The Probe of the Accelerating Universe Survey (PAUS)

First Year Progression Review (2017/2018)

Giorgio Manzoni

Abstract

In this document I report all the work done in my first year of the CDT PhD programme, and the work that I plan to do in the next years, under the supervision of Peder Norberg and Carlton Baugh. First I give a short introduction (Section 1) to my PhD project. Then, in Section 2, I explain short-term projects carried out in the first year. This section includes the two months team project for P&G and the work I have been doing on the evolution of the colour-magnitude using VIPERS data and GALFORM semi-analytic models. Section 3 is the core of the report as it explains the work done on my main PhD project in the first year. I conclude with Section 4, which explains how I will continue from the work done to obtain the main results of my PhD, i.e. the development, testing and application of an innovative group finder able to work in redshift space.

Contents

1 Introduction

The main project of my PhD makes use of the data collected by the Probe of the Accelerating Universe Survey (PAUS). PAUS is an ongoing photometric redshift survey which is unique since it aims to bridge the gap between small pencil-beam and sparse wide-area spectroscopic galaxy surveys. In fact PAUS is surveying large contiguous areas with high density of galaxies with sub-percent photometric redshift accuracy (Eriksen et al., 2018). To achieve this precision, the survey has been designed with 40 contiguous narrow bands (NB) filters, ~ 10 nm wide, ranging from 450 nm to 850 nm. This combination of NB filters makes PAUS up to ten times more precise than any other broad band photometric redshift survey with an accuracy of at least $0.004(1 + z)$ for over 50% of the galaxies up to $i_{\rm AB} \approx 22.5$.

Thanks to its depth and number density (tens of thousands of galaxies per deg^2 with a median redshift of $z \sim 0.7$), PAUS will result in one of the most detailed studies of intermediate-scale cosmic structure ever undertaken. With these features, PAUS has the capability of probing the galaxy clustering in the transition from the linear to the non-linear regime (Stothert et al., 2018a). Moreover, the study of galaxy groups will be further improved since spectroscopic surveys are either too sparse (e.g. BOSS), too shallow (e.g. GAMA), too narrow (e.g. COSMOS) or a combination of those characteristics (e.g. VIPERS).

Since the study of galaxy groups is a key science driver for PAUS, it must rely on a well grounded group finder algorithm, and here is where my work comes into play. In particular, Stothert et al. (2018a) showed how a simple Friends of Friends (FoF) algorithm is inadequate to take into account probabilistic redshifts with different accuracies as in the case of PAUS and how a Markov CLustering (MCL) technique could be more suitable for a photometric redshift survey like this. The main purpose of my PhD project is to develop, test and finally apply this innovative MCL algorithm to PAUS data and mocks.

To do this, a deep understanding of the survey is needed. This has been the direction of my efforts in my first year. I have been part of the validation process of the survey and, attending weekly telecons, I have built the experience needed to deal with a group finder aimed to specifically work on photometric redshifts.

As part of the process of building experience within the survey, I also took part to a PAUS observing run which took place at the William Hershel Telescope (WHT) in La Palma.

In Section 2, I describe other activities which were part of my first year training although not directly connected to the main project. These activities include a two months team project for Procter & Gamble (P&G) and the use of GALFORM semi-analytic models to finalise a paper draft related to the quenching of star formation activities in galaxies.

2 Short-term projects

2.1 Two months team project with Procter & Gamble (P&G)

This project is part of the CDT data intensive training. The team were made by three CDT students (Kevin Kwok, Miguel Icaza and I), an internal leader (Richard Bower), and two P&G leaders (Stefan Egan and Arturo Martinez). Note that there is a Non Disclosure Agreement (NDA) in place with P&G. This mean that the content of this section, in particular Equations 6 and 9, can be read by Durham University Staff only.

The final aim of the project was to help the P&G team to find a more reliable analytic form of the function which describe the density of the laundry powder. The P&G team has previously identified a library of *base functions* suggested by chemistry processes. Each base function might depend on 1 up to 13 of the available ingredients:

$$f_i = f_i(x_1, \cdots, x_{13}) \tag{1}$$

where f_i are the base functions ($i = 1, \cdots, N$) and x_j are the ingredients ($j = 1, \cdots, 13$). They made up the density function as a linear combination of those base functions and empirically they have chosen the combination that works better, setting the coefficient of their base functions f_i to their empirical fixed values $\hat{a}_i$:

$$\text{density} = \sum_{i=1}^{N} \hat{a}_i f_i(x_1, \cdots, x_{13}) \tag{2}$$

where N is the number of bases functions used.

The first step to make this process more quantitative is to use Eq. 3 as a fitting function leaving a_i as free parameters and finding them through the minimisation of the χ^2:

$$f_{\text{fit}} = \sum_{i=1}^{N} a_i f_i(x_1, \cdots, x_{13}) \tag{3}$$

$$\chi^2 = \sum_{k=1}^{M} \frac{(y_k - f_{\text{fit}}(x_{k_1}, \cdots, x_{k_{13}}))^2}{M} \tag{4}$$

where N is the number of base functions and M is the number of data-points.

The minimisation of the χ^2, although computationally expensive because of the high number of base functions (hence the high number of parameters a_i), will result in the best fit, i.e. the one which minimise the distance to the data points. The best fit is not necessarily what we want as it might rely on our choice of base functions f_i and the specific data-set that we are using (x_k, y_k). For example if we add a further base function to the current set of base functions, the χ^2 minimisation will also make use of this further base function to obtain a better fit. This makes the resulting fit dependent

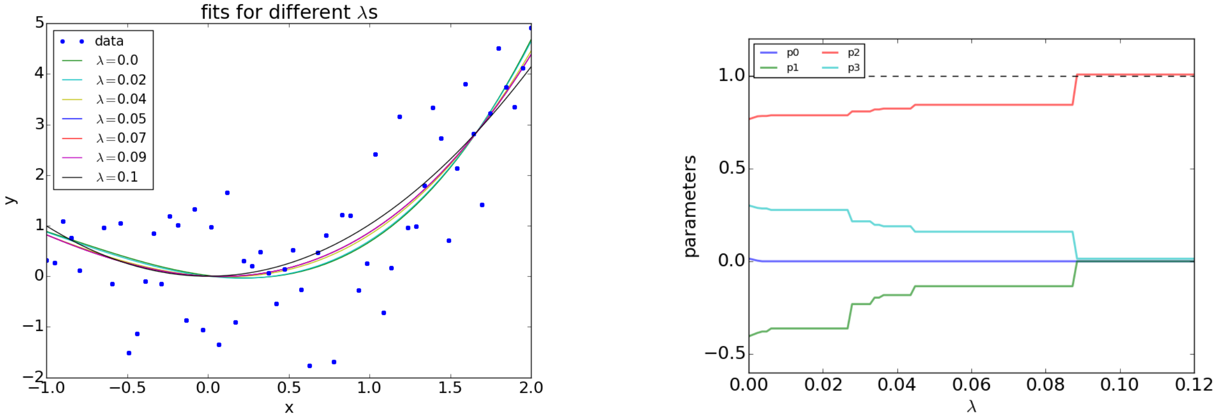

Figure 1: Example of the performances of the modified LASSO in recovering a known function when some gaussian noise is added. **Left-hand panel**: examples of modified LASSO fits for different λs on 60 data points generated from an x^2 function ($f_{\rm true}(x) = x^2$) with the addition of a gaussian noise with standard deviation 0.9. The fitting function is a third order polynomial. **Right-hand panel**: trend of the parameters of the fitting function ($f_{\rm fit} = p_0 + p_1 x + p_2 x^2 + p_3 x^3$) as a function of the λ parameter. Since we know $f_{\tt true}$, we know that the right answer is $p_2 = 1$ (red line) and $p_0 = p_1 = p_3 = 0$. This solution is achieved only for $\lambda \gtrsim 0.09$.

