The Spectral Energy Distribution of Galaxies
Proceedings IAU Symposium No. 284, 2011
R.J. Tuffs & C.C. Popescu, eds.

doi:10.1017/S174392131200868X

Beyond model fitting SEDs

Ignacio Ferreras

Mullard Space Science Laboratory, University College London
Holmbury St Mary, Dorking, Surrey RH5 6NT, UK

email: ferreras@star.ucl.ac.uk

Abstract. Extracting star formation histories from spectra is a process plagued by numerous degeneracies among the parameters that contribute to the definition of the underlying stellar populations. Traditional approaches to overcome such degeneracies involve carefully defined line strength or spectral fitting procedures. However, all these methods rely on comparisons with population synthesis models. This paper illustrates alternative approaches based on the statistical properties of the information that can be extracted from uniformly selected samples of observed spectra, without any prior reference to modelling. Such methods are more useful with large datasets, such as surveys, where the information from thousands of spectra can be exploited to classify galaxies. An illustrative example is presented on the classification of early-type galaxies with optical spectra from the Sloan Digital Sky Survey.

Keywords. methods: statistical, techniques: spectroscopic, galaxies: stellar content

1. The Goal

In order to understand the process of structure formation and evolution in the Universe, it is essential to solve the problem of galaxy formation. *Ab initio* models of galaxy formation come up against the many complexities of the baryon physics transforming gas into stars. Hence, it is desirable to determine from the observations the distribution in age and composition of the stellar populations in galaxies, enabling us to backtrack their formation histories. The spectral energy distribution (SED) of galaxies encodes a treasure trove of information about the underlying stellar populations. It consists of a linear superposition of stellar spectra from a complex distribution of ages, metallicities, and dust. In principle, one could derive the star formation and chemical enrichment histories from an SED. However, the inherent degeneracies prevent us from reaching that goal using straightforward techniques.

2. The Standard Method

The most used approach to the extraction of star formation histories from spectra involves model fitting techniques, whereby the observations are compared with a grid of synthetic stellar populations, parameterised by a distribution of ages, metallicities and dust. A figure of merit is defined, often from a χ^2 function, from which one defines a likelihood, that can be combined with priors in a Bayesian way. The search for the best fit and the uncertainties in the parameters can come in many flavours, such as searches over a large library of models (Gallazzi *et al.* 2005); Metropolis-based algorithms (Cid Fernandes *et al.* 2004); search on data-compressed models (Panter *et al.* 2003), or least-squares solutions (Ockvirk *et al.* 2006). However, all these methods rely on the accuracy of the synthetic models to explain all the subtleties of galaxy spectra.

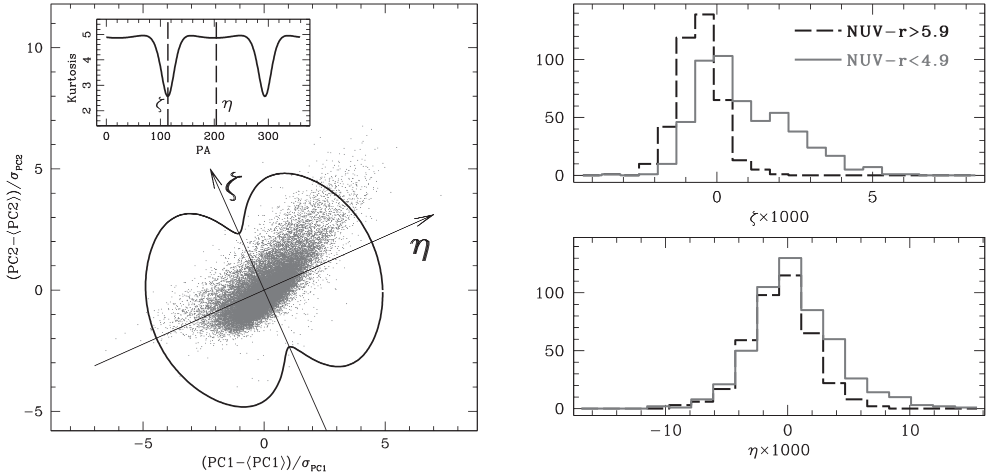

Figure 1. Independent components can be extracted from a PCA decomposition of the data. *Left:* The first two principal components from a sample of SDSS spectra of early-type galaxies is projected in different orientations, and the kurtosis of the projections (shown as a contour in the main panel and as a function of position angle in the inset) shows the preferred orientation to separate the signals from PC1 and PC2. *Right:* The distribution of the projected components (η an ζ) are shown for a subsample where GALEX photometry is available. The histograms correspond to NUV-bright (solid grey) and NUV-faint galaxies (dashed black). Notice only optical spectra are used for the definition of the components.

