

A DATA-DRIVEN APPROACH TO FINDING YOUNG STELLAR POPULATIONS IN EARLY-TYPE GALAXIES FROM THEIR OPTICAL SPECTRA

Louisa Nolan¹, Ata Kabán², Somak Raychaudhury³, and Markus Harva⁴

¹School of Physics and Astronomy, University of Birmingham

²School of Computer Science, University of Birmingham

³School of Physics and Astronomy, University of Birmingham

⁴Helsinki University of Technology

ABSTRACT

We present the results of a novel application of Bayesian modelling techniques, which, although purely data driven, have a physically interpretable result, and will be useful as an efficient data mining tool. We base our pilot study on the UV-to-optical spectra (observed and synthetic) of early-type galaxies, with known star-formation histories. A probabilistic latent variable architecture is formulated, and a rigorous Bayesian methodology is employed for solving the inverse modelling problem from the available data. A powerful aspect of our formalism is that it allows us to recover a limited fraction of missing data due to incomplete spectral coverage, as well as to handle observational errors in a principled way. We find that our data-driven Bayesian modelling allows us to identify those early-types which contain a significant stellar population ≈ 1 Gyr old. We then apply our technique to the optical early-type galaxy spectra in the Sloan Digital Sky Survey. With this substantially larger data set (26,000 galaxies), our method is sufficiently sensitive to identify those early-types which have undergone significant star formation in the last 4 Gyr, which allows us to explore how star formation is triggered and extinguished in early-type galaxies. This method can be extended to other data sets, and is therefore a very useful tool for automatically discovering various interesting sub-classes of galaxies, via rapid analysis of their spectra.

Key words: Galaxies: elliptical and lenticular, cD; ; Galaxies: evolution; Galaxies: stellar content; Methods: data analysis; Virtual Observatory.

1. INTRODUCTION

Hierarchical structure formation models suggest that large objects form from the merging of smaller objects. Hence, in hierarchical structure formation, small objects form at the earliest epochs (e.g. Baugh, Cole & Frenk

1996; Kauffmann 1996; Baugh et al. 1998; Kauffmann & Charlot 1998b). However, one of the most important challenges remaining in understanding the formation and evolution of early-type galaxies (ETGs) is to reconcile this hierarchical model with the observational evidence that smaller present-day ETGs have, on average, younger stellar populations than more massive ETGs (e.g. Cowie et al. 1996, Treu et al. 2005). This suggests that mass assembly and star formation do not have a straightforward relationship. To understand this dichotomy, therefore, we need to know the star-formation history of ETGs, so that we may probe the relationship between mass assembly, environment, and the triggering and quenching of star formation.

Conventional methods of determining star-formation history utilise detailed physical modelling of stellar population spectra, which can be numerically expensive and model-dependent. Hence, with the millions of spectra now available from recent large surveys (e.g. 2dFGRS, SDSS), significant research has recently been focusing on the use of data-driven approaches to astrophysical analysis, especially multivariate statistical analyses (e.g. Connolly et al. 1995; Folkes, Lahav & Maddox 1996; Ronen, Aragon-Salamanca & Lahav 1999; Madgwick et al. 2003a,b). Here, we explore the application of component analyses (e.g. principal components analysis, PCA) to the data-driven identification of distinct stellar sub-populations in the spectra of galaxies. We use a Bayesian framework, and assess our assumptions carefully for each version of the analysis model that we consider. Our framework allows the modelling to be accomplished in a flexible manner, taking account of missing data values and measurement errors.

For the pilot project, we use the UV-optical spectra of 21 near-by early-type galaxies (Figure 1, see Nolan et al. 2006 for details), with star-formation histories known from applying our powerful 2-component stellar population modelling technique (Nolan et al. 2007). Thus, we can compare the parameters derived from the data-driven analysis with those obtained from detailed astrophysical

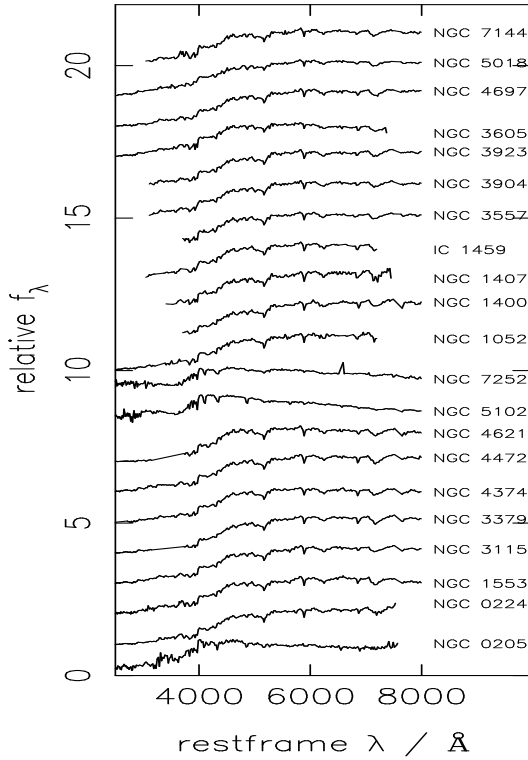


Figure 1. Rest-frame spectra of the 21 nearby early-type galaxies in our sample. The spectra are arbitrarily shifted along the flux axis for the sake of clarity.

modelling, which allows us to place a physical interpretation on our data-driven results.