on our arbitrary choice to add a base function. At the same time, if we change data-set (x_k, y_k), the fit obtained with the old data-set will not minimise the χ^2 any longer (as the best fit is the best for a specific data-set). This makes our results data dependent, hence not robust. That is why, it is better to find a fit that is slightly worse than the χ^2 one in a specific data-set but that is still good when we change data-set.

The two problems (arbitrary choice of base functions and having a data-dependent fit) are related. In particular, the more base functions we use, the more the χ^2 fit will be data-dependent (the so-called *overfitting problem*). Therefore it seems that minimising the number of base functions can help to have a more robust fit.

If we succeed in reducing the number of base functions f_i (setting the relative parameter $a_i = 0$) we might obtain a further advantage coming from the fact that not all of the f_i depend on all the 13 ingredients x_j. For example there might be some of them depending on a single ingredient (e.g. $f_{i_0} = f_{i_0}(x_7)$ instead of $f_{i_0} = f_{i_0}(x_1, \cdots, x_{13})$). In that case (assuming that the ingredient x_7 does not appear in any other base functions f_i with $i \neq i_0$) we learn that this particular ingredient doesn't affect the density at all, and this is particularly relevant for the efficiency of the industrial process[1].

A possible way to address (at least partially) this problem is to use the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani 1996) which basically add a *penalty term* to the usual χ^2 in order to penalise the sum of the parameters, i.e. favouring a fit with a small value of $\sum_{i=1}^{N} |a_i|$. The classic analytic form of LASSO is the following:

$$\text{LASSO} = \boxed{\sum_{k=1}^{M} \frac{(y_k - f_{\rm fit}(x_{k_1}, \cdots, x_{k_{13}}))^2}{M} + \lambda \sum_{i=1}^{N} |a_i|} \equiv \chi^2 + \text{Penalty} \tag{5}$$

where, as always, M is the number of data points, N is the number of base functions and λ is a free parameter that we can tune case by case (the choice of the perfect λ is still an open debate and one possible idea might be to use a machine learning approach to train the algorithm on cases with a known solution). This helps to discard the base functions that are not extremely relevant as, when minimising Eq. 5, a fit with a lot of bases function will result in a bigger penalty term than a fit with less bases functions.

However, we noticed that the traditional LASSO as in Eq. 5 may be improved to work in a P&G-like case, where we want to minimise only the number of functions and not the actual value of the

[1] Although it sounds like a very unlikely situation, it is very likely for the real P&G case.

function coefficients $|a_i|$ (while in the traditional LASSO we are minimising both). In fact minimising the current penalty will result in favouring solutions with small values of $|a_i|$ that is different from favouring a solution with exactly $a_i = 0$ (having some $a_i = 0$ means having less base functions). So we developed a new *modified LASSO* to deal with this problem. In particular we slightly changed the penalty function in this way:

$$\text{New Penalty} = \lambda \left[\left(\sum_{i=1}^{N} |a_i| \right)^2 - \left(\sum_{i=1}^{N} a_i^2 \right) \right] \tag{6}$$

In this new penalty function parameters that are zero are favoured as they appeared only in cross terms. For example, let us assume we have just 3 parameters. The new penalty will be: $(|a_1| + |a_2| + |a_3|)^2 - a_1^2 - a_2^2 - a_3^2 = |2a_1a_2| + |2a_1a_3| + |2a_2a_3|$. If one of the parameters is null, it will make null the product with another parameter and this will make the penalty smaller than having small non zero parameters.

Another improvement that can be done in order to make the LASSO technique more effective is the normalisation of the base functions. Although logically the normalisation is the first thing to do, I introduce it only now because is part of our improvement to LASSO that until now has always been used in the literature without any normalisation. Specifically, the problem that we want to address is that as long as we do not normalise the base functions, different parameters could affect in very different ways the LASSO functions, hence the result of the fit. For example if a base function is $f_{i_0} = x$ or $f_{i_1} = 5x$, the parameter a_{i_0} and a_{i_1} will have the same weight in the penalty function ignoring the fact that they carry different information. That is why we decided to normalise the data and the base function image in the following way:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \hat{y} = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \quad \hat{f}(\hat{x}) = \frac{f(x(\hat{x})) - f_{\min}(x(\hat{x}))}{f_{\max}(x(\hat{x})) - x_{\min}(x(\hat{x}))} \tag{7}$$

where $x \in [x_{\min}, x_{\max}] \Rightarrow \hat{x} \in [0,1]$, $y \in [y_{\min}, y_{\max}] \Rightarrow \hat{y} \in [0,1]$ and $f(x) \in [f_{\min}(x), f_{\max}(x)] \Rightarrow \hat{f}(\hat{x}) \in [0,1]$. To deal with normalised variables, also the χ^2 needs to be normalised in order to have the same order of magnitude of the penalty function. We introduce the ϵ^2 parameter which is the minimum value of the χ^2 (without the penalty) to normalise our modified LASSO to 1 when $\lambda = 0$:

$$\epsilon^2 = \min \left\{ \sum_{\mathrm{k}=1}^{\mathrm{M}} \frac{(\mathrm{y_k} - \mathrm{f_{fit}}(\mathrm{x_{k_1}}, \cdots, \mathrm{x_{k_{13}}}))^2}{\mathrm{M}} \right\} \tag{8}$$

With those definitions, our modified lasso acquire the following form:

$$\text{MODIFIED LASSO} = \boxed{\frac{1}{\epsilon^2} \sum_{k=1}^{M} \frac{(y_k - f_{\text{fit}}(x_{k_1}, \cdots, x_{k_{13}}))^2}{M} + \lambda \left[\left(\sum_{i=1}^{N} |a_i| \right)^2 - \left(\sum_{i=1}^{N} a_i^2 \right) \right]} \tag{9}$$

To test our new implementation of LASSO, we created increasingly difficult known functions (rising the number of variables and combining polynomials with exponentials terms). After having added various kinds of noise with different amplitudes, we tried to recover the original functions, fitting with the *modified LASSO* of Eq. 9.

