

# 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 upcoming years, under the supervision of Peder Norberg and Carlton Baugh.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Short-term projects</b>	<b>2</b>
2.1	Two months team project with Procter & Gamble (P&G) . . . . .	2
2.2	The quenching of star-formation in VIPERS and GALFORM . . . . .	4
<b>3</b>	<b>Main project</b>	<b>7</b>
3.1	PAUS Validation . . . . .	7
3.2	PAUS observations . . . . .	7
<b>4</b>	<b>Future project</b>	<b>8</b>
4.1	Group finder . . . . .	8

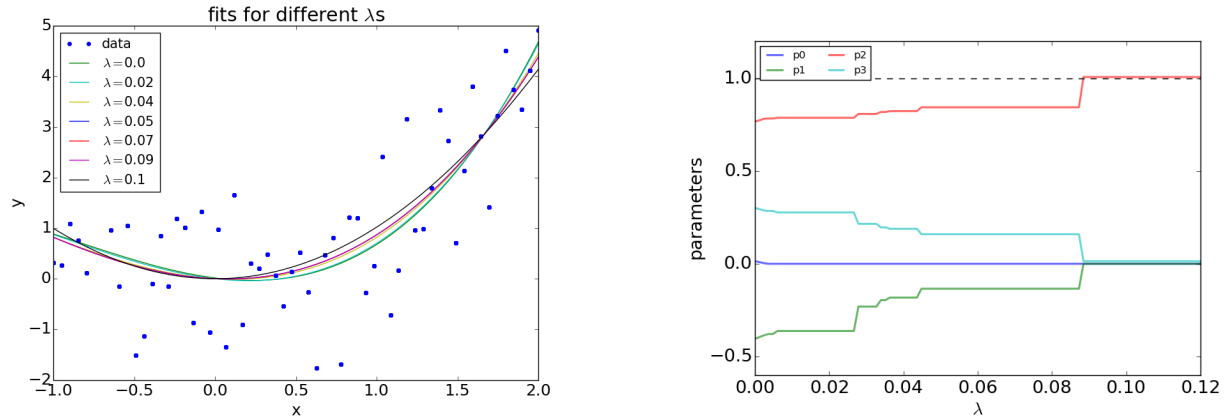
## 1 Introduction

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  $i_{AB} \approx 22.5$ .

Thanks to its depth and number density (tens of thousands of galaxies per  $\text{deg}^2$  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., 2018). Moreover, the study of galaxy groups will be further improved as 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 code, and here is where my work comes into play. In particular, Stothert et al. (2018) 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 is more suitable for a photometric redshift survey like this. Hence, the main purpose of my PhD project is to develop, test and finally apply this innovative MCL algorithm on the PAUS data.

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,



(a) Examples of modified LASSO fits for different  $\lambda$ s on 60 data points generated from an  $x^2$  function ( $f_{\text{true}}(x) = x^2$ ) with the addition of a gaussian noise with standard deviation 0.9. The fitting function is a third order polynomial.

(b) Trend of the parameters of the fitting function ( $f_{\text{fit}} = p_0 + p_1x + p_2x^2 + p_3x^3$ ) as a function of the  $\lambda$  parameter. Since we know  $f_{\text{true}}$ , 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  $\lambda \gtrsim 0.09$ .

Figure 1: Example of the performances of the modified LASSO in recovering a known function when some gaussian noise is added.

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 the last PAUS observational run which took place at the William Herschel Telescope (WHT) in La Palma.

In the rest of the report 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).

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:

$$f_i = f_i(x_1, \dots, x_{13}) \quad (1)$$

They made up the density function as a linear combination of those base functions and empirically they have chosen the combination that works better:

$$\text{density} = \sum_i^N a_i f_i(x_j) \quad (2)$$

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

$$\chi^2 = \sum_{k=0}^M \frac{(y_k - f_{\text{fit}}(x_k))^2}{M} \quad (3)$$

The minimisation of the  $\chi^2$ , 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 I modify  $f_{\text{fit}}$  adding a further base function, the  $\chi^2$  minimisation will try to use also this further parameter in order to obtain a better fit. The problem of finding the very best fit is that in another data-set it is very unlikely to be the best fit as well. This makes our result not robust. What we want instead is to realistically find which base functions the density rely on.

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. In that case (assuming that this ingredient does not appear in other base functions) 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.

