernican Principle holds.

At the high and low redshift ends of the SFR plot, we have excellent agreement with other studies. There is a slight difference in our recovered values in the mid-range of redshift. We see that the period of star formation is longer than previously thought, and that the peak occurs at a lower redshift $z \simeq 0.6$, rather than 1 or more. Specifically, we find that 26% of the mass of stars in

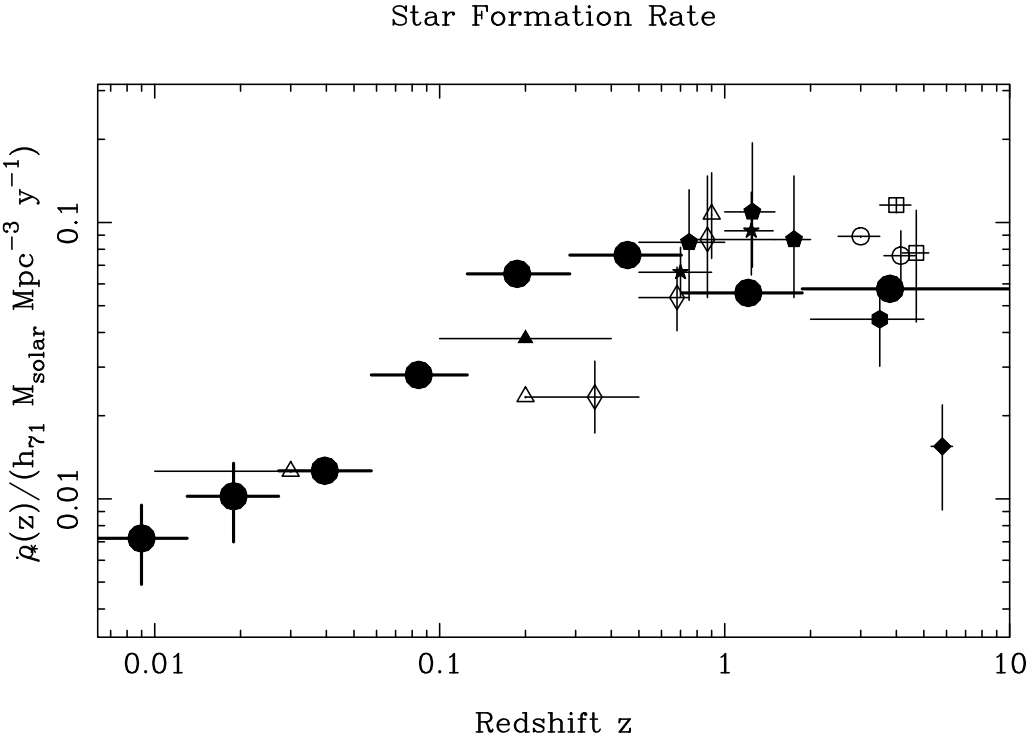


Figure 4.1: The SFR of the SDSS DR1 recovered using MOPED is shown by the eight large filled circles. The horizontal error bars represent the size of the bin in redshift. Vertical errors are bootstrap errors, and are invisibly small for most bins. The other symbols correspond to independent determinations using instantaneous measurements of the star formation rate, as follows; $H\alpha$ measurements are open triangles at $z \simeq 0.03$ (Gallego et al., 1995), $z \simeq 0.2$ (Tresse & Maddox, 1998), $z \simeq 0.9$ (Glazebrook et al., 1999); UV from Subaru (Ouchi et al., 2004, open squares) , GOODS (Stanway et al., 2003, filled diamond), HST etc (Steidel et al., 1996, open circles), CFRS (Lilly et al., 1996, open diamonds), HDF (Connolly et al., 1997, filled pentagons), galaxies (Cowie et al., 1999, stars), galaxies (Sullivan et al., 2000, filled triangle). The filled hexagon at $z = 3.5$ represents a new estimate of the star-formation density provided by sub-mm galaxies in the redshift range $2 < z < 5$. This was derived by integrating the sub-mm source number counts (Scott et al., 2002) down to $S_{850\mu m} = 1$ mJy, and assuming that 75% of such sources lie at $z > 2$, in line with recent redshift measurements (Chapman et al., 2003)

the present-day universe was formed at $z > 2$ (cf. Dickinson et al., 2003).

As we explain below, we believe our result differs because it includes the contributions made by all galaxies over a very wide mass range, extending down to galaxies with $L \sim 2 \times 10^{-3} L_*$. Note that virtually all 96545 galaxies contribute to the $z > 0.3$ bins, so the statistical errors shown for each estimate, determined via a bootstrap analysis, are negligible in comparison with modelling uncertainties and residual uncertainties in flux calibration of the SDSS spectra. As a reality check, I have computed the average metallicity of gas which makes stars: it rises from $\frac{Z}{Z_\odot} = 0.44$ at high redshift to a peak of 0.8 at $z \sim 1$ before declining to a level around 0.25 at the present day. We also note that because we are not dominated by statistical errors, the errors are smaller in this approach than by analyzing, for example, some appropriately-weighted average of the spectra themselves.

4.2 Splitting SFR by Stellar Mass

If we accept the Copernican principle, which we believe we have evidence to support, how can we reconcile the intermediate redshift range variation between our fossil analysis and the traditional high-z instantaneous SFR indicators? An obvious difference between those studies and ours is the sample of galaxies which are under study. In the high-z regime, only galaxies which are very bright will be observable - the smaller galaxies are too dim to be observed at such a great distance. If there is no variation in SFR with mass it is simple to calculate the cosmic SFR from these large galaxies by integrating along the mass function, which in turn can be estimated from dark matter studies. To investigate the validity of the assumption that the SFH is the same over differently massed systems, we split our SFR into bins of observed stellar mass. Immediately it is obvious that the assumption is invalid - figure 4.2 shows that the redshift at which star-formation activity peaks is an essentially monotonically increasing function of final stellar mass. Star-formation activity in the galaxies in the lowest mass bins ($M_* \simeq 10^{10} M_\odot$) peaked at $z \simeq 0.2$, while for galaxies an order of magnitude more massive the peak lies at $z \simeq 0.5$. At still higher masses, galaxies with masses comparable to a present-day L_* galaxy appear to have experienced a peak in activity at $z \simeq 0.8$, while the highest-mass systems ($M_* > 10^{12} M_\odot$) show a monotonic decline in SFR in our data, with any peak constrained to lie at $z > 2$.

This clearly accounts for the difference between the SFR estimates of high z surveys and our own - in the high z case, only the largest galaxies are observed, but the SFR in these massive galaxies does not track the SFR in smaller galaxies. Figure 4.3 shows the relative contributions of the different massed systems with no offset. This provides a natural explanation for why the most massive star-forming systems, such as the luminous sub-mm selected galaxies, should be largely found to lie at high-redshift ($z > 2$; Chapman et al. (2003)) while at the same time providing further evidence that the bright sub-mm galaxies are indeed the progenitors of today's massive ellipticals (Lowenthal, 2001). The importance of low-mass systems in low-redshift star formation has been noted by Pérez-González et al. (2003) and Fujita et al. (2003), but only with this large sample and with MOPED do we obtain a complete picture.

Indeed, the strong mass-dependence of the star formation history provides a natural explanation of the high redshift of peak star formation activity seen in other surveys, since they are sensitive to the most massive objects only. The fact that we have now discovered that global star-formation activity in fact peaks at rather modest redshift is due to the fact that the peak

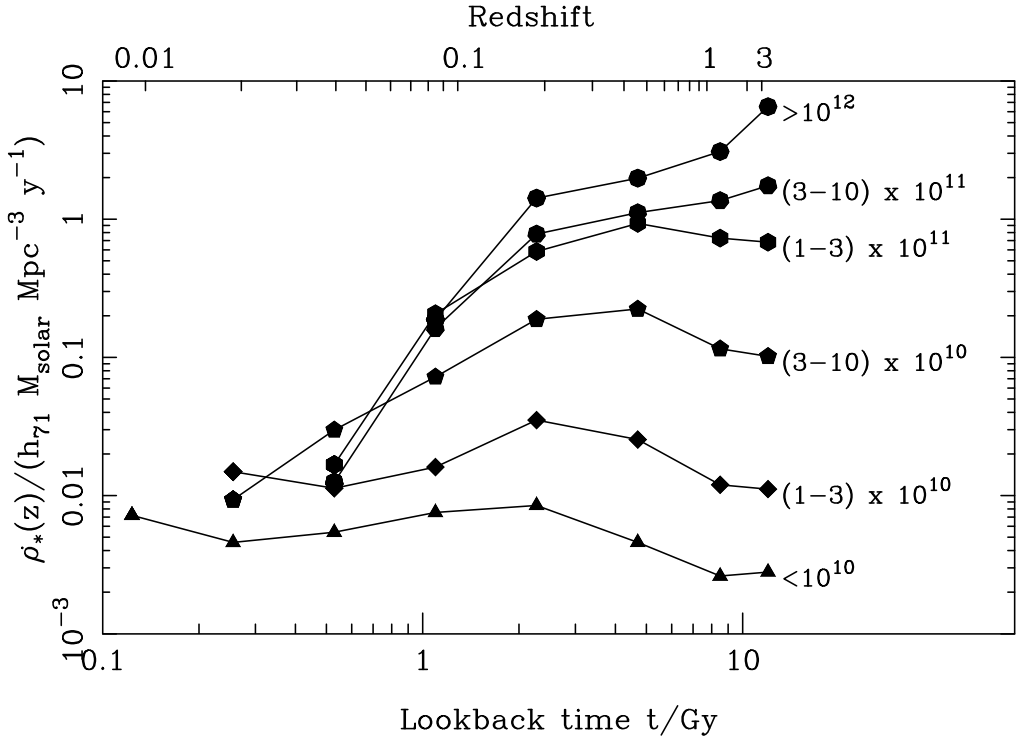


Figure 4.2: The star formation rate as a function of the observed stellar mass of the galaxy. Labels are in Solar Masses. For clarity, the curves are offset vertically successively by 0.5 in the log, except for the most massive galaxies, which are offset by an additional 1.0. Note the clear trend for galaxies with larger present-day stellar mass to have formed their stars earlier. The bulk of the star formation rate at $z \simeq 0.5$ comes from galaxies with present-day stellar masses in the range $3 - 30 \times 10^{10} M_{\odot}$. Note that the graph makes no statement about when the masses were aggregated.