A simple scenario is shown in Fig. 1 where we generated 60 data-points using the function $f_{\text{true}} = x^2$ with additional Gaussian noise with a standard deviation of 0.9. We fitted the data-point using a third order polynomial $f_{\text{fit}}(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ knowing that the solution is $a_2 = 1$ and $a_0 = a_1 = a_3 = 0$. Although, looking at the left panel, the fits for different λ do not look very different, in the right panel we notice that we recover the true solution only for $\lambda \gtrsim 0.09$. Picking the right λ is not trivial, and currently there are no rules in the literature to decide which is the right value. In the future, it might be interesting to train our algorithm to choose the right λ using known case studies like the x^2 one.

2.2 The quenching of star-formation in VIPERS and GALFORM

In September 2018 I attended the workshop "Understanding Emission-line galaxies for the next generation of cosmological survey" in Teruel (Spain). On that occasion I presented the work done in my master project which studied the redshift evolution of the colour-magnitude relation in galaxies with the aim of getting insights on the quenching of star-formation (SF) activity. For this project I have used data from the VIMOS Public Extragalactic Redshift Survey (VIPERS, Guzzo et al. 2014; Scodeggio et al. 2016)[2], a spectroscopic survey which covers 23.5 deg^2 at intermediate redshifts ($0.4 < z < 1.3$).

Specifically I studied the evolution of the bright-edge of the colour-magnitude for 7 different redshift bins, as shown in Fig. 2. The bright-edge is a good tracer of the star formation activity as it is determined by the population of bright galaxies that are getting fainter because of a change in star-formation activity. The speed at which the bright-edge moves with redshift might be related to the quenching of star-formation activity that pushes individual galaxies to move into the red sequence of quiescent galaxies. The faint-end of the colour-magnitude diagram instead does not bring any information at all as it is determined by the depth of the survey (in the case of VIPERS, the cut is at $i_{\rm AB} < 22.5$).

To consistently identify the bright-edge, I have developed an algorithm which, after having splitted the colour-magnitude diagram in colours bins, it fixes the edge when the distribution of magnitudes drops below a certain percentage of the most populated magnitude bin (see thick lines in Fig. 2). The trend in Fig. 2 shows that the bright-edge at high redshifts is defined by galaxies that are clearly brighter than those at low redshift (galaxies at high redshift are brighter than what is observed at low redshift).

One way of getting some information on the star-formation activity of galaxies exploiting what we know about the location of the bright-edge is to build some stellar population synthesis models. In particular, I have used the PEGASE 2 code to generate these models in order to try and reproduce the evolution of the bright-edge, testing different star formation histories (SFH, i.e star-formation rate (SFR) as a function of time) with and without a quenching event[3] occurring. We can see the results of this calculation in Fig. 3. In particular, Fig. 3 shows the comparisons between the observed VIPERS galaxies in the lowest redshift bin ($0.4 < z < 0.5$) drawn as yellow points (which are the same in the three panels) and galaxies in the highest redshift bin ($1.0 < z < 1.3$) after having experienced a synthetic evolution of 3.26 Gyr (green points). This is the interval time necessary to bring a typical galaxy from $z \sim 1.11$ to $z \sim 0.47$ (median redshift in the highest and lower bin respectively). The three different panels show different synthetic evolutions, i.e. different star-formation histories. Specifically, the panel on the left-hand side is obtained with a simple delayed exponential SFH as described by Gavazzi et al. (2002), i.e. following Eq. 10,

$$\mathrm{SFR(t,}\tau) = \frac{\mathrm{t}}{\tau^2} \exp\left[-\frac{\mathrm{t}^2}{2\tau^2}\right] \tag{10}$$

where t is the cosmic time and τ is the unique parameter which sets the delay. In this scenario, there is no quenching occurring. In the center panel I introduced a quenching (SFR instantaneously[4] brought to zero) for all the galaxies at the epoch of observation, that in this case is $z \sim 1.11$. The rights panel instead shows a scenario in which each galaxy experience the quenching at a random time between the epoch of observation and a time interval of 3 Gyrs afterwards. This guarantee that after 3 Gyrs all the galaxies in the sample are quenched.

The agreement between the observed bright-edge and the synthetic one (blue and red line respectively, in Fig. 3) can be used as an indicator of which synthetic SFH better reproduce the data, hence which quenching scenario is more realistic. In the "no quenching scenario" (left panel) the evolution of galaxies seems too mild, leaving a lot of blue and bright star-forming galaxies surviving at low redshift, in contrast with observations. In the central panel instead, where all the galaxies are quenched exactly at the epoch of observations, the evolution looks too fast bringing all the synthetic galaxies at low redshift to be in the red sequence with no galaxies left in the blue cloud.

[2] The data I have used comes from the second (and complete) Public Data Release (PDR-2) which is now available at `http://vipers.inaf.it/`.

[3] When I speak about a quenching event, I refer to the suppression of the star-formation of a galaxy in a maximum interval of time of 100,000 years (which corresponds to the resolution of the PEGASE models).

[4] Instantaneously means a time of 100,000 years, i.e. the resolution of PEGASE models.