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  $\chi^2$  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:

$$\text{LASSO} = \sum_{k=0}^M \frac{(y_k - f_{\text{fit}}(x_k))^2}{M} + \lambda \sum_{i=1}^N |a_i| \equiv \chi^2 + \text{Penalty} \quad (4)$$

where M is the number of data points and N is the number of base functions and  $\lambda$  is a free parameter that we can tune case by case (the choice of the perfect  $\lambda$  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. 4, a fit with a lot of bases function will result in a bigger penalty term, hence favouring a solution made by a linear combination with a lot of parameter set to zero.

However, we realised that the traditional LASSO as in Eq. 4 is not perfectly calibrated to work with all the possible scenarios as it minimises the sum of the absolute values of the parameters and not the number of them (small  $a_i$  is different from  $a_i = 0$ ). So we developed a new *modified LASSO* to deal with this problem. In particular we slightly changed the penalty function in this way:

$$\text{New Penalty} = \lambda \left[ \left( \sum_{i=0}^N |a_i| \right)^2 - \left( \sum_{i=0}^N a_i^2 \right) \right] \quad (5)$$

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 problem we faced was the fact 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:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \hat{y} = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \quad \hat{f}(\hat{x}) = \frac{f(x(\hat{x})) - f_{\min}(x(\hat{x}))}{f_{\max}(x(\hat{x})) - f_{\min}(x(\hat{x}))} \quad (6)$$

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  $\chi^2$  needs to be normalised in order to have the same order of magnitude of the penalty function. We managed it introducing the  $\epsilon^2$  parameter which is essentially the minimum value of the  $\chi^2$  (without the penalty) to normalise our modified LASSO to 1 when  $\lambda = 0$ :

$$\epsilon^2 = \min \left\{ \sum_{k=0}^M \frac{(y_k - f_{\text{fit}}(x_k))^2}{M} \right\} \quad (7)$$

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

$$\text{MODIFIED LASSO} = \frac{1}{\epsilon^2} \sum_{k=0}^M \frac{(y_k - f_{\text{fit}}(x_k))^2}{M} + \lambda \left[ \left( \sum_{i=0}^N |a_i| \right)^2 - \left( \sum_{i=0}^N a_i^2 \right) \right] \quad (8)$$

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. 8.

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  $\lambda$ s 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  $\lambda$  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  $\lambda$  using known case studies like the  $x^2$  one.

## 2.2 The quenching of star-formation in VIPERS and GALFORM

In September 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)<sup>1</sup>, a spectroscopic survey which covers 23.5 deg<sup>2</sup> 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 progressively disappearing. The speed at which it 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 deep of the survey (in the case of VIPERS, the cut is at  $i_{\text{AB}} < 22.5$ ).

To consistently identify the bright-edge, I have developed an algorithm which marks this feature when the number density in magnitude bins drops down a certain percentage of the most populated 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.

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<sup>2</sup> 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

<sup>1</sup>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/>.

<sup>2</sup>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).

