Composite Spectral Energy Distributions and Bayesian Machine Learning for Spectral Data

Ming-Feng Ho[1]

[1]Department of Physics and Astronomy, University of California, Riverside; email: mho026@ucr.edu

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–15

https://doi.org/10.1146/((please add article doi))

Keywords

galaxies: evolution, methods: data analysis, methods: Bayesian non-parametric

Abstract

Composite SEDs and Bayesian non-parametric clustering. Photometry and medium bands: surveys Spectral Energy Distributions: fitting template, FAST, EAZY Composite SEDs: evolution from grouping methods Bayesian non-parametric on functional data: 1. Dirichlet Processes for clustering 2. Gaussian Processes on Spectral data 3. Clustering on functional data

Contents

1. INTRODUCTION

One fundamental dilemma in observational astronomy is that: do we want to take a spectrum or take a multi-wavelength photometric observation? Taking a spectrum will give us more detailed information about the properties of the object, e.g., emission lines and absorption lines. However, the exposure time of taking a spectrum is much longer than taking a multi-wavelength photometric observation.

Using multi-wavelength photometry, on the other hand, is less expensive than taking a spectrum. The other advantage of multi-wavelength observations is that we are able to reach a deeper redshift and a wider spatial range. However, the trade-off of multi-wavelength observations is that we lose the resolution on the SEDs of data.

The dilemma here is similar to the problem of **exploration or exploitation** situation[1]: whether you want to focus on taking accurate observations on small number of objects or your want to explore a wider and deeper area but lose the accuracy? Another analogy to the exploration or exploitation problem is the "Battleship" game[2]. In the Battleship game, we have to destroy our enemy's ships by guessing the ship locations with a few trail shootings. We have to make an intelligent choice to balance exploration and exploitation to win. In common cases, we would start with exploring the whole area; after gaining some low-resolution knowledge, we could start to focus on the plausible regions to find our enemy's hidden ships.

The game played by astronomers here is similar: should we choose to explore the space with a lower resolution by multi-wavelength photometry or exploit on a limited number of objects for taking expensive spectra? To make an intelligent decision, we need to know more. Since the budget is limited and the number of spectra we can take is also limited, gaining information from multi-wavelength photometric data before taking spectra is important.

Astronomers developed techniques including UVJ color-color diagram and SED fitting to gain more information with solely photometric data in order to make optimal decisions. The color-color diagram classifies galaxies into quiescent or star-forming types; SED fitting estimates the properties of galaxies and redshifts (e.g., EAZY (Brammer, van Dokkum & Coppi 2008)) with the help of spectral energy modeling and synthetic templates built by matrix factorization (Blanton & Roweis 2007).

[1] See Roman's slide (Garnett 2018) `https://www.cse.wustl.edu/~jain/cse591-18/ftp/garnett591.pdf` or `https://en.wikipedia.org/wiki/Active_learning_(machine_learning)`