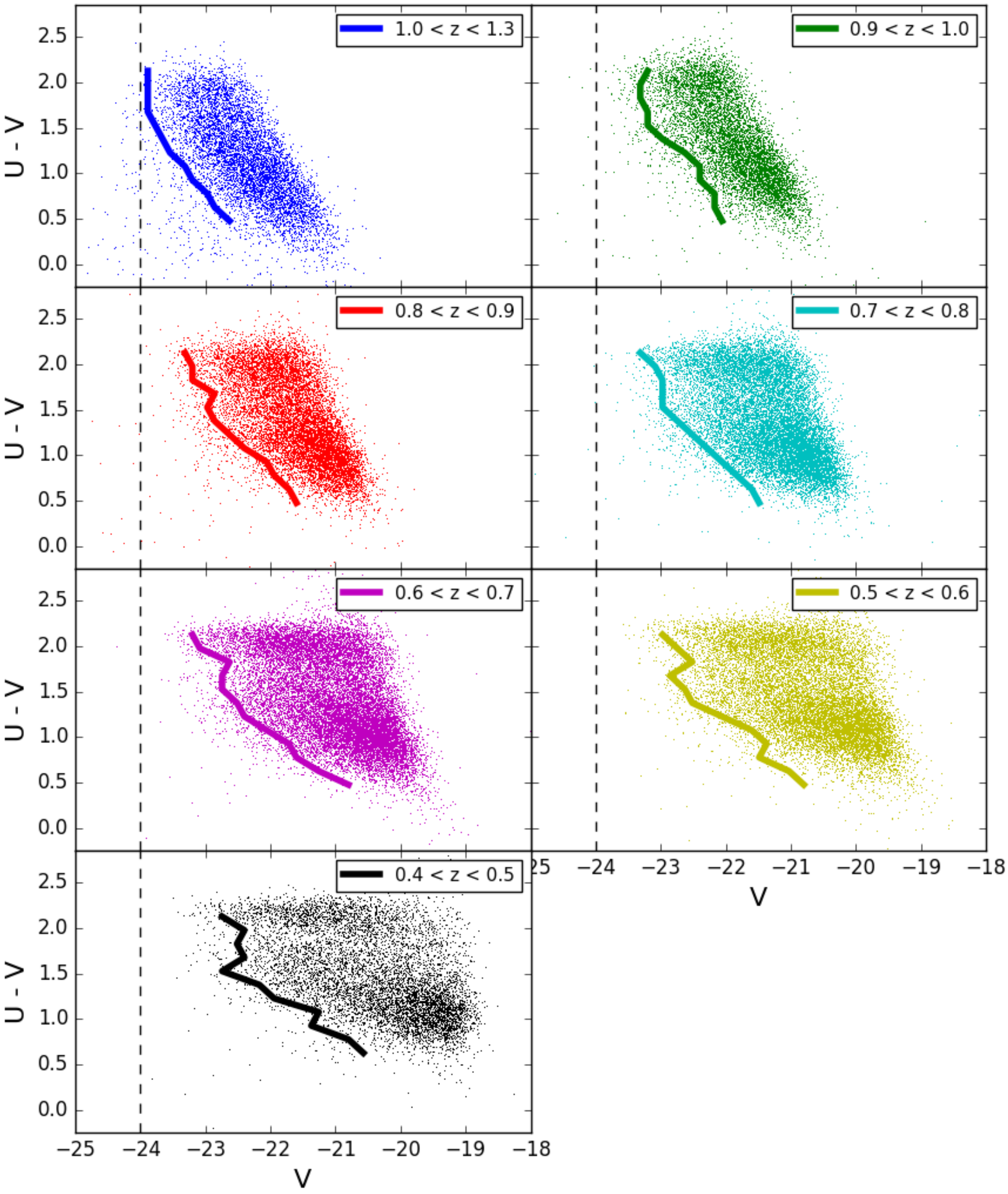

Figure 2: VIPERS colour-magnitude diagram in 7 different redshift bins. The thick line marks the bright-edge as defined by the drop in number counts to the 15% of the most populated bin.

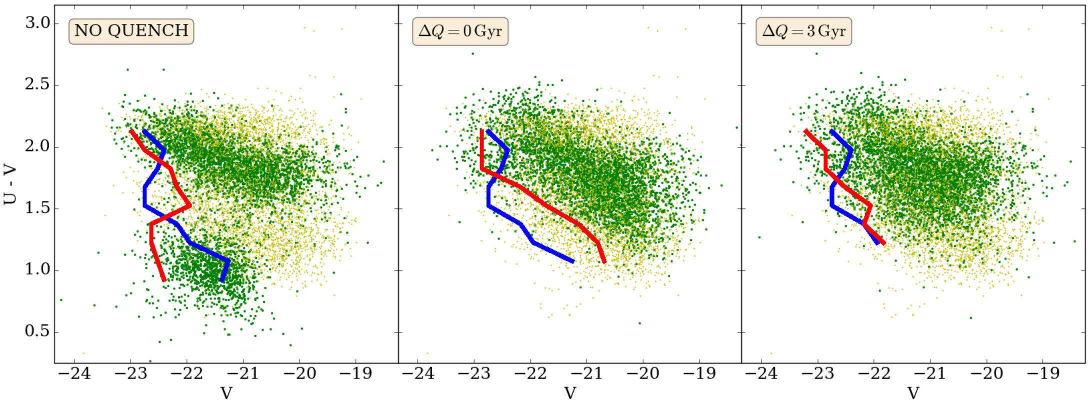

Figure 3: Results of the synthetic evolution of the colour-magnitude relation following PEGASE models with 3 different star formation histories. In the three panels, the yellow points are the same, i.e. the observed galaxies in the lower redshift bin ($0.4 < z < 0.5$). The green points instead are the galaxies in the highest redshift bin ($1.0 < z < 1.3$) after having experience a synthetic evolution of 3.26 Gyr, that is the time needed to move from the $z \sim 1.11$ to $z \sim 0.47$ (the median redshift in the respective bins). The blue line is the observed bright-edge (relative to the yellow points) while the red line is the synthetic bright-edge (relative to the green points). The three panel are respective (from left to right) a simple delayed exponential SFH, SFH with quench at the epoch of observation ($z \sim 1.11$ in this case) and SFH with a random quench spread in an interval time of 3 Gyr after the epoch of observation.

Finally the conclusion of this work is that a scenario in which galaxies are randomly quenched within an interval time of 3 Gyrs (right panel) seems to be the one that produce a better agreement between the observed and the synthetic reproduced bright-edge, hence a better modelling of the SFH.

For this project I have an advanced paper draft (attached to this document). At the Teruel workshop I received some good feedback about how to make these results more robust. One of the suggestions I am still working on is the comparison of these results with the ones obtained from a semi-analytic galaxy formation model, in particular I am using the Gonzalez-Perez et al. (2014) version of GALFORM (GP14).

The advantage of semi-analytics models is that we can actually see the behaviour of progenitor galaxies following a specific branch of the merger tree. In this way we can study how related galaxies move in the colour-magnitude diagram. This is something we can not do with real data as we do not have information about the connection between galaxies at different redshift.

Another advantage of semi-analytics models like GALFORM is that we can run them multiple times changing the parameters that are more important. In our investigation, the parameters we want to change are those relative to the quenching mechanism.

For this reason it is important to be aware of which quenching process is potentially more important in affecting the colour-magnitude diagram. This test can be done identifying the population of galaxies close to the bright-edge. Using GP14, I realised that central galaxies are dominant with respect to satellites. That means that the quenching in satellites (usually ram pressure stripping) would have just a minor effect in the location of the bright-edge of the colour-magnitude plane. Instead AGN feedback (main quenching process for centrals in GALFORM) would play a crucial role in the evolution of the bright-edge.

Knowing this, the final test that needs to be done is running GALFORM switching the AGN feedback on and off, and testing which scenario better agrees with VIPERS observations.

A realistic timescale to finish off this project and submit the paper (while continue to work on my main project described in Section 3) is late January.