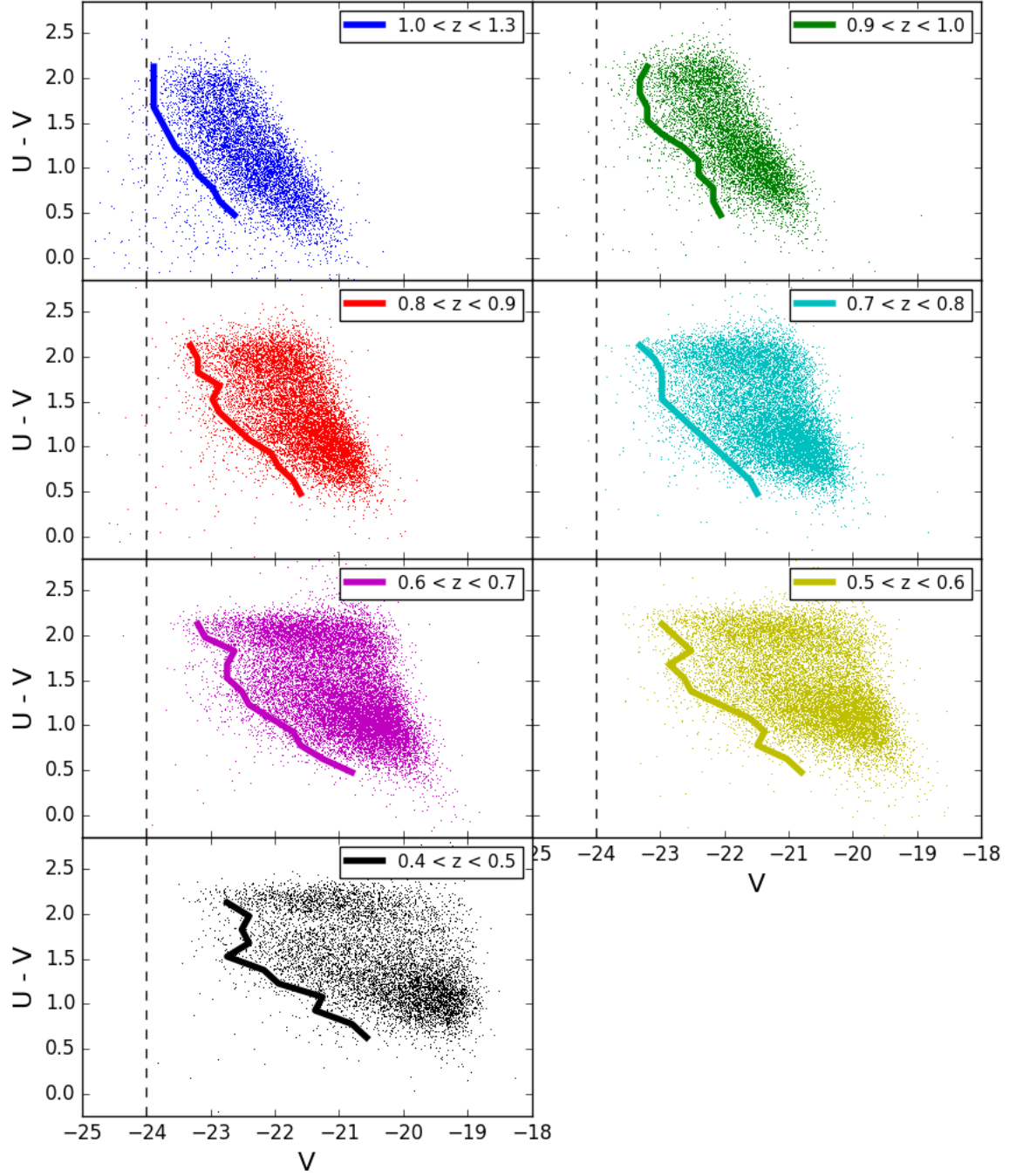


Figure 2: VIPERS colour-magnitude diagram in 7 different redshift bins. The thick line marks the bright-edge as defined by the drop in number counts to the 15% of the most populated bin.

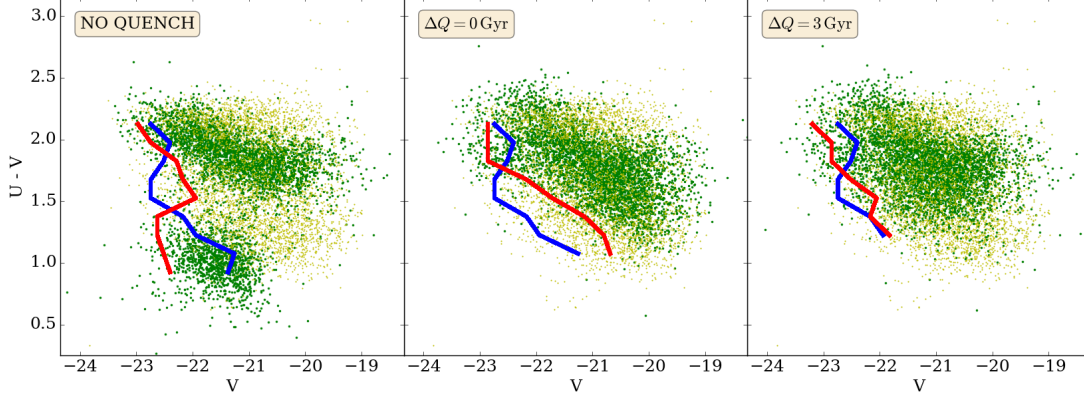


Figure 3: Results of the synthetic evolution of the colour-magnitude relation following PEGASE models with 3 different star formation histories. In the three panels, the yellow points are the same, i.e. the observed galaxies in the lower redshift bin ( $0.4 < z < 0.5$ ). The green points instead are 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.

Gavazzi et al. (2002), i.e. following Eq. 9,

$$\text{SFR}(t, \tau) = \frac{t}{\tau^2} \exp \left[ -\frac{t^2}{2\tau^2} \right] \quad (9)$$

where  $t$  is the cosmic time and  $\tau$  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<sup>3</sup> 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 to 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.

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). In Teruel workshop I received some good feedback about how to make these results more robust. One of the suggestion I am still working on is the comparison of these results with the ones obtained from semi-analytics models, 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.

<sup>3</sup>Instantaneously means a time of 100,000 years, i.e. the resolution of PEGASE models.