3. Beyond Model Fitting SEDs

Large spectroscopic surveys, such as the Sloan Digital Sky Survey (SDSS, York *et al.* 2000) have opened up the possibility of extracting information on a purely statistical basis. One could consider a sample of galaxy spectra as a set of multi-dimensional vectors, converting this problem into a multi-variate analysis method. For instance, these vectors can be expressed as linear combinations of a reduced set of “basis vectors” that constitute fundamental stellar populations from which one can disentangle the observations. Such methods have been applied to perform general classifications of spectra from surveys such as 2dFGRS (Madgwick *et al.* 2003) or SDSS (Yip *et al.* 2004). They can also be used to improve data reduction, such as in the removal of night sky lines (Wild & Hewett 2005). More relevant to this conference, these techniques can be exploited to disentangle the information from the stellar populations (see e.g. Ronen *et al.* 1999, Ferreras *et al.* 2006).

In this paper, we present and extend recent results following this approach, focused on a volume-limited sample of early-type galaxies (ETGs) from SDSS (see e.g. Rogers *et al.* 2007 and Rogers *et al.* 2010). We emphasize that the method does not rely on information from any model in order to extract information about the underlying stellar populations.

3.1. *Independent Component Analysis*

In Fig. 1, we illustrate the technique of Independent Component Analysis (Hyvärinen *et al.* 2001), where the observations are assumed to be created from superpositions of spectra that are statistically independent. One could naively relate these independent components to the populations that define the star formation histories of the galaxies under scrutiny. In the left panel of Fig. 1, we show galaxy spectra as projections on to the first and second components obtained from Principal Component Analysis (see Rogers *et al.* 2007, for details). Principal components are decorrelated signals extracted from

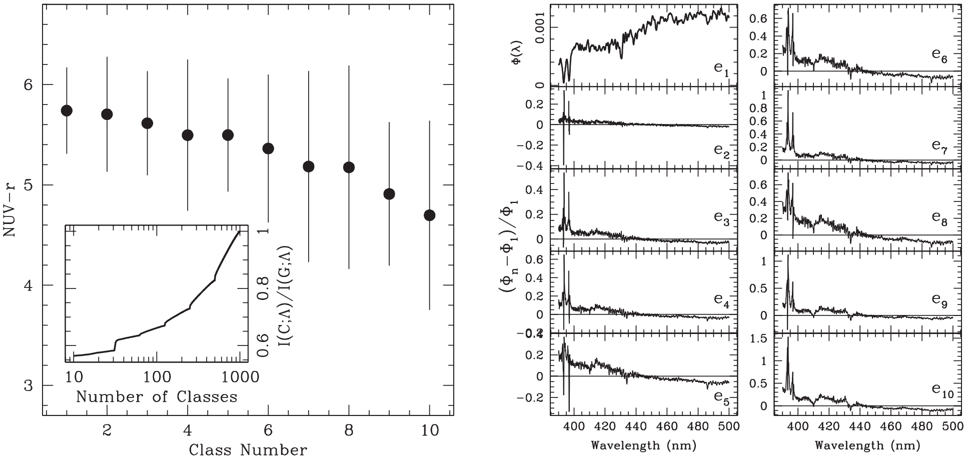

Figure 2. A sample of 1,000 galaxies from the sample of SDSS early-type galaxies, with available GALEX photometry, is presented to an Information Bottleneck sorter (see text for details). *Left:* The inset shows the decrease in information as the number of classes is reduced from the trivial set (as many classes as galaxies) to a target of 10 classes. The main panel shows the NUV–r colour of each of the 10 classes. *Right:* Spectral information of the 10 classes. The average spectrum of the first class is shown in the top-left panel. For the other 9 classes, we show the relative change with respect to the first one.

the original spectra. In order to go beyond a simple decorrelation, one can explore the non-gaussianity of the distribution: if we assume that the observations are superpositions of more fundamental spectra, according to the central limit theorem, the latter should be less gaussian. We show as a solid line the contour of the kurtosis of the (PC1,PC2) projections, as a function of the projeciton angle. There is a preferred set of directions where the kurtosis reach extrema (see inset). These directions define a new pair of "co-ordinates" for each spectra, which are the projections on to the new axes, given here as η and ζ. It is possible to interpret the meaning of these new components when we compare the distribution of η and ζ as a function of NUV–r colour, combining SDSS and GALEX photometry (rightmost panels of Fig. 1). Out of the two components, ζ is found to correlate strongly with NUV excess, a sign of recent star formation (Kaviraj *et al.* 2007).