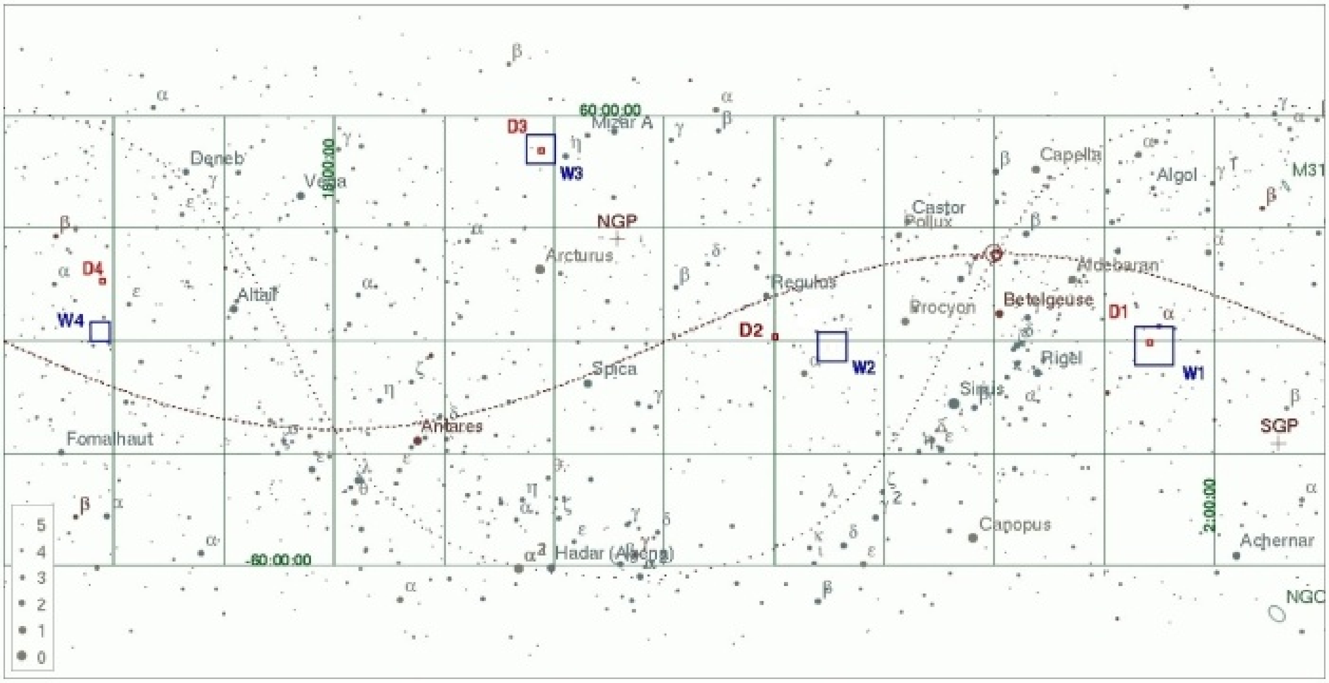

Figure 4: CFHTLS Wide (W1, W2, W3, W4) and deep (D1, D2, D3, D4) fields. Among those fields, PAUS is observing all of the Wide and D2 which corresponds to the COSMOS field.

3 Main project

3.1 PAUS Validation

The PAU Survey is currently collecting data using the specifically designed PAUCam (Castander et al., 2012), an optical wide field imaging camera (1 deg^2 Field-of-View, FoV) installed at the prime focus of the William Hershel Telescope (WHT). The camera is installed with 6 large ugrizY filters (not used in PAUS), each of those covering the total focal plane and 5 trays with 8 narrow band (NB) filters each (covering the 8 central CCDs) for a total of 40 NB filters. These contiguous NB filters, which are the main feature of PAUS, are 10 nm wide and cover a wavelength range from 450 to 850 nm.

PAUS is aiming to cover the COSMOS field (COSMOS/D2) and the four wide fields of CFHTLS (Fig. 4: W1, W2, W3 and W4). For each field, PAUS is collecting photometric images in all of the NBs, with the aim of performing forced aperture photometry. This means that independently on the signal to noise ratio, PAUS takes the position of the galaxies from a parent survey and extract the flux within a certain radius. Information about the radius to be used are also coming from the parent catalogue. For example one choice is to use the half-light radius of a galaxy.

Early photometric results for galaxies in the COSMOS field, which used COSMOS as a parent survey (Scarlata et al., 2007), are already available in Eriksen et al. (2018).

When reducing data from the Wide fields, another parent survey needs to be chosen as there is no longer overlap with COSMOS. The choice of this new parent survey is very important and it may introduce any sort of bias, especially if the definition of the radius is different from the one that has been used in the COSMOS field.

The choice of this new parent catalogue has been an important part of the first year of my PhD. In particular, I analysed two parent surveys: CFHTLS and CFHTLenS. While CFHTLS covers both the Wide and the Deep fields (D1, D2, D3, D4, W1, W2, W3, W4), CFHTLenS covers just the Wide ones (W1, W2, W3, W4). Since COSMOS overlap only with Deep-2 (D2), a comparison *object by object* between PAUS/D2 and CFHTLenS is not possible. hence, the first analysis that I did was a statistical comparison between the distribution of radii in COSMOS/D2 and in CFHTLenS/W3. After that, I used CFHTLS as a "bridge" comparing *object by object* COSMOS/D2 with CFHTLS/D2 and then (assuming that all the CFHTLS Deep fields adopt the same definition of radius) I compared

CFHTLS/D3 with CFHTLenS/W3.

Both the distribution from CFHTLS radii and from CFHTLenS are consistent to the COSMOS one. However to make the result more precise we can apply a linear transformation to the radii. I found this relation imposing the median and the σ_{68} of the distribution to be the same:

$$\mathrm{median}(a + b \times r_{\mathrm{new}}) := \mathrm{median}(r_{\mathrm{COSMOS}})$$

$$\sigma_{68}(a + b \times r_{\mathrm{new}}) := \sigma_{68}(r_{\mathrm{COSMOS}}) \tag{11}$$

Doing the math, I worked out that:

$$a = \mathrm{median}(r_{\mathrm{COSMOS}}) - \frac{\sigma_{68}(r_{\mathrm{COSMOS}})}{\sigma_{68}(r_{\mathrm{new}})} \times \mathrm{median}(r_{\mathrm{new}})$$

$$b = \frac{\sigma_{68}(r_{\mathrm{COSMOS}})}{\sigma_{68}(r_{\mathrm{new}})} \tag{12}$$

The PAUS collaboration decided to use CFHTLenS because it has been designed for lensing purposes and this is one of the goals also for PAUS. Using Equations 11 and 12 I found the values to transform CFHTLenS radii in COSMOS-like radii:

$$r_{\mathrm{COSMOS-like}} = -0.043 + 1.757 \times r_{\mathrm{CFHTLenS}} \tag{13}$$

This is the relation that the PAUS collaboration has used to obtain radii for the forced photometry in W3.

One important science goal for PAUS is clustering studies. To study the clustering of galaxies, it is essential that observations have been carried out uniformly in every part of the observed field. Every sort of inhomogeneity will affect the clustering. I did tests on the PAUS/D2 (i.e. the COSMOS field) to check that instead there are no bias due to observations. Two factors are crucial to test this. First, the redshift accuracy

$$\delta z = \frac{|z_{\mathrm{spec}} - z_{\mathrm{phot}}|}{1 + z_{\mathrm{spec}}}$$

needs to be consistent in all the areas of the field. This implies that we have chosen a spectroscopic parent catalogue from which obtaining z_{spec}. z_{spec} is used here as a *true value* since the accuracy in spectroscopic redshift is better than photometric redshift by construction. Second, each area of the sky has been observed with a certain number of NB filters (the maximum number being 40). To guarantee that the counting of galaxies (hence the clustering) is reliable, we want to have this number homogeneous. While the redshift accuracy is mainly determined by the quality of observations (one key factor being the seeing, i.e. a parameter quantifying the amount of turbulence in the atmosphere), the different number of filters in different areas comes also from the observation strategy. In fact, to minimise the number of changes of the filter tray (hence to optimise the telescope time), it is preferable to observe contiguous areas of the sky with the same filter tray before changing it. Moreover, each filter tray contains 8 different NB filters in different part of the focal plane. This makes the counting of NB filters per area very important, particularly at the borders of the observed field. In order to repeat all of these tests on every production released by the PAUS data management team, I am developing a pipeline to produce all the relevant plots automatically.