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.

[I am here](#)

## 3 Main project

### 3.1 PAUS Validation

Excluding lectures and CDT events, I invested the most of my first year PhD getting involved in the PAUS collaboration, in particular in the validation of the survey. PAUS is a photometric redshift survey with the main feature being the 40 contiguous narrow band filters. Since it is an ongoing survey, one of the task was to make sure that observations have been carried out with all of the filters. Because of the survey strategy and the format of the CCD, we don't observe each galaxy with all the filters in a row. Instead it is more convenient to move the CCD in a nearly spiral way (observing in dithers rather than single exposures) letting galaxies fall in the 8 different filters that are part of each of the 5 trays. That is why it becomes important to count the number of times that a galaxy have been observed in one particular filter. The number of filters is just one of many potential biases that might arise in an ongoing survey and my task is to check that instead there are no bias both at the observation stage and at the processing stage (nightly, memba and photo-z pipelines). I did all of this tests in the COSMOS area observed by PAUS, with the possibility of comparing PAUS photometric redshifts with COSMOS spectroscopic redshifts. Now the survey has collected enough data in the W3 field so that in the close future I am ready to do all the tests in this area as well. The others two fields that PAUS is planning to cover are W1 and W4. For those fields the advantage is the overlapping with the VIPERS survey (which has been the main focus of my master project) which offers spectroscopic redshift to be used for comparison with the photometric ones.

Before processing the W3 data, an important choice about the parent survey had to be done. Since we are dealing with forced photometry, what drove this choice was the half-light radius. In particular we needed to find a survey which overlaps with W3 and that has a similar definition of half-light radius as the one in the COSMOS survey (because we want consistency among all the PAUS fields). So, in this scenario I made both statistical and (where possible) object-by-object comparison among the COSMOS, CFHTLS and CFHTLenS surveys. I finally worked out that the CFHTLenS survey has a radius consistent with the COSMOS data and so suitable for the PAUS/W3 field.

One further analysis was aimed to improve the photo-z capability prediction through SED fitting. Specifically the current SED fitting procedure used by the PAUS team is based on a set of fixed emission-line ratio assumptions. A more physical scenario would allow the emission line ratios to vary following specific BPT diagrams. For this reason I have investigated in the literature various BPT diagrams for surveys at similar redshifts with the aim of building an empirical relations. The PAUS collaboration has approved the new scheme of emission line relations and they are going to use it in the next run of the photo-z pipeline. Of course, once they have done it, my next step is to check how much this variation contributes to the estimate of photometric redshifts, comparing the different productions.

In a very long term view, the experience gained in the PAU survey validation might be useful for the upcoming Multi-Object Optical and Near-infrared Spectrograph (MOONS) survey.

### 3.2 PAUS observations

In order to understand the nature of the data and to build experience within the survey, last March I took part to the PAUS observational run. The PAU survey is using the William Herschel Telescope (WHT), a 4.2 m diameter telescope based in La Palma (Spain) combined with a large field camera (PAUCam) built appositely for the survey. 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. The set up consisted of taking flat fields and bias images while the observation were aimed in covering part of the W3 and W1 fields accordingly to the spiral observing strategy. Galaxies observed with a seeing worse than 2.0 and transparency worse than 0.5 are flagged as “bad exposure” and the following night they are selected to be observed again. This observational experience was meant to be a training for the next PAUS observations. In fact, I have been assigned 4 nights for next December observational run.

## 4 Future project

### 4.1 Group finder

As mentioned in the Introduction, the main purpose of my PhD is to test and (if needed) to improve a group finder algorithm based on the Markov CLustering (MCL) technique. This technique seems to be more reliable (in terms of completeness, purity and variation of information, VI) than a simple Friend of Friend (FoF) technique which preferentially consider large structures as a single group even if they are barely connected. Until now the MCL has been tested only on mock catalogues and it looks promising for a real survey like PAUS. What makes different a real survey from a simulation is the fact that we are limited to work in the redshift space instead of the simulated real space and this has an impact in the MCL code. Running successfully this group finder on the PAUS data might provide some new insights relative to the clustering of galaxies.

## References

- 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. Tal-lada, N. Tonello, and L. Tortorelli. The PAU Survey: Early demonstration of photometric redshift performance in the COSMOS field. *ArXiv e-prints*, Sept. 2018.
- 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. Malek, F. Marulli, H. J. McCracken, L. Paioro, J. A. Peacock, M. Polletta, A. Pollo, H. Schlagenhauser, 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.
- 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 Fèvre, 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. 2018.
- 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>.