3.2. *The Information Bottleneck*

In Fig. 2, we present another approach based on the concept of mutual information between sets of data. The Information Bottleneck technique (IB, Slonim *et al.* 2001) focuses at a classification of input data based on an algorithm that aims at minimising the complexity of the set (i.e. the number of classes), while minimising the information lost by the classification. For this example, we start with 1,000 early-type galaxies from the above mentioned sample, and classify them into ten bins, according to the IB method. In the inset of the left panel, the mutual information between spectra and classes is plotted as a function of the total number of classes in each step, from full information at the top-right corner, where we start with the trivial choice of having as many classes as galaxies, to the bottom-left corner, where only ten classes are used to describe the entire set, and 60% of the initial information retained. On the left of Fig. 2, the distributon of NUV–r colours is shown for each class, showing that there is a clear trend in the classification, with respect to NUV flux, meaning that, to first order, the most important factor driving the differences in the spectra of massive ETGs is the presence of recent star formation.

https://doi.org/10.1017/S174392131200868X Published online by Cambridge University Press

Note that the classification of the spectra uses the optical range ($\lambda = 3800$-7000Å), where the presence of recent star formation, as detected by an NUV excess, leaves a very weak trend, which is hard to detect when using traditional methods of spectral fitting. The panels on the right show the average spectra of all ten classes (from class 2 to 10, we show the fractional change with respect to the spectrum of the first class). The excess in blue light as one progresses down the class number is evident. Notice also the bumps in the higher order classes, at the positions of the Hδ ($\lambda = 4102$Å) and Hγ ($\lambda = 4340$Å) Balmer absorption lines, which is also a characteristic feature of age.

4. Epilogue

The methods described here are nothing but a tip in the iceberg of possible techniques explored in the field of multivariate analysis. Machine learning methods such as neural networks (Abdalla *et al.* 2008) or support vector machines (Tsalmantza *et al.* 2009) can be applied to galaxy data, when prior information about the classes is robust. Clustering methods have also been tentatively applied to galaxy spectra (Sánchez Almeida *et al.* 2011). However, a proper disentanglement of the underlying stellar populations still remains an open problem.

References

Abdalla, F. B., Mateus, A., Santos, W. A., Sodrè, L., Ferreras, I., & Lahav, O. 2008, *MNRAS*, 387, 945

Cid Fernandes R., Gu Q., Melnick J., Terlevich E., Terlevich R., Kunth D., Rodrigues Lacerda R., & Joguet B. 2004, *MNRAS*, 355, 273

Ferreras, I., Pasquali, A., de Carvalho, R. R., de la Rosa, I. G., & Lahav, O. 2006, *MNRAS*, 370, 828

Hyvärinen, A., Karhunen, J., & Oja, E., *Independent Component Analysis*, 2001, Wiley

Gallazzi, A., Charlot, S., Brinchmann, J., White, S. D. M., & Tremonti, C. A. 2005, *MNRAS*, 362, 41

Rogers, B., Ferreras, I., Lahav, O., Bernardi, M., Kaviraj, S., & Yi, S. K. 2007, *MNRAS*, 382, 750

Rogers, B., Ferreras, I., Pasquali, A., Bernardi, M., Lahav, O., & Kaviraj, S. 2010, *MNRAS*, 405, 329

Kaviraj, S. *et al.* 2007, *ApJ Supp. Ser.*, 173, 619

Madgwick, D., Somerville, R., Lahav, O., & Ellis, R. 2003, *MNRAS*, 343, 871

Ocvirk, P., Pichon, C., Lançon, A., & Thiébaut, E. 2006, *MNRAS*, 365, 46

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

Ronen, S., Aragón-Salamanca, A., & Lahav, O. 1999, *MNRAS*, 303, 284

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & Huertas-Company, M., 2011, *ApJ*, 735, 125

Slonim, N., Somerville, R., Tishby, N., & Lahav, O. 2001, *MNRAS*, 323, 270

Tsalmantza, P. *et al.* 2009, *A& A*, 504, 1071

Wild, V. & Hewett, P. C. 2005, *MNRAS*, 358, 1083

Yip, C. W. *et al.* 2004, *AJ*, 128, 2603

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

https://doi.org/10.1017/S174392131200868X Published online by Cambridge University Press