Unlike spectroscopic redshift which comes directly from the measure of the shift of specific features in the spectrum, photometric redshifts are derived from the fit of the observed Spectral Energy Distribution (SED) of the galaxy with theoretical stellar population synthesis (SPS) models. To improve the quality of the fit, information about emission lines can be taken into account. In Eriksen et al. (2018) the choice that has been done for PAUS is to fix a set of emission line ratios as in Ilbert et al. (2009). The values of this ratios are shown in Table 2 of Eriksen et al. (2018). However a better modelling of emission lines can improve even more the determination of photometric redshifts.

For this reason I have considered empirical relations for some specific Baldwin-Philips-Terlevich (BPT) diagrams (Baldwin et al., 1981). These diagram shows the relation between some specific emission line ratios. The more relevant, considering the PAUS wavelength range, are:

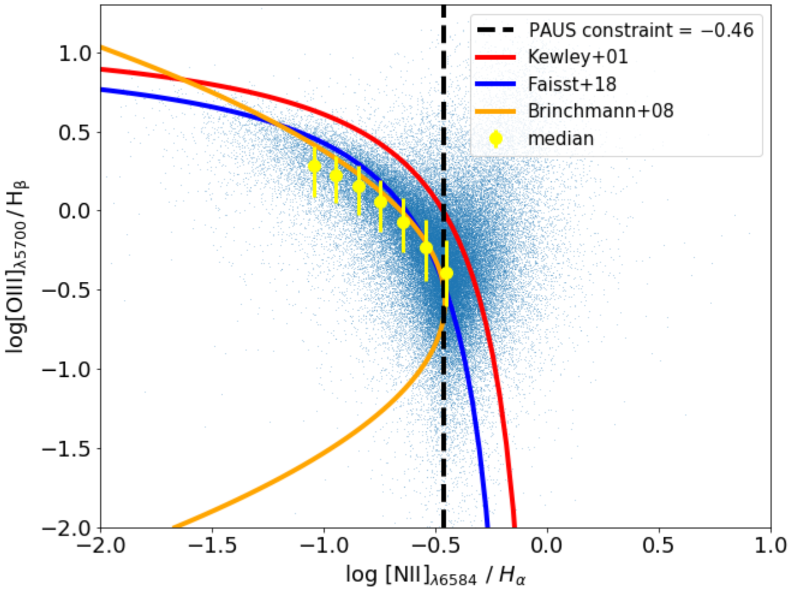

Figure 5: Empirical modelling for a particular BPT diagram based on specific emission line ratios, i.e. $[\mathrm{NII}]_{\lambda6584}$ / H_α versus $[\mathrm{OIII}]_{\lambda5700}$ / H_β. The data (blue points) comes from the GAMA survey (Driver et al., 2011). The vertical dashed black line is the PAUS assumption which brings no information about the [OIII]/H_β ratio as they just fix the [NII]/H_α ratio. The yellow filled circles are median values of [OIII]/H_β in bins of the [NII]/H_α ratio. The error bars represent the 25th and 75th percentiles. Brinchmann et al. (2008) and Faisst et al. (2018) (orange and blue line respectively) provide empirical relations for this specific BPT diagram. Kewley et al. (2001) instead (red line) provide an empirical relation to separate star forming galaxies (below the red line) and AGN (above the red line).

- $[\mathrm{NII}]_{\lambda6584}$ / H_α versus $[\mathrm{OIII}]_{\lambda5700}$ / H_β
- $[\mathrm{SII}]_{\lambda\lambda6717}$ / H_α versus $[\mathrm{OIII}]_{\lambda5700}$ / H_β
- $[\mathrm{OI}]_{\lambda6300}$ / H_α versus $[\mathrm{OIII}]_{\lambda5700}$ / H_β

In Fig. 5, there is an example of the work I have done on the first of the BPTs I have listed above. The data in Fig. 5 shows data (light-blue points) from the GAMA survey (Driver et al., 2011). The vertical dashed line instead is the PAUS assumption, which is a fixed value for the [NII]/H_α ratio. The yellow filled circles represents the median of [OIII]/H_β for bins of [NII]/H_α, each 0.1 dex wide. The error bars are obtained from the 25th and 75th percentile. The other coloured line instead comes from the literature as indicated in the legend. We can see that the PAUS assumption, although roughly describe the bulk of the population, may be used just for a first approximation. Using instead a relation like Faisst et al. 2018 (dark blue line) can help the SED fitting with a stronger constraint on the emission line ratios, hence a better photometric redshift estimation. I have done a similar work on the other BPT diagrams listed before, providing the PAUS team with a set of empirical relations to be used in the next photometric production. Once this new production is ready, the next step is to analyse how much the new emission lines scheme affect the photo-z results and if it actually enhance the redshift accuracy.

3.2 PAUS observations

To understand the nature of the data and to build experience within the survey, I took part in a PAUS observing run. The PAU survey is using the William Hershel Telescope (WHT), a 4.2 m diameter telescope based in La Palma (Spain) combined with a large field camera (PAUCam) specially designed for this survey (see Castander et al. 2012 for further details). I have been observing for 7 nights sided by expert observers who taught me how to practically set up the telescope in the afternoon and how to carry out observations during the night.

Afternoon calibrations include first checking the temperature and pressure of the camera and then acquiring BIAS and FLAT FIELD images. This procedures are aimed to calibrate the CCD in particular we are setting the right parameters in order to convert a number of electron[5] into a number of photons (i.e. the flux). Specifically the BIAS is an image with no light entering in the focal plane and it is obtained simply keeping close the dome and the petals of the telescope. In this way we are setting a certain value of the CCD electric current to be the zero point of our observation. We do this because the CCD works in a linear regime (number of incident photon proportional to the number of electron released) only for certain values of the electric current. The results will be that in case of no photons striking the CCD then the electric current released will be the bias current. The BIAS is just the CCD image generated by this bias current. Hence, every scientific image needs to be bias subtracted to obtain real values of the flux. The reason why the BIAS is an image and not just a number is that every pixel behaves slightly differently. Subtracting the BIAS takes into account all of this imperfection of the CCD pixels. Since the BIAS is a calibration with no lights, we also want to calibrate the CCD when it is actually measuring photons. That is why we measure the FLAT FIELD images. These are images of a uniform light source in order to calibrate the slope of the proportionality between photons and electrons released. There are some different techniques to obtain a uniform source of light. One of those is to directly observe the lights of the sunset. In the case of PAUS, there are specifically designed lamps inside the dome which produce uniform light. It is important at this stage to leave the dome close, so that no spurious light from other sources can enter, and open the petals of the telescope, so that the light from the lamp can reach the focal plane.

After that BIAS and FLAT FIELDS are taken, everything is set to start observing. Observations take place 20 minutes before the end of twilights. This is because the first thing we want to observe is a calibration star. The calibration star is an SDSS star which flux is known in order to compare the measured flux and obtain the so called zero point (that is used to convert the CCD output for the number of photons in units of flux). The observation strategy, i.e. deciding which filter tray, which exposure time and which field to observe, was mainly driven by the moon phase and position on the sky. Basically the purpose is to observe at any time of the night the field that is further from the moon and using a redder filter tray for larger fraction of the moon[6]. The exposure time instead depends on the quality of the sky with worse condition requiring longer time of exposure.