§2 outlines the Bayesian model formulation. §3 presents the results of the pilot project, and §4 briefly discusses the application of our data analysis model to ETG spectra in the SDSS. §5 summarises our conclusions.

2. THE BAYESIAN MODEL

Our hypothesis is that each of the N observations of galaxy spectra (here, $N = 21$) can be described by a superposition of K underlying components, where $K < N$. These components are not observable by direct measurements, but only by some (as yet) unknown linear mapping. The K components are assumed to be statistically independent, and we include terms for both the noise (or stochastic component) of the model, and also for the measurement errors.

We consider 4 different factor models in the variational Bayesian (VB) framework. Three of these models were available in existing literature, and the 4th has been recently developed by two of the authors (Harva & Kabán 2005). We briefly list the main differences between the models below. Full details of these models are available in Nolan et al. (2006).

VB-PCA (Bishop 1999) is the baseline hypothesis. There

is no positivity constraint, so components can be negative, which is non-physical when describing fluxes, and the noise term is isotropic. **VB-FA** (Bishop 1999), the factor analysis model, relaxes the constraint of isotropic noise imposed on the VB-PCA model. All other aspects are identical with VB-PCA. **VB-PFA** (Miksin 2000), the positive factor analysis model, imposes the positivity constraint on both the hidden components and their mixing coefficients (weights). **VB-RFA** (Harva & Kabán 2005), the rectified factor analysis model, also imposes the positivity constraint on the components and mixing coefficients, but the model allows more flexibility regarding the shape of the factors compared with VB-PFA.

We have thus constructed four different flavours of model, all of which we could consider equally likely a priori. The Bayesian framework allows us to use the evidence (strictly, the negative log evidence bound) to decide which hypothesis is best supported by the data, and the most appropriate value of K . The VB-RFA model outperforms the other three. There is no significant improvement in the evidence going from $K = 2$ to $K = 3$ for this model, which suggests that the bulk of the spectra of ETGs can be reconstructed using only two components.

Figure 2 shows the shape of the components recovered from the four factor models, compared with the synthetic spectra (Jimenez et al. 2004) of a young (0.7 Gyr, Z_{\odot}) and old, metal-rich (10 Gyr, $2.5 Z_{\odot}$) stellar population. The similarity between the recovered components and the synthetic spectra is striking. In all cases, one component contains many of the features and the general shape of a young stellar population, and the other the features and shape of an old, metal-rich population. The VB-RFA components demonstrate the least amount of mixing between the young and old features.

Figure 3 shows the correlation of the age of the younger stellar population (as determined from two-component fitting to the observed spectra) with the weight of the first component (a_1 , the 'young' component) of the best-performing VB-RFA model of the observed spectra, with $K = 2$. It is clear that a high ($\gtrsim 1.0$) value of a_1 indicates the presence of a significant young (< 1 Gyr) stellar population, and that we have recovered components with a physical significance. In addition, predictions for the missing values provided by the VB-RFA model agree well those from the physical model. This is a strong indication that these two entirely independent approaches are well-matched. A full discussion of these results is presented in Nolan et al. (2006)

3. EARLY-TYPE GALAXIES WITH YOUNG STELLAR POPULATIONS IN THE SDSS

Following the success of the VB-RFA model in identifying young stellar components in near-by ETGs with UV-optical spectra, we then run the analysis on $> 13,000$ optical spectra of massive ETGs in the SDSS at redshift