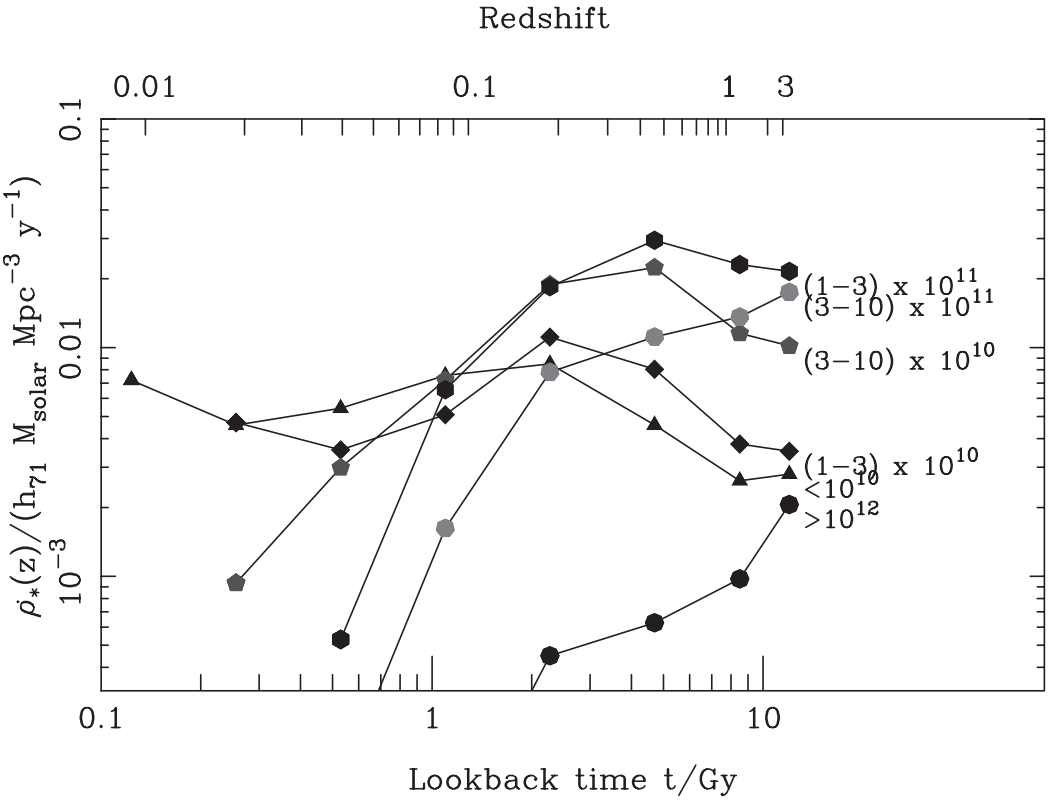


Figure 4.3: The star formation rate as a function of the observed stellar mass of the galaxy, with no offset. Labels give the observed stellar mass in units of $M_\odot$. Plotting the star formation rate in this way allows the dominant star forming objects at each epoch to be seen. It is clear that until a redshift of ~ 0.2 the majority of star formation occurred in galaxies with present day stellar mass $(1-3) \times 10^{11} M_\odot$, while at the present day galaxies $\sim 10\times$ smaller dominate.

epoch of star-formation activity in objects of lower (present-day stellar) masses ($< 3 \times 10^{11} M_{\odot}$) was at $z \lesssim 0.5$, and that such lower-mass galaxies make a significant contribution to the overall star-formation density. Given present-day observational capabilities, this result could only be revealed by a method such as used here, due to the fact that the fossil record approach allows as to explore the star-formation history of galaxies spanning over two decades in mass.

The mass dependence of peak star-formation epoch revealed in figure 4.2 appears to mirror the mass dependence of black-hole activity as recently seen in redshift surveys of both radio-selected (Waddington et al., 2001) and X-ray selected (Hasinger, 2003) active galactic nuclei. Such apparently anti-hierarchical behaviour ("downsizing") is in fact quite consistent with the standard cosmological model, in which galaxies form in small units and merge - our method makes no statement about whether the stellar mass at high redshift was in smaller units or not. The behaviour we see is based on the *present-day* stellar mass of the galaxies, and generally we would expect more massive systems to be part of large-scale over densities, whose first star formation would occur earlier. Furthermore, these results suggest a very different formation history of low- and high-mass systems, explored later.

4.3 Splitting SFR by Concentration Index

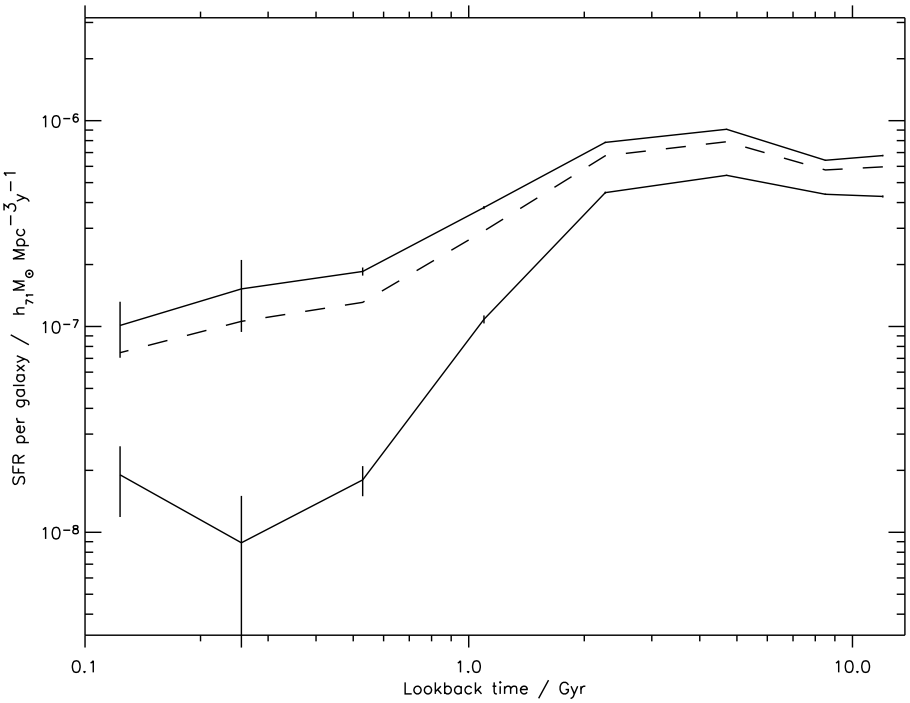


Figure 4.4: Splitting the SFR by concentration index. The upper solid line gives the average SFR per galaxy of galaxies with $c < 2.86$, classified as late type. The lower solid lines gives that of galaxies with $c > 2.86$, classified as early type. The dashed line shows the average star SFR per galaxy for the entire sample.

The Sloan catalogue contains much more than just spectra. By combining measured photometric properties of galaxies with our results we are able to probe star formation as a function of

a property or combination of properties. As an example, there are a number of indicators which may assist in morphological determinations. The one we choose to use is the concentration index, c, defined as

$$c = \frac{R_{90}}{R_{50}} \tag{4.1}$$

where R_{90} and R_{50} are the radii which contain 90 and 50% of the Petrosian flux respectively. It has been shown by Shimasaku et al. (2001) and Strateva et al. (2001) that in bright galaxies the concentration index is closely linked to Hubble type, as classified by eye. The boundary value between early and late type galaxies is poorly defined - Strateva et al. (2001) claim that the break is at 2.6, while Nakamura et al. (2003) suggest 2.86. We have followed the choice made in Shen et al. (2003) and take the divide at $c = 2.86$. Nakamura et al. (2003) report that the completeness of the separation for a subsample of 1500 Sloan galaxies classified with this break is 0.82 and Shen et al. (2003) find it compares favourably to a further indicator of type, the Sersic profile index (Blanton et al., 2003b). Figure 4.4 shows the SFH (normalized to one galaxy) for early and late type galaxies, compared to the average. Although, on average, the early type galaxies have the majority of their star formation at an early age and could be reasonably well modelled by a single burst, there is still some recent star formation. On the other hand, in late type galaxies there is still star formation occurring at the present day, and a single burst does not adequately model the SFH of that galaxy.

4.4 Mass Function of the Stellar Component of Galaxies

Using the stellar mass information from the MOPED analysis of the DR1 data it is possible to calculate the stellar mass function between 3.08×10^7 and $1.05 \times 10^{12} h^{-2} M_\odot$, as shown in figure 4.5. There are 25 galaxies in the sample outside this range, but with 1-4 galaxies per bin the errors are large and these have been omitted. Between about 10^9 and $10^{11} h^{-2} M_\odot$ we find excellent agreement with results obtained by previous studies of SDSS and 2dFGRS galaxies (Bell et al., 2003; Cole et al., 2001), where the stellar mass is estimated more simply from infrared data. We are able to extend the mass range considerably, by around a decade in mass at the upper mass end, and about two decades at the lower-mass end. The stellar mass function of SDSS galaxies is now accurately determined between $10^{7.5}$ and 10^{12} $h^{-2} M_\odot$, where h is the Hubble parameter in units of 100 km s^{-1} Mpc^{-1}.

We fit the galaxy stellar mass function with a Schechter function (Schechter, 1976)

$$\phi(M_s) dM_s = \phi^* \left(\frac{M_s}{M^*}\right)^{\alpha} \exp\left(\frac{-M_s}{M^*}\right) dM_s \tag{4.2}$$

with best-fitting parameters $\phi^* = (7.8 \pm 0.1) \times 10^{-3} h^3 Mpc^{-3}$, $\alpha = -1.159 \pm 0.008$ and $M^* = (7.64 \pm 0.09) \times 10^{10} h^{-2} M_\odot$. This fit is shown overplotted in Fig.4.6, and is a good fit up to $M_s = 10^{11.5} h^{-2} M_\odot$.

There is evidence for an excess over the Schechter fit at the high-mass end (which seems to be confirmed from dynamical measurements of the mass of SDDS galaxies, Bernardi et al. private communication), which can be well modelled by the addition of a power law over the range

$11.5 < \log_{10}(h^2 M_s/M_\odot) < 12.6$ giving a fitting equation of the form

$$\phi_c = \phi + F_C \left(\frac{M_s}{M^*}\right)^\beta \tag{4.3}$$

Setting $\beta = -3.2$ gives $F_C = (7.1 \pm 0.5) \times 10^{-4} h^3 \mathrm{Mpc}^{-3}$ as shown in Figure 4.6. This excess could be due to cD galaxies or a failure of the modelled correction to total magnitudes for these extremely large galaxies due to their large angular size. The magnitudes, modelled spectra and redshift distribution of these high mass galaxies were examined, and nothing unusual was found which could skew the analysis.