This observing experience was meant to be a training for next PAUS observations and to get hands on experience with the acquisition of the data that is to be analysed as part of my PhD thesis. In fact, I have been assigned 4 nights for next December 2018-B observing run.

4 Future project: Galaxy groups in PAUS

PAUS will result in one of the most detailed studies of intermediate-scale cosmic structure ever undertaken. Even though PAUS will cover a modest area compared to large wide imaging surveys like DES and KiDS, PAUS will increase the number density of galaxies with sub-percent accuracy redshifts by nearly two orders of magnitude, reaching tens of thousands of redshifts per deg^2. This will make possible to measure the clustering of galaxies in the transition from the linear to the non-linear regime (Stothert et al., 2018b). For these reasons, one of the key science drivers for PAUS is the study of galaxy groups. Galaxy groups are the observable counterparts to dark matter halos, so detecting galaxy groups can help us infer more about the galaxy-halo connection.

A group finder is an algorithm able to identify galaxy groups having as an input a galaxy catalogue. A variety of group finder algorithm have already been developed, with the most common being the so called Friend-of-Friend (FoF) algorithm. However Stothert et al. (2018a) pointed out that such an algorithm is not optimised nor ideal to work with probabilistic redshifts, as in the case of a photometric redshift survey like PAUS.

One innovative alternative is to develop a group finder based on the Markov CLustering (MCL) algorithm (Van Dongen, 2000). Stothert et al. (2018a), tested the MCL algorithm in real space on a mock galaxy catalogue constructed from an N-body simulation using the GALFORM semi-analytic

[5] Photons coming from the source strike the CCD that releases electrons that needs to be converted again into a number of photons in order to obtain the flux.

[6] Since the moon light peaks at blu wavelengths, a filter sensitive to redder wavelengths is less sensitive to the presence of the moon.

model (Gonzalez-Perez et al., 2014). His results show that the widely used FoF algorithm is just a subset of MCL.

The idea for my future work is to use Stothert et al. (2018a) as a starting point to develop, test and finally apply an MCL group finder able to work in redshift space, allowing for galaxy redshifts with different uncertainties. I will finally apply this new MCL group finder on the PAUS data and if successful it will be possible to run the algorithm on other photometric redshift surveys.

The benefits of having a reliable group finder reflects on a multiplicity of science goals. In the cosmology fields, for example, weak lensing around groups can help to have stronger constraints on cosmological parameters. In the galaxy formation and evolution field, we can study the mass-to-light ratio for dark matter halos in order to measure where the galaxy formation is most efficient. See for example Eke et al. (2004) and Viola et al. (2015). This will allow us to identify Milky Way analogues for a possible deeper observational follow up.

References

J. A. Baldwin, M. M. Phillips, and R. Terlevich. Classification parameters for the emission-line spectra of extragalactic objects. *PASP*, 93:5–19, Feb. 1981. doi: 10.1086/130766.

J. Brinchmann, M. Pettini, and S. Charlot. New insights into the stellar content and physical conditions of star-forming galaxies at z = 2-3 from spectral modelling. *MNRAS*, 385:769–782, Apr. 2008. doi: 10.1111/j.1365-2966.2008.12914.x.

F. J. Castander, O. Ballester, A. Bauer, L. Cardiel-Sas, J. Carretero, R. Casas, J. Castilla, M. Crocce, M. Delfino, M. Eriksen, E. Fernández, P. Fosalba, J. García-Bellido, E. Gaztañaga, F. Grañena, C. Hernández, J. Jiménez, L. López, P. Martí, R. Miquel, C. Neissner, C. Padilla, C. Pío, R. Ponce, E. Sanchez, S. Serrano, I. Sevilla, N. Tonello, and J. de Vicente. The PAU camera and the PAU survey at the William Herschel Telescope. In *Ground-based and Airborne Instrumentation for Astronomy IV*, volume 8446 of *Proc. SPIE*, page 84466D, Sept. 2012. doi: 10.1117/12.926234.

S. P. Driver, D. T. Hill, L. S. Kelvin, A. S. G. Robotham, J. Liske, P. Norberg, I. K. Baldry, S. P. Bamford, A. M. Hopkins, J. Loveday, J. A. Peacock, E. Andrae, J. Bland-Hawthorn, S. Brough, M. J. I. Brown, E. Cameron, J. H. Y. Ching, M. Colless, C. J. Conselice, S. M. Croom, N. J. G. Cross, R. de Propris, S. Dye, M. J. Drinkwater, S. Ellis, A. W. Graham, M. W. Grootes, M. Gunawardhana, D. H. Jones, E. van Kampen, C. Maraston, R. C. Nichol, H. R. Parkinson, S. Phillipps, K. Pimbblet, C. C. Popescu, M. Prescott, I. G. Roseboom, E. M. Sadler, A. E. Sansom, R. G. Sharp, D. J. B. Smith, E. Taylor, D. Thomas, R. J. Tuffs, D. Wijesinghe, L. Dunne, C. S. Frenk, M. J. Jarvis, B. F. Madore, M. J. Meyer, M. Seibert, L. Staveley-Smith, W. J. Sutherland, and S. J. Warren. Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release. *MNRAS*, 413:971–995, May 2011. doi: 10.1111/j.1365-2966.2010.18188.x.

V. R. Eke, C. S. Frenk, C. M. Baugh, S. Cole, P. Norberg, J. A. Peacock, I. K. Baldry, J. Bland-Hawthorn, T. Bridges, R. Cannon, M. Colless, C. Collins, W. Couch, G. Dalton, R. de Propris, S. P. Driver, G. Efstathiou, R. S. Ellis, K. Glazebrook, C. A. Jackson, O. Lahav, I. Lewis, S. Lumsden, S. J. Maddox, D. Madgwick, B. A. Peterson, W. Sutherland, and K. Taylor. Galaxy groups in the Two-degree Field Galaxy Redshift Survey: the luminous content of the groups. *MNRAS*, 355: 769–784, Dec. 2004. doi: 10.1111/j.1365-2966.2004.08354.x.

M. Eriksen, A. Alarcon, E. Gaztanaga, A. Amara, L. Cabayol, J. Carretero, F. J. Castander, M. Delfino, J. De Vicente, E. Fernandez, P. Fosalba, J. Garcia-Bellido, H. Hildebrandt, H. Hoekstra, B. Joachimi, P. Norberg, R. Miquel, C. Padilla, A. Refregier, E. Sanchez, S. Serrano, I. Sevilla-Noarbe, P. Tallada, N. Tonello, and L. Tortorelli. The PAU Survey: Early demonstration of photometric redshift performance in the COSMOS field. *ArXiv e-prints*, Sept. 2018.

