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

1 Introduction **1**

2 Short-term projects **1**
2.1 Two months team project with Procter & Gamble (P&G) 1
2.2 Workshop on Emission-Line Galaxies (Teruel - Spain) 3
2.3 GALFORM semi-analytic models 3

3 Long-term and future projects **3**
3.1 PAUS Validation 3
3.2 PAUS observations 4
3.3 Group finder 4

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 $i_{\rm AB} \approx 22.5$.

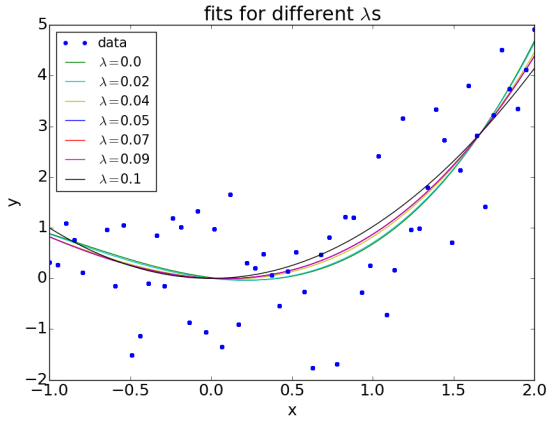
Thanks to its depth and number density (tens of thousands of galaxies per deg^2 with a median redshift of $z \sim 0.7$), PAUS will result in one of the most detailed studies of intermediate-scale cosmic structure ever undertaken. With these features, PAUS has the capability of probing the galaxy clustering in the transition from the linear to the non-linear regime (Stothert et al., 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.

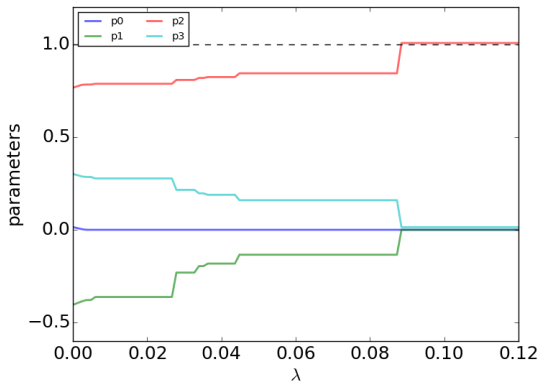
To do this, a deep understanding of the survey is needed. This has been the direction of my efforts in my first year. I have been part of the validation process of the survey and, attending weekly telecons, I have built the experience needed to deal with a group finder aimed to specifically work on photometric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



(a) Examples of modified LASSO fits for different λs 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 ($f_{\text{fit}} = p_0 + p_1 x + p_2 x^2 + p_3 x^3$) as a function of the λ parameter. Since we know f_{true}, we know that the right answer is $p_2 = 1$ 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.

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:

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

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

$$\text{density} = \sum_{\text{i}}^{\text{N}} \text{a}_\text{i} \text{f}_\text{i}(\text{x}_\text{j}) \tag{2}$$

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 χ^2:

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

The minimisation of the χ^2, although computationally expensive because of the high number of bases 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_{fit} adding a further base function, the χ^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. What we want instead is to realistically find which base functions the density rely on. A further advantage of eliminating some base functions f_i (i.e. setting the relative parameter $a_i = 0$) comes from the fact that not all of them depend on all the 13 ingredients x_j, 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 industrial processes.

A possible way to address this problem is LASSO $\Rightarrow$ I am here

That is why we have chosen to use the Least Absolute Shrinkage and Selection Operator (LASSO) 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 $\sum_{i=1}^{N} a_i$. This helps to discard the base functions that are not extremely relevant to the fit because a linear combination with a lot of parameter set to zero will have a smaller penalty term. However, we realised that the traditional LASSO is not perfectly calibrated to work with all the possible scenarios as it minimises the sum of the 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 that problem, however we agreed with the P&G team not to disclose the analytic form of the new penalty term. To test our new implementation of LASSO, we created increasingly difficult known functions (rising the number of variables and combining polynomials with exponentials terms) and adding various kinds of noise with different amplitudes we tried to recover the original functions. Since the penalty term depends on a free parameter λ, our case studies can be used in the future to train our algorithm to identify the best choice of λ.

2.2 Workshop on Emission-Line Galaxies (Teruel - Spain)

Last September I took part to the workshop named "Understanding Emission-line galaxies for the next generation of cosmological survey". In that occasion I presented, in a 15 minutes talk, 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 activity. For this project I have used data from the VIMOS Public Extragalactic Redshift Survey (VIPERS), a spectroscopic survey which covers 23.5 deg^2 at intermediate redshifts[1]. Specifically I studied the evolution of the bright-edge of the colour-magnitude developing an algorithm to quantify consistently this edge. Then, I have used the PEGASE 2 code to generate stellar population synthesis models to try and reproduce the evolution of the bright-edge, testing different star formation histories with and without a quenching event occurring. For this project I have an advance draft of a paper (attached to this document) which just need a comparison with simulations (see short term projects) in order to be complete.

2.3 GALFORM semi-analytic models

As mentioned in Section 2.2, the VIPERS draft paper need some comparison with simulations in order to be complete. One possible way to accomplish this is to use GALFORM semi-analytic models. I firstly used the Gonzalez-Perez 2014 (GP14) version to check that the evolution of the colour-magnitude is similar to the one observed in VIPERS. With GP14 we can also check the behaviour of the progenitor galaxies in the colour magnitude (still under investigation). One other test to check which quenching process is potentially more important in affecting the colour-magnitude diagram is to identify the population of galaxies close to the bright-edge. Using GP14, we 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 colour-magnitude. Instead AGN feedback (main quenching process for centrals in GALFORM) would play a crucial role in the evolution of the bright-edge in the colour-magnitude plane. For this reason, the only thing needed to complete this project is to run again GALFORM, switching on and off the AGN feedback and see which scenario better represents VIPERS observations.

3 Long-term and future projects

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 redhift 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 being 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

[1]The second (and complete) Public Data Release (PDR-2) is now available at `http://vipers.inaf.it/`.

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 Hershel 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.

3.3 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. Tallada, N. Tonello, and L. Tortorelli. The PAU Survey: Early demonstration of photometric redshift performance in the COSMOS field. *ArXiv e-prints*, Sept. 2018.

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