4.5 Evolution of Mass Function with Redshift

By splitting the DR1 sample by redshift we can probe the evolution of the stellar mass function in recent times. Fig. 4.7 shows the stellar mass function for galaxies within narrow redshift ranges. Because of the flux limit, there is essentially a minimum mass which can be probed at each redshift, but this is not a sharp cutoff because the galaxies have a range of star formation histories so the mapping from stellar mass to luminosity is not one-to-one. It is apparent from the figure that within the limits of the survey there is very little, if any, evolution in the redshift range $z < 0.34$. The only notable deviation from this is an apparent deficiency of high-mass galaxies ($M_s > 10^{11} h^{-2} M_\odot$) at very low $z < 0.05$. The high mass results from the lowest redshift sample should be treated with caution. The galaxies at the high mass end of the mass function are generally large in their angular size. This leads to a problem with "shredding" by the SDSS photometric pipeline, where large galaxies are treated as many smaller sources. It is thought that this is only really a problem for $z < 0.02$, but for $z < 0.01$ as many as 10% of the detections could be affected (SDSS Collaboration, private communication).

The lack of evolution of the mass function with redshift is in contrast to the significant evolution found in the luminosity function, where the characteristic luminosity has become fainter by around 0.3 magnitudes since $z = 0.2$, and the number density of bright galaxies has declined by a factor of two or more (Loveday, 2004; Blanton et al., 2003c). The most natural explanation is that the stellar mass content has hardly changed, but that the galaxies have just become significantly fainter; this is expected given the drop-off in star formation activity, and can be illustrated by Fig. 4.8, which shows the evolution of the average stellar mass with redshift, for galaxy populations of different luminosities. We see clearly an increase in the average mass with decreasing redshift.

4.6 Cosmological Baryon Density in Stars

The stellar mass function can be used to give a further constraint on the contribution to the density parameter from baryons in stars, $\Omega_{\mathrm{b}*}$. By integrating the mass over the range of the mass function we deduce a value of $\Omega_{\mathrm{b}*} h = 2.39 \pm 0.08 \times 10^{-3}$. This value is in broad agreement with results obtained previously (Cole et al., 2001; Bell et al., 2003; Fukugita et al., 1998; Kochanek

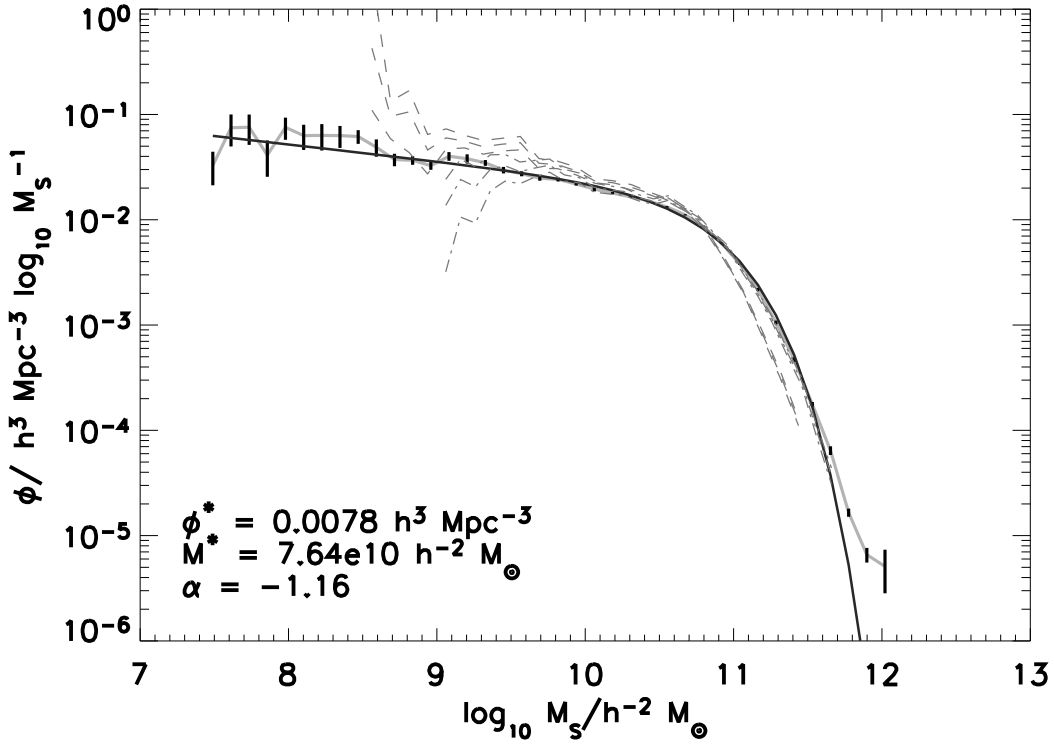


Figure 4.5: Stellar mass function for the SDSS DR1, with a Schechter Function fit overplotted (solid black line). 18 galaxies with $M_S < 3.08 \times 10^7 h^{-2} M_\odot$ and 7 galaxies with $M_S > 1.05 \times 10^{12} h^{-2} M_\odot$ have been omitted as there are fewer than 9 galaxies per bin. Also shown are the mass functions and errors recovered by Cole et al. (2001) (dashed grey line) and Bell et al. (2003) (dashed-dotted grey line). The mass offset of the Bell data is due to a different choice of IMF.

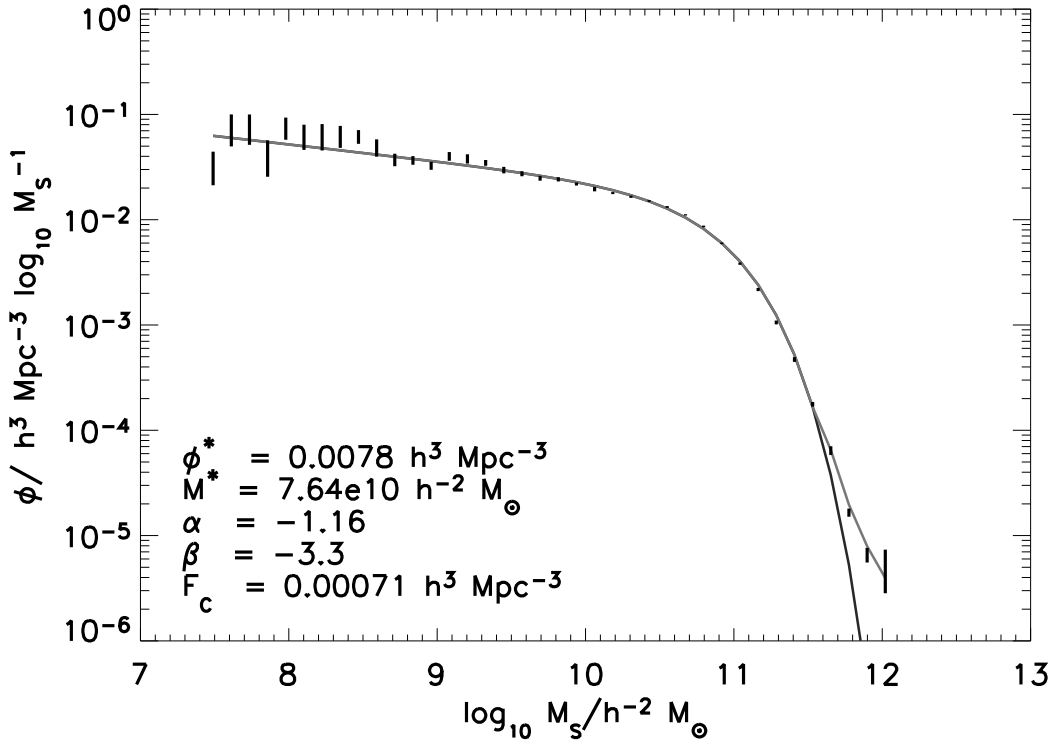


Figure 4.6: Mass Function with Modified Fit. Here we present a well-fitting modification to the Schechter function (grey line) which gives excellent agreement with our results at high mass. Again, the first four bins have been excluded from the fit. The Schechter function alone is a good fit up to $10^{11.5}h^{-2}M_\odot$

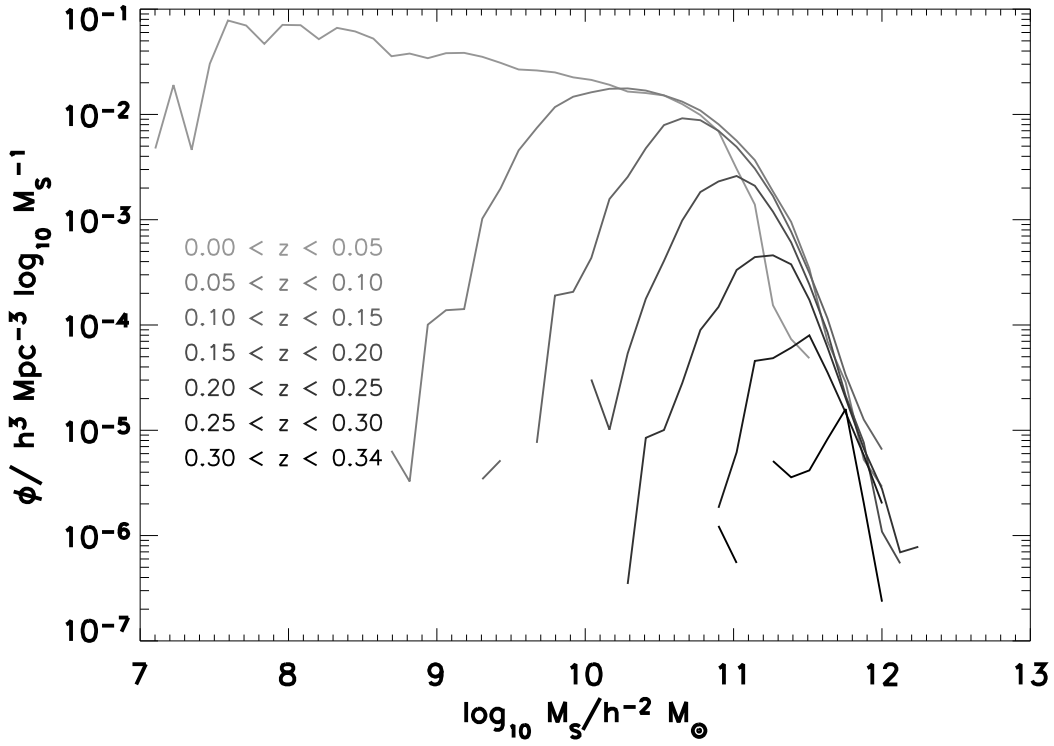


Figure 4.7: Mass functions for different redshift ranges. The agreement is generally very good where the samples overlap, indicating that there is little evolution over the redshift range $0 - 0.3$. There is some discrepancy at the high-mass end in the lowest redshift range, thought to be due to Sloan photometric pipeline shredding large galaxy images.