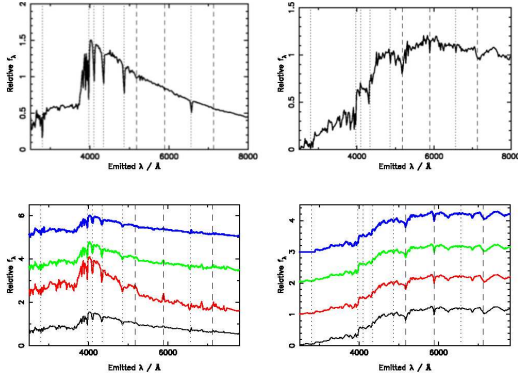


Figure 2. **Top:** Synthetic stellar population spectra (Jimenez et al. 2004), at 0.7 Gyr, Z_{\odot} (left) and 10 Gyr, $2.5 Z_{\odot}$ (right). The dotted lines mark some of the absorption features in the spectrum which are typically strong in young stellar populations, and the dashed lines mark those which are typically strong in old, metal-rich stellar populations. From left to right, the absorption line species are: MgII (2799 Å), H ϵ (3970 Å), H δ (4102 Å), H γ (4340 Å), H β (4861 Å), Mgb (5175 Å), NaD (5893 Å), H α (6563 Å), TiO (7126 Å). **Bottom:** The two components found from the variational Bayesian analyses, with the number of components $K = 2$, of the observed early-type galaxy spectra. The methods are, from bottom to top: VB-FA, VB-PFA, VB-PCA, VB-RFA. The recovered spectra are convincingly disentangled into one component (bottom left) with young stellar population features (dotted lines) and shape, and a second (bottom right) with the features (dashed lines) and shape of an old, high-metallicity stellar population. The young / old populations separation is most convincing for the VB-RFA components.

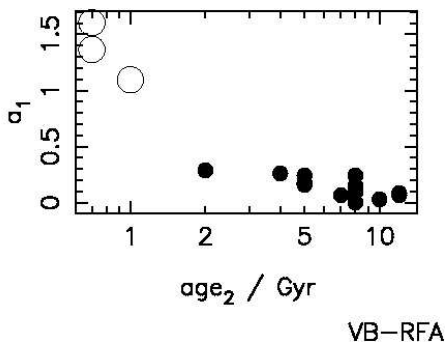


Figure 3. The correlation of the age of the younger stellar population (as determined from two-component fitting to the observed spectra) with the weight of the first component (a_1) of the VB-RFA linear basis transformation analyses of the observed spectra, with $K = 2$. The open circles are for those galaxies with a secondary population of age less than 1 Gyr. A high value of a_1 clearly corresponds to a young (≈ 1 Gyr) stellar population.

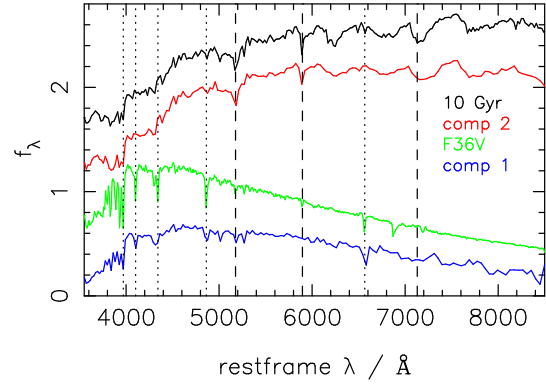


Figure 4. From **top to bottom:** a 10 Gyr, $2.5 Z_{\odot}$ single stellar population model (Bruzual & Charlot 2003, **black**); component 2 from our analysis (**red**); a super-solar metallicity F star spectrum (Santos et al. 1995 **green**); component 1 from our analysis (**blue**). The dotted lines mark some of the absorption features in the spectra which are typically strong in young stellar populations, and the dashed lines mark some of the absorption features which are typically strong in old, metal-rich stellar populations. The dotted and dashed lines are as in figure 2. It can clearly be seen that the first component (**blue, bottom**) represents a younger stellar contribution (**green, second from bottom**), whereas the second component (**red, second from top**) represents a mature stellar population (**black, top, Nolan, Raychaudhury & Kabán 2007**).

0.06–0.14. Figure 4 shows the recovered components. Again, one component represents an old, metal-rich population, and the other a young stellar population. With this data set, the ‘young’ component looks more like the spectrum of an F-star, at ~ 4 Gyr, rather than the younger-looking component recovered from the analysis of the non-statistical sample of 21 galaxies in the pilot project. Hence, we can construct a population of ETGs which contain a significant ($\gtrsim 10\%$) young ($\lesssim 4$ Gyr) stellar component, by choosing those ETGs for which the weight of the first, ‘young’ component (a_1) is greater than that of the second component (a_2). Details of the sample selection and a detailed discussion of the components and identification of ETGs with young stellar populations can be found in Nolan, Raychaudhury & Kabán (2007).