A. L. Faisst, D. Masters, Y. Wang, A. Merson, P. Capak, S. Malhotra, and J. E. Rhoads. Empirical Modeling of the Redshift Evolution of the [{\rm{N}}\,{\rm{II}}]/Hα Ratio for Galaxy Redshift Surveys. *ApJ*, 855:132, Mar. 2018. doi: 10.3847/1538-4357/aab1fc.

G. Gavazzi, C. Bonfanti, G. Sanvito, A. Boselli, and M. Scodeggio. Spectrophotometry of Galaxies in the Virgo Cluster. I. The Star Formation History. *ApJ*, 576:135–151, Sept. 2002. doi: 10.1086/341730.

V. Gonzalez-Perez, C. G. Lacey, C. M. Baugh, C. D. P. Lagos, J. Helly, D. J. R. Campbell, and P. D. Mitchell. How sensitive are predicted galaxy luminosities to the choice of stellar population synthesis model? *MNRAS*, 439:264–283, Mar. 2014. doi: 10.1093/mnras/stt2410.

L. Guzzo, M. Scodeggio, B. Garilli, B. R. Granett, A. Fritz, U. Abbas, C. Adami, S. Arnouts, J. Bel, M. Bolzonella, D. Bottini, E. Branchini, A. Cappi, J. Coupon, O. Cucciati, I. Davidzon, G. De Lucia, S. de la Torre, P. Franzetti, M. Fumana, P. Hudelot, O. Ilbert, A. Iovino, J. Krywult, V. Le Brun, O. Le Fèvre, D. Maccagni, K. Małek, F. Marulli, H. J. McCracken, L. Paioro, J. A. Peacock, M. Polletta, A. Pollo, H. Schlagenhaufer, L. A. M. Tasca, R. Tojeiro, D. Vergani, G. Zamorani, A. Zanichelli, A. Burden, C. Di Porto, A. Marchetti, C. Marinoni, Y. Mellier, L. Moscardini, R. C. Nichol, W. J. Percival, S. Phleps, and M. Wolk. The VIMOS Public Extragalactic Redshift Survey (VIPERS). An unprecedented view of galaxies and large-scale structure at $0.5 < z < 1.2$. *A&A*, 566: A108, June 2014. doi: 10.1051/0004-6361/201321489.

O. Ilbert, P. Capak, M. Salvato, H. Aussel, H. J. McCracken, D. B. Sanders, N. Scoville, J. Kartaltepe, S. Arnouts, E. Le Floc'h, B. Mobasher, Y. Taniguchi, F. Lamareille, A. Leauthaud, S. Sasaki, D. Thompson, M. Zamojski, G. Zamorani, S. Bardelli, M. Bolzonella, A. Bongiorno, M. Brusa, K. I. Caputi, C. M. Carollo, T. Contini, R. Cook, G. Coppa, O. Cucciati, S. de la Torre, L. de Ravel, P. Franzetti, B. Garilli, G. Hasinger, A. Iovino, P. Kampczyk, J.-P. Kneib, C. Knobel, K. Kovac, J. F. Le Borgne, V. Le Brun, O. Le Fèvre, S. Lilly, D. Looper, C. Maier, V. Mainieri, Y. Mellier, M. Mignoli, T. Murayama, R. Pellò, Y. Peng, E. Pérez-Montero, A. Renzini, E. Ricciardelli, D. Schiminovich, M. Scodeggio, Y. Shioya, J. Silverman, J. Surace, M. Tanaka, L. Tasca, L. Tresse, D. Vergani, and E. Zucca. Cosmos Photometric Redshifts with 30-Bands for 2-deg^2. *ApJ*, 690: 1236–1249, Jan. 2009. doi: 10.1088/0004-637X/690/2/1236.

L. J. Kewley, M. A. Dopita, R. S. Sutherland, C. A. Heisler, and J. Trevena. Theoretical Modeling of Starburst Galaxies. *ApJ*, 556:121–140, July 2001. doi: 10.1086/321545.

C. Scarlata, C. M. Carollo, S. Lilly, M. T. Sargent, R. Feldmann, P. Kampczyk, C. Porciani, A. Koekemoer, N. Scoville, J.-P. Kneib, A. Leauthaud, R. Massey, J. Rhodes, L. Tasca, P. Capak, C. Maier, H. J. McCracken, B. Mobasher, A. Renzini, Y. Taniguchi, D. Thompson, K. Sheth, M. Ajiki, H. Aussel, T. Murayama, D. B. Sanders, S. Sasaki, Y. Shioya, and M. Takahashi. COSMOS Morphological Classification with the Zurich Estimator of Structural Types (ZEST) and the Evolution Since z = 1 of the Luminosity Function of Early, Disk, and Irregular Galaxies. *ApJS*, 172:406–433, Sept. 2007. doi: 10.1086/516582.

M. Scodeggio, L. Guzzo, B. Garilli, B. R. Granett, M. Bolzonella, S. de la Torre, U. Abbas, C. Adami, S. Arnouts, D. Bottini, A. Cappi, J. Coupon, O. Cucciati, I. Davidzon, P. Franzetti, A. Fritz, A. Iovino, J. Krywult, V. Le Brun, O. Le Févre, D. Maccagni, K. Malek, A. Marchetti, F. Marulli, M. Polletta, A. Pollo, L. A. M. Tasca, R. Tojeiro, D. Vergani, A. Zanichelli, J. Bel, E. Branchini, G. De Lucia, O. Ilbert, H. J. McCracken, T. Moutard, J. A. Peacock, G. Zamorani, A. Burden, M. Fumana, E. Jullo, C. Marinoni, Y. Mellier, L. Moscardini, and W. J. Percival. The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2). *ArXiv e-print 1611.07048*, Nov. 2016.

L. Stothert, P. Norberg, and Baugh. Statistic in large galaxy redshift surveys. *PhD thesis, Durham University*, 1, Dec. 2018a.

L. Stothert, P. Norberg, C. M. Baugh, A. Alarcon, A. Amara, J. Carretero, F. J. Castander, M. Eriksen, E. Fernandez, P. Fosalba, J. Garcia-Bellido, E. Gaztanaga, H. Hoekstra, C. Padilla, A. Refregier, E. Sanchez, and L. Tortorelli. The PAU Survey: spectral features and galaxy clustering using simulated narrow-band photometry. *MNRAS*, 481:4221–4235, Dec. 2018b. doi: 10.1093/mnras/sty2491.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

S. Van Dongen. A cluster algorithm for graphs. *National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam*, Technical Report INS-R0010, May 2000.

M. Viola, M. Cacciato, M. Brouwer, K. Kuijken, H. Hoekstra, P. Norberg, A. S. G. Robotham, E. van Uitert, M. Alpaslan, I. K. Baldry, A. Choi, J. T. A. de Jong, S. P. Driver, T. Erben, A. Grado, A. W. Graham, C. Heymans, H. Hildebrandt, A. M. Hopkins, N. Irisarri, B. Joachimi, J. Loveday, L. Miller, R. Nakajima, P. Schneider, C. Sifón, and G. Verdoes Kleijn. Dark matter halo properties of GAMA galaxy groups from 100 square degrees of KiDS weak lensing data. *MNRAS*, 452:3529–3550, Oct. 2015. doi: 10.1093/mnras/stv1447.