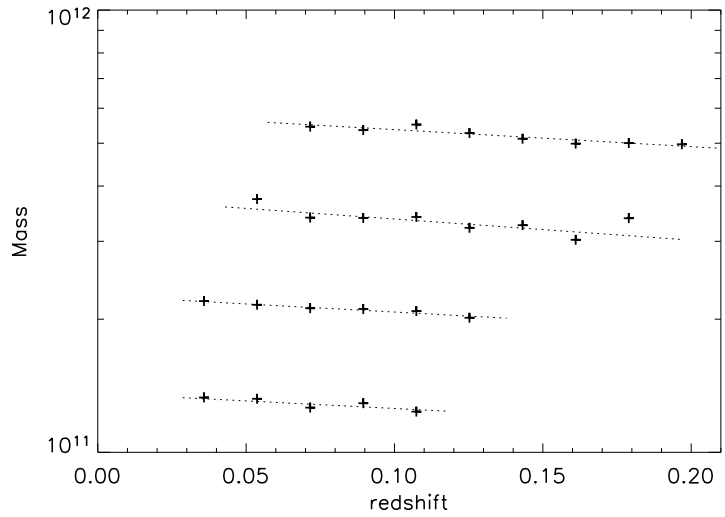


Figure 4.8: The evolution of the average galaxy stellar mass for galaxies with (from the bottom) $\log_{10}(L_R/L_\odot) = 10.25 - 10.3, 10.45 - 10.5, 10.65 - 10.7, 10.85 - 10.9$, where L_R is the $R-$band luminosity, $K-$corrected using kcorrect v3_1B (Blanton et al., 2003a)

et al., 2001; Glazebrook et al., 2003; Persic & Salucci, 1992; Salucci & Persic, 1999). Our error is a jackknife estimate, and is purely statistical; systematic errors such as the choice of IMF have not been included.

4.7 Baryonic Conversion Tree

The results in this section are the results of collaborative work with Raul Jimenez, University of Pennsylvania, Philadelphia. My contribution to the results is the data for the luminous matter side of the comparison; he produced the dark matter halo merger information. The methods he used to build the halo merger trees follow Wechsler et al. (2002), and are published in Jimenez et al. (2004b).

The star formation histories of individual galaxies can be integrated and averaged to give an estimate for the buildup of baryons in galaxies with certain masses. By comparing this information with the buildup of dark matter, it is possible to gain some understanding of the feedback processes which affect star formation. Using 10 time bins, spaced equally in the log, we find that more massive galaxies transform their gas into stars *earlier* than less massive ones. The top panel of figure 4.9 shows for each time bin how many baryons (as a fraction of the observed total stellar mass) were converted into stars, f_s, as a function of observed stellar mass $M_* = M_*(t = 0)$. The different lines correspond to different time bins: dash oldest, dots second-oldest and continuous lines denote younger bins. The bottom panel shows the stellar mass assembly history $\delta M_*(t_i)$ for different observed stellar masses.

There is a clear correlation between baryon conversion efficiency and present-day stellar mass. This can be see from figure 4.9. In galaxies with M_* larger than 3×10^{11} $M_\odot$, more than 60% of their present stellar mass was already in place at redshift 1.7. In terms of look-back time it

means that the stars of massive galaxies ($M_* \sim 10^{12}$ M$_\odot$) were essentially formed (if not in place) more than 9 Gyr ago, or just ~ 4 Gyr after the big bang. Conversely, for $M_* < 3 \times 10^9$ M$_\odot$, 80% of their stellar mass remains unformed at $z = 0.2$ or ~ 3 Gyr ago.

4.7.1 Dark Matter Assembly History

To compute the mass of dark matter halos as a function of time we use two approaches: first, we generate multiple realizations (10^4) of the merger history of dark matter halos of several masses in the range 10^9 to 10^{13} M$_\odot$. This is done using the algorithm described in Somerville & Kolatt (1999) for the standard LCDM cosmology ($\Omega_0 = 0.27$, $\Lambda_0 = 0.73$, $h = 0.71$, Spergel et al. (2003)). This algorithm reproduces the merger histories of halos seen in N-body simulations of structure formation (Somerville et al., 2000; Wechsler et al., 2002), especially at low redshift, which is the range of interest. One free parameter in the algorithm is the value of the mass that is considered accreted instead of merged. Agreement with CDM simulations is achieved when everything below 1% of the final halo mass is considered accreted and this is the value we use. Second, we use the fitting formula for the mass accretion history from Wechsler et al. (2002). This is obtained from numerical N-body simulations performed with the ART code (Kravtsov et al., 1997). We find that both approaches show the same qualitative behaviour illustrated in Fig. 4.10, however since the Extended Press-Schechter approach, at the base of the Somerville & Kolatt (1999) algorithm, is not a perfect fit to N-body simulations, especially at high-redshift and for very massive halos, we will present here only results obtained using the second approach.

We compute the dark matter mass of the most massive progenitor that is virialised in each redshift bin as a function of the total mass of the dark halo. The top panel of Fig. 4.10 shows the fraction of the final dark matter mass that has been virialised in the most massive progenitor, f_D, in a given redshift bin (line styles same as in Fig. 4.9) as a function of the final stellar mass in the dark halo. The stellar mass of the dark halo is obtained from the dark one assuming the universal baryon fraction f_b as determined by WMAP ($\Omega_{\rm CDM}/\Omega_b = 4.8 \equiv f_{\rm DM/b}$; Spergel et al. (2003)), and that at $z \sim 0$ only about 6% of the baryons are in stars (Fukugita, 2004).

The bottom panel of Fig. 4.10 shows the mass assembly history of the dark halo.

A comparison of Fig. 4.9 and Fig. 4.10 indicates that dark matter and baryons do not follow the same assembly process. For example, for $M_* > 10^{12}$ M$_\odot$ less than 50% of the dark matter is assembled in the main progenitor at $z > 1.7$ ($t_{\rm lookback} \sim 9.7$ Gyr), while more than 75% of the stellar mass is already formed. On the other hand, for stellar masses smaller than 10^{11} M$_\odot$, 20% of the stellar mass is formed in the same time bin while already 60% of the dark matter is in place. This hints at a role of early stellar feedback in these halos as we discuss later in §4.

While in the hierarchical LCDM model for structure formation the more massive CDM structures form late, we find observational evidence for early star formation of giant galaxies. This can happen because of two reasons: *a)* massive galaxies are formed by mergers with smaller ones, each carrying an evolved stellar population *b)* these massive galaxies are already in place at high-redshift and the dark matter halo has been assembled at the same time as the stellar population. If the dark matter halo collapse triggers star formation then one would expect *a)* to be the case; however observations of e.g., old elliptical galaxies at high redshift (e.g. Dunlop et al., 1996; Nolan et al., 2003; Daddi et al., 2004; Saracco et al., 2004) seem to support the second scenario, at least in some cases.

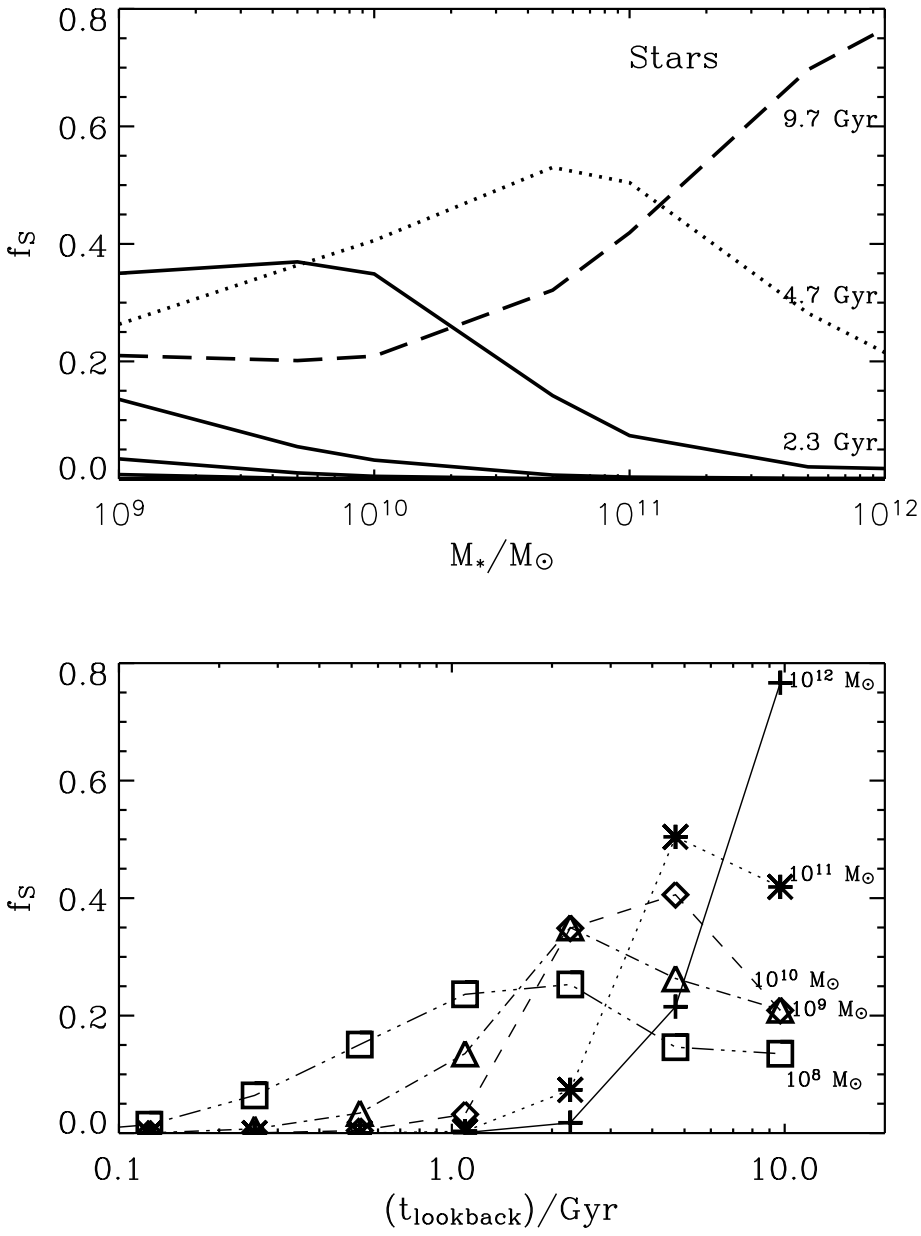


Figure 4.9: Top panel: fraction of stellar mass of the galaxy as a function of total stellar mass for different time bins. Dashed line corresponds to the oldest bin (9.7 Gyr), dotted line to the second oldest (4.7 Gyr) and solid line to the third oldest (2.3 Gyr). The other solid lines correspond to younger time bins (1.1, 0.53, 0.26). Bottom panel: fraction of the observed total stellar mass of the galaxy created as a function of time for different galaxy stellar masses (determined at the observed redshift).