We can use our large (> 2000) sample of ETGs with young stellar populations to explore the physical mechanisms which trigger starbursts in ETGs. Figure 5 shows the distribution of the local galaxy surface density, Σ_5 , for these galaxies, compared with the distribution of ETGs which have formed at least 10 % (by mass) of their stars in a merger-induced starburst, from the semi-analytic simulations of Baugh et al. (2005), at $z = 0.089$. The similarity of the two distributions suggest a galaxy-galaxy merger / interaction origin for our ETGs with young stellar populations. The peak at $\Sigma_5 \sim 0.1 \text{ Mpc}^2$ may indicate the optimum local galaxy surface density for galaxy-galaxy mergers and interactions, representing a trade-off between sufficient density for galaxy-galaxy

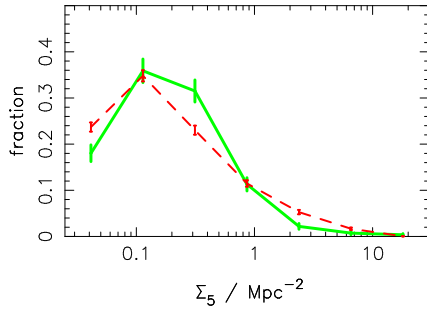


Figure 5. Normalised distributions of the local galaxy surface density, Σ_5 , of ETGs with young ($\lesssim 4$ Gyr) stellar populations (red, dashed) compared with early-types which have formed at least 10% (by mass) of their stars in a merger-induced starburst, from the semi-analytic simulations of Baugh et al. (2005), at $z = 0.089$ (green, solid). The populations are normalised so that the area under each line represents 100% of the galaxies in that sample (Nolan, Benson & Raychaudhury 2007, in preparation).

encounters, and slow enough relative velocities that these encounters lead to a meaningful interaction (Nolan, Raychaudhury & Kabán 2007; Nolan, Benson & Raychaudhury 2007, in preparation).

4. CONCLUSIONS

Using rigorous Bayesian modelling, we have successfully developed a time-efficient method for the analysis of the spectra of early-type galaxies. Our formulation incorporates observational errors, and allows us to predict missing data values. By comparing the results of this analysis with the results of our unique two-component synthetic stellar population fitting to the long-baseline data, we have shown that the independent components analysis is physically interpretable. Our variational Bayesian rectified factor analysis performs best, both in terms of the evidence, and the physical interpretation, of the various analyses we tested.

Two linear components are sufficient to describe the bulk of the stellar content of early-type galaxies. One of the components represents a young stellar population, and the other, an old, metal-rich population. The relative contribution of these two components allows us to robustly identify which early-type galaxies host significant young stellar populations.

We have used our Bayesian technique to explore early-type galaxy spectra in the Sloan Digital Sky Survey, and have thus identified a class of early-types with young stellar populations. Studying the relationship between this galaxy population and local galaxy surface density allows us to investigate the physical mechanisms which trigger and/or quench star formation in early-types. Preliminary results suggest that the dominant mechanism in early-types is galaxy-galaxy interactions and mergers.

We expect that this technique will be useful in uncovering other interesting classes of galaxies, for example starburst galaxies, and intend to make this powerful tool available for the Virtual Observatory.

REFERENCES

- Baugh, C. M., Lacey, C. G., Frenk, C. S., Granato, G. L., Silva, L., Bressan, A., Benson, A. J., & Cole, S., 2005, MNRAS, 356, 1191
- Baugh C. M., Cole S., Frenk C. S., 1996, MNRAS, 283, 1361
- Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 1998, ApJ, 498, 504
- Bishop C. M., 1999, in Proc. 9th Int. Conf. on Artificial Neural Networks (ICANN), Vol. 1. IEEE, p. 509
- Connolly A. J., Szalay A. S., Bershadsky M. A., Kinney A. L., Calzetti D., 1995, AJ, 110, 1071
- Cowie L. L., Songaila A., Hu E. M., Cohen J. G., 1996, AJ, 112, 839C
- Folkes S.R., Lahav O., Maddox S.J., 1996, MNRAS, 383, 651
- Harva M., Kabán A., 2005, in Proc. IEEE Int. Conf. Neural Networks, p. 185
- Jimenez R., MacDonald J., Dunlop J. S., Padoan P., Peacock J. A., 2004, MNRAS, 349, 240
- Kauffmann G., 1996, MNRAS, 281, 487
- Kauffmann G., Charlot S., 1998, MNRAS, 297, L23
- Madgwick D. S., Somerville R., Lahav O., Ellis R., 2003a, MNRAS, 343, 871M
- Madgwick D. S. et al., 2003b, ApJ, 599, 997
- Miskin J., 2000, PhD thesis, Univ. Cambridge
- Nolan L. A., Harva M. O., Kabán A., Raychaudhury S., 2006, MNRAS, 366, 321
- Nolan, L. A., Dunlop, J. S., Panter, B., Jimenez, R., Heavens, A., & Smith, G., 2007, MNRAS, 375, 371
- Nolan, L. A., Raychaudhury, S., & Kabán, A., 2007, MNRAS, 375, 381
- Ronen S., Aragon-Salamanca A., Lahav O., 1999, MNRAS, 303, 284
- Treu T., Ellis R. S., Liao T. X., van Dokkum P. G., 2005, ApJ, 622, L5