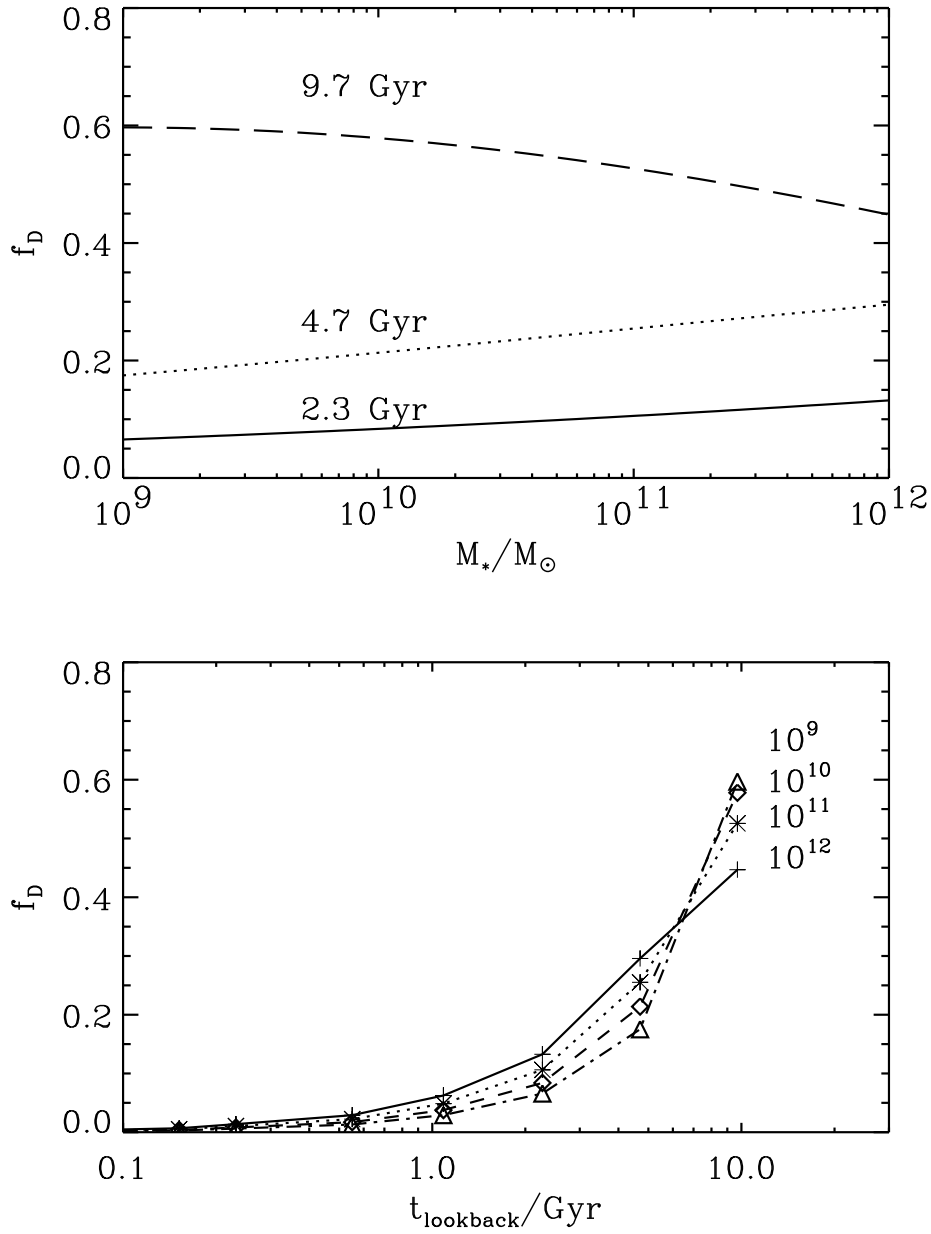


Figure 4.10: Top panel: fraction of the final dark matter mass that has been virialized in the most massive progenitor, as a function of the total stellar mass of the galaxy (at the observed redshift). The different lines correspond to the amount of dark matter that is virialized in the given time bin. Bottom panel: fraction of the final dark matter mass that has been virialized in the most massive progenitor as a function of $t_{\rm lookback}$. The different lines correspond to different total stellar masses of the galaxy at the observed redshift.

4.7.2 The Number of Progenitors of Galaxies

It is clear that since massive dark halos are predicted, *on average*, to assemble later than the stellar population they contain, the stars may naturally have formed previously in smaller dark halos that subsequently merged. By considering the (dark) mass accretion history of the most massive progenitor of a halo, and the star-formation rate, we can constrain the minimum number of progenitors forming a halo, and the star formation efficiency as outlined below. In other words, if the most massive progenitor at a given redshift z carries enough baryons to form all the stars that should be in place by z then only one progenitor (the most massive one) is needed. If the minimum number of progenitors forming a halo is larger than 1, massive galaxies must have form by mergers with smaller ones, each carrying an evolved stellar population. However, if the minimum number of progenitors is 1, then the LCDM paradigm can accommodate old elliptical galaxies at high redshift.

For example, consider a galaxy with stellar mass $M_* = 10^{12}$ $M_\odot$ (see Fig. 4.9 and Fig. 4.10). In the oldest time bin 76% of the stellar mass is already in place ($M_*(t_{10}) = 0.76 M_*$), yet the virialised fraction of the dark matter halo is less than 50%. If we assume case a) (that is star formation happens uniformly in all halos and sub-halos that ultimately end up forming the final galaxy, and that 100% of the dark matter is virialised at all times) then the fraction of the total baryonic mass converted into stars is 4.6%. On the other hand, observations of old elliptical galaxies at high redshift can be explained within the LCDM paradigm if we assume that the main virialised progenitor harboured all the stars observed, in this case the fraction of mass converted to stars in the dark matter halo must have been $\sim 10\%$. This requires an enhancement in star formation efficiency at high redshift, as the typical efficiency observed in giant molecular clouds today is $< 10\%$ (e.g. Padoan & Nordlund, 2002).

More specifically, we can assume that the stellar mass $M_*(t_{10})$, was in progenitors (the most massive progenitor and possibly other sub halos), whose cumulative dark matter must have been at least $M_*(t_{10}) \times f_{\rm DM/b}/f$, where $f \leq 1$ parameterises the star formation efficiency, and the fraction of baryons that gets turned into stars, and depends on the mass of the sub-halos. We approximate $f(M_{\rm dark})$ as $f_s(M_{\rm dark}/f_{\rm DM/b} \times 0.1)$ (as 0.1 is the minimum efficiency in the oldest time bin).

The minimal number of progenitors can thus be obtained by minimisation as a function of the sub halo dark mass M_{SH}:

$$N(M_{\rm SH}) = 1 + \max\left\{\frac{f_{\rm DM/b} M_*(t_{10})/f(M_{\rm SH}) - M_{\rm dark}(t_{10})}{M_{\rm SH}}, 0\right\} \tag{4.4}$$

We obtain a minimum number of 1 and $f \sim 10\%$, in agreement with the more heuristic argument above. If the main progenitor did not have enough baryons to accommodate $M_*(t_{10})$, then the minimum number of progenitors would be greater than one. If the minimum number of progenitors is one and the star formation efficiency is constrained to be reasonable, then all the old stellar population could have been formed in the main progenitor.

Thus this suggests that in the LCDM paradigm, the massive old galaxies that we see today could have been made of mergers of few progenitors, each carrying an old stellar population, in agreement with other indications that elliptical galaxies are already formed at $z > 1$ (e.g. Bower et al., 1992; Peacock et al., 1998; Lilly et al., 1998; Brinchmann & Ellis, 2000; Im et al., 2002; Mulchaey et al., 2004; Gao et al., 2003; Saracco et al., 2004; Glazebrook et al., 2004). A small

number of mergers can also naturally explain the tightness of the observed colour-magnitude relation (Bower et al., 1992).

4.7.3 Time Evolution of Star Formation Efficiency

For each of the time bins we compute the ratio of the newly-formed stellar mass to the baryonic mass added to the main progenitor, assuming the nucleosynthesis baryon fraction. Fig. 4.11 shows the above ratio as a function of look-back time ($t_{\rm lookback}$) for different masses: crosses correspond to a stellar mass of 10^{12} $M_\odot$, asterisks to 10^{11} $M_\odot$, diamonds to 10^{10} $M_\odot$, triangles to 10^{9} $M_\odot$ and squares to 10^{8} $M_\odot$. This is a measurement of how much gas is transformed into stars as a function of the newly-added baryons. A value of 1 indicates that the mass of baryons accreted to the main progenitor matches the mass converted into stars. A value higher than 1 shows that the accretion or merger was accompanied by a greater mass of triggered star formation somewhere in the galaxy. Fig. 4.11 clearly shows that for stellar masses above 10^{10} $M_\odot$ this ratio is never greater than 1. Clearly we are comparing mass accreted on to the main progenitor with stars created in any of the progenitors, and this should be borne in mind in interpreting Fig. 4.11. However, since the main progenitor contains $\sim 50\%$ of the final mass even at a lookback time of 10 Gyr, and this fraction is weakly dependent on mass, the efficiency of conversion of baryons to stars in the galaxy as a whole is unlikely to alter the main conclusions of Fig. 4.11, where the differences between objects of different present-day masses typically far exceed a factor of 2.

For the most massive galaxies at early times this measure of star formation efficiency is close to 40%. For stellar masses below $10^{10}M_\odot$, the efficiency is only of about 6-8% at $t_{\rm lookback} = 10$ Gyr, but grows to 100% at $t_{\rm lookback} = 2$ Gyr, and then decreases.

For galaxies with stellar mass smaller than 10^{9} $M_\odot$, this increase in star formation efficiency rises until $t_{\rm lookback} \sim 0.5$ Gyr, at which point it reaches 300% efficiency, which means that more gas is transformed into stars than the baryons brought into the parent dark halo by accretion. This points toward a picture where these low-mass gas-rich galaxies see a lot of their gas reservoir transformed into stars due, for example, to a merger or accretion event. Another possibility is that star formation is proceeding rapidly in other sub-halos, which subsequently merge with the parent. This scenario is not strongly supported by the merger histories of low-mass halos.

Thus massive galaxies have a high star-formation efficiency at early times and then evolve "passively", with fresh infall of gas being suppressed or turned into stars with low efficiency, possibly because it is likely to be too hot. Small galaxies seem to accrete mass passively at early times and form stars very efficiently later.

Conversely, the probability distribution for dark halo merger events peaks at higher redshift for small halos and at lower redshift for large halos. In a Λ-dominated Universe, merger probability is suppressed at $z < 1$ especially in low density regions, where small galaxies are most likely to be. Thus there seems to be no correlation between halo virialization or dark matter merger events and star formation efficiency.

However, one could imagine explaining this trend, for example, by postulating the existence of a "threshold" for star formation: once this threshold is crossed, all available baryons are turned into stars (as in an "infall model"), then afterward galaxies evolve passively (as in a "closed box model"). In this toy model, if this threshold is crossed at very early times in the progenitors of massive galaxies, one would expect these galaxies to form stars very efficiently early on then

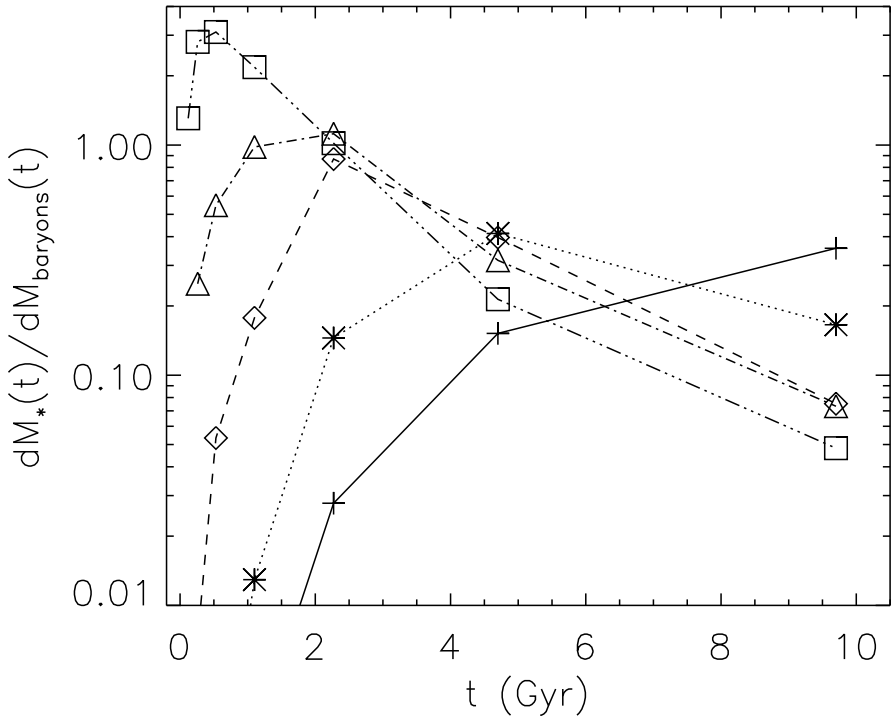


Figure 4.11: Ratio of mass transformed into stars in a galaxy to the baryons that are accreted onto the main progenitor, which typically contains $>$ half the mass. Crosses correspond to a stellar mass of 10^{12}, asterisks to 10^{11}, diamonds to 10^{10}, triangles to 10^{9} and squares to 10^{8} $M_{\odot}$. A value for $dM_*(t)/dM_{\rm baryons}(t)$ smaller than 1 means that not all the baryons (the nucleosynthesis value as recently constrained by WMAP) have been transformed into stars, while a value larger than 1 indicates that gas previously available in the galaxy has been turned into stars: either the recent accretion has triggered star formation in the main progenitor, or it is going on elsewhere. Note that for stellar masses below 10^{10} $M_{\odot}$, the star formation efficiency peaks at $t_{\rm lookback} \sim 1-2$ Gyr.

evolve passively. For progressively smaller galaxies this threshold is met at increasingly later times.

An alternative explanation is that stellar feedback is responsible for the lack of star formation in small galaxies at early times. Since the escape velocity in these systems is smaller than in more massive galaxies, gas can leave the dark matter halo more easily. Only the very massive systems are able to retain their gas and convert a majority of their gas into stars. If so we find that $M_* \sim 10^{10}$ M$_\odot$ is the characteristic mass that defines the border between efficient and inefficient feedback.

4.7.4 Summary

We have determined for the first time the baryonic conversion tree for galaxies. We were able to do this with observations at $z < 0.3$, using 96545 galaxies from the SDSS DR1 spectroscopic sample.

In the hierarchical structure formation model, massive *dark halos* form (i.e. virialize) later than smaller halos, from mergers of smaller units (e.g. Lacey & Cole, 1993, 1994; Lin et al., 2003). This model has been thoroughly tested against numerical cold dark matter simulations. Naively one could expect that the dark matter halo collapse should trigger baryonic gas transformation into stars; in addition subsequent dark matter mergers should produce star formation episodes.

Instead, for the the stellar assembly history we find that: the more massive galaxies have old stellar population and massive, old elliptical galaxies are already in place at $z \sim 1$. This has been known for a long time and sometimes it is referred to as "down-sizing", (Cowie et al., 1996). We find that massive galaxies have transformed more gas into stars at higher redshift (in agreement with high z observations e.g., Kodama et al. (2004)), and then star formation was suppressed, while less massive galaxies transform more gas into stars at low redshift.

So one should not be surprised to see abundant red objects at high redshift in the LCDM paradigm, these objects can form in virialised halos if star formation efficiency is high. Indeed, Jimenez et al. (1999b) have shown that single-halo hydro-dynamical models would require an increased star formation efficiency for more massive galaxies, higher than the few percent found today in giant molecular clouds, in agreement with the value determined from the "fossil record" in the present work. Our findings, based only on observations at $z < 0.35$ (the "fossil record"), are in agreement with a suite of independent, $z > 1$, observations: observations of old elliptical galaxies at high redshift (Dunlop et al., 1996; Spinrad et al., 1997; Nolan et al., 2003), indications that elliptical galaxies are already formed at $z > 1$ (e.g. Bower et al., 1992; Peacock et al., 1998; Lilly et al., 1998; Brinchmann & Ellis, 2000; Im et al., 2002; Mulchaey et al., 2004; Gao et al., 2003; Glazebrook et al., 2004). It also nicely explains the tightness of the colour-magnitude relation (Bower et al., 1992).

On the other hand, we find that small galaxies seem to accrete mass passively at early times and see a lot of their gas reservoir transformed into stars at late times. Since the probability distribution for dark halo mergers peaks at low redshift for massive halos and high redshift for small halos, we conclude that dark matter mergers and star formation are not correlated. We speculate that one possibility to explain the anti-hierarchical nature of the stellar assembly history is the existence of a threshold for star formation: once the threshold is crossed all available baryons are turned into stars ("infall model", see Binney & Merrifield, 1998) and afterward galaxies approximatively evolve passively. The threshold is met at very early time for massive galaxies and a

later time for less massive ones (see e.g. Heavens & Jimenez (1999)).

A star formation threshold has been observed in disk galaxies by Martin & Kennicutt (2001) and there has been some recent additional evidence from the formation of dust lanes in disk galaxies (Dalcanton et al., 2004) that this threshold may take place at $V_c \sim 100$ km s^{-1}, in agreement with the findings of Verde et al. (2002) and Kannappan (2004) who also found a transition at about 100 km s^{-1} for star formation efficiency. This V_c value corresponds to the characteristic mass found here ($M_* \sim 10^{10}$ M$_\odot$), that defines the border between efficient and inefficient star formation. This characteristic mass has been related to feedback efficiency threshold (e.g. Dekel & Silk, 1986; Dekel & Woo, 2003, and references therein).

As we do not yet have a fundamental theory for galaxy formation and given the complexity involved in studying the process with hydrodynamic-N body simulations, we hope that this new determination of the baryonic conversion history will be a useful observable to gauge galaxy formation models against.

CHAPTER 5

Conclusions

5.1 Summary of Results

The main aim of this thesis has been to calculate the star formation history of galaxies from the fossil record. By determining the fractions of different stellar populations which contribute to the galaxy spectrum using MOPED, a data compression algorithm, and ROAM, a parallel computing environment, I have analyzed the ~ 150000 galaxies in the SDSS DR1 dataset. Choosing a suitable subsample of this set I have been able combine the star formation histories (SFH) with other observed SDSS parameters to determine the stellar mass component of each galaxy, which allows both the cosmic star formation rate and the galaxy stellar mass function to be calculated. The breadth of the survey allows average SFH to be calculated for different mass ranges and morphological types of galaxies to give the build up of stellar material - the Baryonic Conversion Tree (BCT). Comparing this buildup to that of the dark matter, modelled by halo mergers, gives some insight into the poorly understood feedback processes which control star formation in galaxies.

This thesis contains several results which contribute to our understanding of the formation of galaxies. I have shown that:

- Using MOPED, SPEED and a Markov Chain technique it is possible to explore, and break, the age metallicity degeneracy in galaxy spectra for the population as a whole.
- The overall shape of the cosmic SFR is in good agreement with previous estimates - the SFR has peaked and is now in decline across the Universe.
- The peak in SFR is much broader than previously thought, and possibly only peaks at $z \sim 0.6$. This is in disagreement with previous studies, probably because instantaneous SF indicators at high z probe only high mass galaxies.
- The SFH is not uniform across different galaxy types or masses - the peak increases in redshift with observed galaxy mass. The redshift of the peak in star formation ranges from $z > 2$ for the highest mass galaxies ($M_S \gtrsim 10^{12} M_\odot$) to $z \sim 0.2$ for the lowest ($M_S \lesssim 10^{10} M_\odot$).Since high redshift surveys study only the most massive systems (Malmquist Bias), they will misinterpret the overall star formation rate.
- The variation in SFR shape with galaxy mass is a convincing observation of "downsizing". Such apparently anti-hierarchical behaviour is in fact quite consistent with the standard cosmological model, in which galaxies form in small units and merge. The behaviour we see is based on the *present-day* stellar mass of the galaxies, and generally we would expect more massive systems to be part of large-scale overdensities, whose first star formation would occur earlier.
- The recent SFH of early type galaxies ($c < 2.86$) is different to that of late type ($c > 2.86$). The ellipticals have most of their star formation initially, where the spirals still have starformation at the current day.
- The agreement between the observed star formation rate from high redshift surveys and this work gives support to the Copernican Principle, that we are not privileged observers.
- I have extended the range of the mass function of galaxies' stellar mass by 2 decades of mass. The mass function is very well fitted by a Schechter function over 4 decades of mass,

in agreement with previous more limited studies. In particular the area around the break is fitted in exquisite detail, although the scaling of the mass axis probably depends on choice of IMF. The Schechter fit parameters are $\phi^* = (7.8 \pm 0.1) \times 10^{-3} h^3 \mathrm{Mpc}^{-3}$, $M^* = (7.64 \pm 0.09) \times 10^{10} h^{-2} M_\odot$ and $\alpha = -1.159 \pm 0.008$.

- A slight modification of the Schechter function at high mass provides an excellent fit over 5 decades of mass, although with the current sample there are insufficient galaxies to properly constrain the parameters of the modification. The galaxies that contribute to the excess could be cannibalizing cD monster galaxies in clusters.
- The contribution to the density parameter Ω from baryons in stars, $\Omega_{\mathrm{b}^*} h$, is $2.39 \pm 0.08 \times 10^{-3}$. This value is in broad agreement with previous estimates.
- There is no noticeable evolution of the mass function in the redshift range $0.05 < z < 0.34$. This indicates that almost all stars were already formed at $z \sim 0.34$, in good agreement with the sharp decline in SFR. This lack of mass evolution suggests that the evolution seen in the luminosity function must be largely due to stellar fading, as expected from models of stellar evolution.
- Comparing the BCT to the dark matter merger tree indicates that star formation efficiency (the measure of how much available gas has actually been processed into stars) at $z > 1$ had to be high (as much as 10%) in galaxies with present-day stellar mass larger than $2 \times 10^{11} M_\odot$, if this early star formation occurred in the main progenitor.
- Conversely, in galaxies with present-day stellar mass less than $10^{11} M_\odot$, efficient star formation seems to have been triggered at $z \sim 0.2$.
- These two observations lead to there being characteristic mass ($M \sim 10^{10} M_\odot$) for feedback efficiency (or lack of star formation). For galaxies with masses lower than this, feedback (or star formation suppression) is very efficient while for higher masses it is not.

5.2 Discussion

The analysis techniques given in this thesis have performed well in all tests applied to them: application of SDSS noise; variation of the continuum slope; H-δ line-filling; testing for low redshift bias. To all extents possible we have tested that there is neither a hidden bias to the results given by individual galaxies or some problem with our $V_{\max}$ method. Even so, there are some factors which are difficult to test - in particular the stellar models are, after all, only models, and there are still lines in the spectra of galaxies which are not included in the models. Although we have some sensitivity to the IMF, we are far less dependent on it than, for example, UV continuum flux methods, where (especially at high redshifts) large corrections are required to extrapolate from high mass stars to low. The different dust models differ very little in their shape in the rest frame range of the SDSS galaxies: although there is some evidence that assuming the same model for all galaxy types could be misleading in some circumstances, there is very little change between recovered SFR between models. In reality, the dust is mixed in with stars and this could be modelled to a greater degree of complexity: but although a massively complex model will tell you more about the dust in the galaxy it is unclear whether it will give more information on

the recovered SF. This issue of paramaterization is a tricky one - too few and the results are over simplified, too many and the dimensionality of the hypersurface is too high to properly interpret. Our current paramaterization assumes that at a given time the gas which makes all stars throughout the galaxy has the same metallicity - in reality there will be a mix of metallicities, but without either a large increase in the number of parameters or making some assumption about the relative fractions of metallicity it is impossible to reflect this condition. It is very reassuring however that totally different methods of recovering the SFR (both at high and low redshift) give good agreement with our results. There is no reason that this should be the case, and the discrepancy at intermediate redshifts can be explained by the mass dependence of SFR.

5.3 Future Work

The database of SFH information developed in this thesis opens up the possibility of investigating the SFR of galaxies in different environments. It is possible to construct the local galaxy density from the SDSS itself or from the complete 2dF redshift survey in overlap regions, and investigate the effect of environment on star formation.

Subdivision of the sample will allow us to investigate star formation rate by spectral type, concentration index, colour - in fact any measured parameter of the Sloan survey. I also plan to compare my database with one which identifies those galaxies with an AGN component. Using the two datasets it should be possible to determine the effect of AGN on the star formation efficiency of galaxies. My overall aim is to improve the understanding of SF processes and provide a set of firm constraints for galaxy formation models.

The MOPED technique has been proven here on the huge SDSS dataset. This survey has a massive size but only extends to a redshift of ~ 0.34. Much deeper surveys such as the Gemini Deep Deep Survey (Abraham et al., 2004) are currently collecting some thousands of high redshift spectra, using these will allow improvements in the time resolution of the SFR at high redshift.

Although applying MOPED to new datasets is possible at present, the MOPED code itself could be further developed. There are now high-resolution (~ 1 Å) stellar models available. Fortunately, there is virtually no processing overhead with MOPED in using higher-resolution models (Bruzual & Charlot, 2003), and some (possibly significant) information gain. More advanced dust models are available which consider dust in a more complex fashion than a simple screen (Charlot & Fall, 2000), and there is continual development on the IMF which is incorporated in our method for the determination of SFH.

Bibliography

Abazajian, K. et al. 2003, AJ, 126, 2081

Abazajian, K. et al. 2004, AJ, 128, 502

Abraham, R. G. et al. 2004, AJ, 127, 2455

Bell, E. F. 2003, ApJ, 586, 794

Bell, E. F., McIntosh, D. H., Katz, N., Weinberg, M. D. 2003, ApJ, 585, L117

Bennett, C. L. et al. 2003, ApJ, 583, 1

Binney, J., Merrifield, M. 1998, Galactic astronomy, Princeton University Press, 1998. (Princeton series in astrophysics) QB857 .B522 1998

Blanton, M. R., Brinkmann, J., Csabai, I., Doi, M., Eisenstein, D., Fukugita, M., Gunn, J. E., Hogg, D. W., Schlegel, D. J. 2003a, AJ, 125, 2348

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

Blanton, M. R. et al. 2003c, ApJ, 592, 819

Boggess, N. W. et al. 1992, ApJ, 397, 420

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

Brinchmann, J., Charlot, S., White, S. D. M., Tremonti, C., Kauffmann, G., Heckman, T., Brinkmann, J. 2004, MNRAS, 351, 1151

Brinchmann, J., Ellis, R. S. 2000, ApJ, 536, L77

Bruzual, A. G. 1983, ApJ, 273, 105

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

Burstein, D., Faber, S. M., Gaskell, C. M., Krumm, N. 1984, ApJ, 287, 586

Calzetti, D. 1997, AJ, 113, 162

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

Chapman, S. C., Blain, A. W., Ivison, R. J., Smail, I. R. 2003, Nature, 422, 695

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

Charlot, S., Kauffmann, G., Longhetti, M., Tresse, L., White, S. D. M., Maddox, S. J., Fall, S. M. 2002, MNRAS, 330, 876

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

Cole, S. et al. 2001, Mon. Not. Roy. Astron. Soc., 326, 255

Colless, M. et al. 2001, MNRAS, 328, 1039

Condon, J. J. 1992, ARA&A, 30, 575

Connolly, A. J., Szalay, A. S., Dickinson, M., Subbarao, M. U., Brunner, R. J. 1997, ApJ, 486, L11+

Cowie, L. L., Songaila, A., Barger, A. J. 1999, AJ, 118, 603

Cowie, L. L., Songaila, A., Hu, E. M., Cohen, J. G. 1996, AJ, 112, 839

Daddi, E. et al. 2004, ApJ, 600, L127

Dalcanton, J. J., Yoachim, P., Bernstein, R. A. 2004, ApJ, 608, 189

de Bernardis, P. et al. 2000, Nature, 404, 955

Dekel, A., Silk, J. 1986, ApJ, 303, 39

Dekel, A., Woo, J. 2003, MNRAS, 344, 1131

Diaz, A. I., Terlevich, E., Terlevich, R. 1989, MNRAS, 239, 325

Dicke, R. H., Peebles, P. J. E., Roll, P. G., Wilkinson, D. T. 1965, ApJ, 142, 414

Dickinson, M., Papovich, C., Ferguson, H. C., Budavári, T. 2003, ApJ, 587, 25

Donas, J., Deharveng, J. M. 1984, A&A, 140, 325

Donas, J., Deharveng, J. M., Laget, M., Milliard, B., Huguenin, D. 1987, A&A, 180, 12

Donas, J., Milliard, B., Laget, M. 1995, A&A, 303, 661

Dressler, A., Shectman, S. A. 1987, AJ, 94, 899

Dunlop, J., Peacock, J., Spinrad, H., Dey, A., Jimenez, R., Stern, D., Windhorst, R. 1996, Nature, 381, 581

Faber, S. M., Friel, E. D., Burstein, D., Gaskell, C. M. 1985, ApJS, 57, 711

Fan, X. et al. 2004, AJ, 128, 515

Freedman, W. L. 1994, Bulletin of the American Astronomical Society, 26, 1473

Freedman, W. L. et al. 2001, ApJ, 553, 47

Fujita, S. S. et al. 2003, ApJ, 586, L115

Fukugita, M. 2004, in IAU Symposium, p. 227

Fukugita, M., Hogan, C. J., Peebles, P. J. E. 1998, ApJ, 503, 518

Gallagher, J. S., Hunter, D. A., Bushouse, H. 1989, AJ, 97, 700

Gallego, J., Zamorano, J., Aragon-Salamanca, A., Rego, M. 1995, ApJ, 455, L1+

Gamow, G. 1946, Phys. Rev., 70, 572

Gao, L., Loeb, A., Peebles, P. J. E., White, S. D. M., Jenkins, A. 2003, ArXiv Astrophysics e-prints 0312499

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. 1996, Markov Chain Monte Carlo in Practice, Markov Chain Monte Carlo in Practice / Boca Raton, FL: Chapman and Hall

Glazebrook, K., Blake, C., Economou, F., Lilly, S., Colless, M. 1999, MNRAS, 306, 843

Glazebrook, K. et al. 2003, ApJ, 587, 55

Glazebrook, K. et al. 2004, Nature, 430, 181

Gordon, K. D., Clayton, G. C., Misselt, K. A., Landolt, A. U., Wolff, M. J. 2003, ApJ, 594, 279

Gunn, J. E. et al. 1998, AJ, 116, 3040

Hanany, S. et al. 2000, ApJ, 545, L5

Hasinger, G. 2003, ArXiv Astrophysics e-prints 0310804

Hastings, W. K. 1970, Biometrika, 57

Heavens, A., Panter, B., Jimenez, R., Dunlop, J. 2004, Nature, 428, 625

Heavens, A. F., Jimenez, R. 1999, MNRAS, 305, 770

Heavens, A. F., Jimenez, R., Lahav, O. 2000, MNRAS, 317, 965

Hopkins, A. M. 2004, ArXiv Astrophysics e-prints 0407170

Hubble, E., Humason, M. L. 1931, ApJ, 74, 43

Huchra, J. P., Brodie, J. P., Caldwell, N., Christian, C., Schommer, R. 1996, ApJS, 102, 29

Im, M., Simard, L., Faber, S. M., Koo, D. C., Gebhardt, K., Willmer, C. N. A., Phillips, A., Illingworth, G., Vogt, N. P., Sarajedini, V. L. 2002, ApJ, 571, 136

Jimenez, R., Bowen, D. V., Matteucci, F. 1999a, ApJ, 514, L83

Jimenez, R., Friaca, A. C. S., Dunlop, J. S., Terlevich, R. J., Peacock, J. A., Nolan, L. A. 1999b, MNRAS, 305, L16

Jimenez, R., MacDonald, J., Dunlop, J. S., Padoan, P., Peacock, J. A. 2004a, MNRAS, 349, 240

Jimenez, R., Panter, B., Heavens, A., Verde, L. 2004b, ArXiv Astrophysics e-prints 0403294

Kannappan, S. J. 2004, ApJ, 611, L89

Kauffmann, G. et al. 2003, MNRAS, 341, 33

Kennicutt, R. C. 1983, ApJ, 272, 54

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

Kennicutt, R. C. 1992b, ApJ, 388, 310

Kennicutt, R. C. 1998, ARA&A, 36, 189

Kennicutt, R. C., Tamblyn, P., Congdon, C. E. 1994, ApJ, 435, 22

Kochanek, C. S., Pahre, M. A., Falco, E. E., Huchra, J. P., Mader, J., Jarrett, T. H., Chester, T., Cutri, R., Schneider, S. E. 2001, ApJ, 560, 566

Kodama, T. et al. 2004, MNRAS, 350, 1005

Kong, X., Cheng, F. Z. 2001, MNRAS, 323, 1035

Kravtsov, A. V., Klypin, A. A., Khokhlov, A. M. 1997, ApJS, 111, 73

Kroupa, P. 2002, Science, 295, 82

Kurucz, R. L. 1979, ApJS, 40, 1

Lacey, C., Cole, S. 1993, MNRAS, 262, 627

Lacey, C., Cole, S. 1994, MNRAS, 271, 676

Lange, A. E. et al. 2001, Phys. Rev. D, 63(4), 042001

Lilly, S. et al. 1998, ApJ, 500, 75

Lilly, S. J., Le Fevre, O., Hammer, F., Crampton, D. 1996, ApJ, 460, L1+

Lin, W. P., Jing, Y. P., Lin, L. 2003, MNRAS, 344, 1327

Litzkow, M., Livny, M., Mutka, M. 1988, in Proceedings of the 8th International Conference of Distributed Computing Systems

Loveday, J. 2004, MNRAS, 347, 601

Lowenthal, J. 2001, PASP, 113, 127

Madgwick, D., Somerville, R., Lahav, O., Ellis, R. 2002, ArXiv Astrophysics e-prints

Martin, C. L., Kennicutt, R. C. 2001, ApJ, 555, 301

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. 1953, Journal of Chemical Physics, 21, 1087

Mulchaey, J. S., Dressler, A., Oemler, A. (eds.) 2004, Measuring and Modeling the Universe

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

Nakamura, O., Fukugita, M., Yasuda, N., Loveday, J., Brinkmann, J., Schneider, D. P., Shimasaku, K., SubbaRao, M. 2003, AJ, 125, 1682

Nolan, L. A., Dunlop, J. S., Jimenez, R., Heavens, A. F. 2003, MNRAS, 341, 464

Oliver, S. et al. 2000, MNRAS, 316, 749

Ouchi, M. et al. 2004, ApJ, 611, 660

Pérez-González, P. G., Gil de Paz, A., Zamorano, J., Gallego, J., Alonso-Herrero, A., Aragón-Salamanca, A. 2003, MNRAS, 338, 525

Padoan, P., Nordlund, Å. 2002, ApJ, 576, 870

Panter, B., Heavens, A. F., Jimenez, R. 2003, MNRAS, 343, 1145

Peacock, J. A. 1999, Cosmological physics, Cosmological physics. Cambridge University Press, 1999. ISBN: 0521422701

Peacock, J. A., Jimenez, R., Dunlop, J. S., Waddington, I., Spinrad, H., Stern, D., Dey, A., Windhorst, R. A. 1998, MNRAS, 296, 1089

Peebles, P. J. E. 1970, AJ, 75, 13

Peebles, P. J. E. 1983, ApJ, 274, 1

Penzias, A. A., Wilson, R. W. 1965, ApJ, 142, 419

Perlmutter, S. et al. 1999, ApJ, 517, 565

Persic, M., Salucci, P. 1992, MNRAS, 258, 14P

Press, W. H., Schechter, P. 1974, ApJ, 187, 425

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. 1992, Numerical recipes in FORTRAN. The art of scientific computing, Cambridge University Press, 1992, 2nd ed.

Rabin, D. 1982, ApJ, 261, 85

Reichardt, C., Jimenez, R., Heavens, A. F. 2001, MNRAS, 327, 849

Ronen, S., Aragon-Salamanca, A., Lahav, O. 1999, MNRAS, 303, 284

Rosa-González, D., Terlevich, E., Terlevich, R. 2002, MNRAS, 332, 283

Salpeter, E. E. 1955, ApJ, 121, 161

Salucci, P., Persic, M. 1999, MNRAS, 309, 923

Saracco, P., Longhetti, M., Giallongo, E., Arnouts, S., Cristiani, S., D'Odorico, S., Fontana, A., Nonino, M., Vanzella, E. 2004, A&A, 420, 125

Scalo, J. M. 1986, Fundamentals of Cosmic Physics, 11, 1

Schechter, P. 1976, ApJ, 203, 297

Scott, S. E. et al. 2002, MNRAS, 331, 817

Shen, S., Mo, H. J., White, S. D. M., Blanton, M. R., Kauffmann, G., Voges, W., Brinkmann, J., Csabai, I. 2003, MNRAS, 343, 978

Shimasaku, K. et al. 2001, AJ, 122, 1238

Smith, A. M., Cornett, R. H. 1982, ApJ, 261, 1

Somerville, R. S., Kolatt, T. S. 1999, MNRAS, 305, 1

Somerville, R. S., Lemson, G., Kolatt, T. S., Dekel, A. 2000, MNRAS, 316, 479

Somerville, R. S., Primack, J. R. 1999, MNRAS, 310, 1087

Spergel, D. N. et al. 2003, ApJS, 148, 175

Spinrad, H., Dey, A., Stern, D., Dunlop, J., Peacock, J., Jimenez, R., Windhorst, R. 1997, ApJ, 484, 581

Stanway, E. R., Bunker, A. J., McMahon, R. G. 2003, MNRAS, 342, 439

Steidel, C. C., Giavalisco, M., Pettini, M., Dickinson, M., Adelberger, K. L. 1996, ApJ, 462, L17+

Stoughton, C. et al. 2002, AJ, 123, 485

Strader, J., Brodie, J. B. 2004, ArXiv Astrophysics e-prints 0407001

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

Strauss, M. A. et al. 2002, AJ, 124, 1810

Sullivan, M., Treyer, M. A., Ellis, R. S., Bridges, T. J., Milliard, B., Donas, J. 2000, MNRAS, 312, 442

Tegmark, M., Taylor, A. N., Heavens, A. F. 1997, ApJ, 480, 22

Thain, D., Tannenbaum, T., Livny, M. 2004, Concurrency and Computation: Practice and Experience

Tresse, L., Maddox, S. J. 1998, ApJ, 495, 691

Verde, L., Oh, S. P., Jimenez, R. 2002, MNRAS, 336, 541

Waddington, I., Dunlop, J. S., Peacock, J. A., Windhorst, R. A. 2001, MNRAS, 328, 882

Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V., Dekel, A. 2002, ApJ, 568, 52

Worthey, G. 1994, ApJS, 95, 107

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

Appendix: Journal Articles

This appendix contains the following papers, written in support of the work of this thesis:

'*Star Formation and Metallicity History of the SDSS Galaxy Survey: Unlocking the Fossil Record*', B. Panter, A.F. Heavens, R. Jimenez, Monthly Notices of the Royal Astronomical Society, Volume 343, Issue 4, pp. 1145-1154

'*The Star-Formation History of the Universe from the Stellar Populations of Nearby Galaxies*', A.F. Heavens, B. Panter, R. Jimenez, J. Dunlop, Nature, Volume 428, Issue 6983, pp. 625-627 (2004)

'*The Mass Function of the Stellar Component of Galaxies in the Sloan Digital Sky Survey*', B. Panter, A.F. Heavens, R. Jimenez, Monthly Notices of the Royal Astronomical Society, in press.

'*Baryonic Conversion Tree: The Global Assembly of Stars and Dark Matter in Galaxies from the SDSS*', R. Jimenez, B. Panter, A.F. Heavens, L. Verde, Monthly Notices of the Royal Astronomical Society, in press.