[2] `https://en.wikipedia.org/wiki/Battleship_(game)`

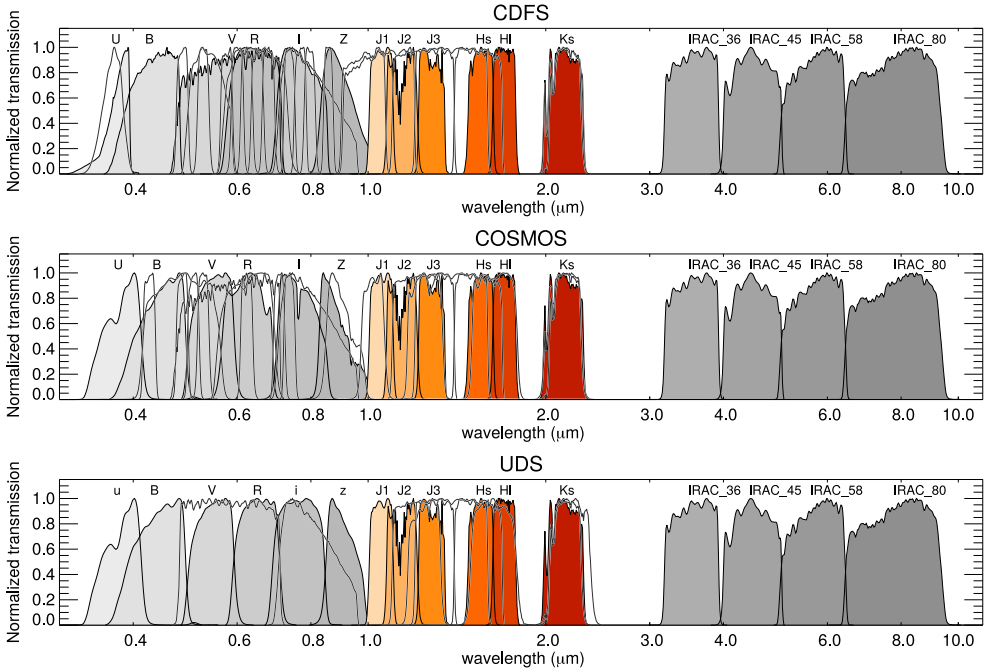

Figure 1

`Fourstar` medium bands (J1, J2, J3, Hs, Hl, Ks) in Straatman et al. (2016) (highlighted as red and yellow). The choice of medium bands would increase the resolution of sampling on the SED.

One thing we can ask is: can we gain more knowledge about our targets beyond the UVJ diagram and SED fitting? Composite SED technique (Kriek et al. 2011; Forrest et al. 2018) provides an empirical way to cluster multi-wavelength targets together based on the shape of SED fittings.

2. GALAXY EVOLUTION IN TERMS OF COMPOSITE SPECTRAL ENERGY DISTRIBUTIONS (seds)

2.1. Medium-Band Photometry

This is dummy text.

2.2. Composite Spectral Energy Distributions (seds)

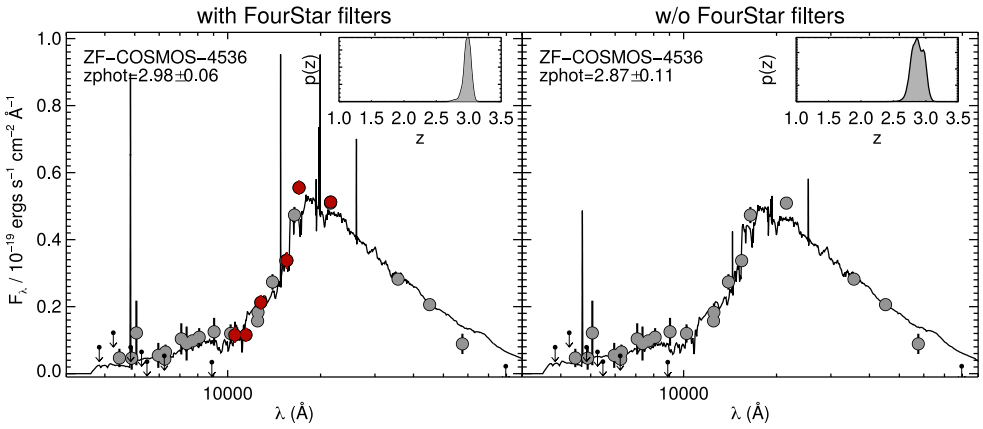

Figure 2

Sampling SED with or without **Fourstar** medium-band filters. With the help of **Fourstar** filters (red), the posterior distribution of photometric redshift (with SED fitting) is way more accurate.

2.2.1. Grouping Method.

Data: Multi-wavelength photometry of galaxies within some redshift range
Result: Composite SEDs of different galaxy types
`// Grab your data: fluxes and wavelengths`
$\mathbf{F}, \Lambda \leftarrow$ Multi-photometry of galaxies
`// Bin: 22 rest-frame filters with equal widths between 1 226 Å< λ < 49 580 Å`
$Bin(\lambda) = \text{map} : \lambda \rightarrow \lambda[\log_{10} 1226 : \frac{1}{22}(\log_{10} 49580 - \log_{10} 1226)i]$
`// Loop: modify each galaxy multi-wavelength flux`
for **f** *in* $\mathbf{F}$, λ *in* $\mathbf{\Lambda}$ **do**
  `// Fit: with EAZY to get continous flux function`
  $f_{\text{SED}}, z_{photo} \leftarrow EAZY(\mathbf{d})$
  `// De-redshift: using photo-z`
  $\lambda_{rest} \leftarrow \lambda/(1 + z_{photo})$
  `// Bin: update discrete flux vector with rest-frame filters`
  $\mathbf{F}[i, :] \leftarrow f_{\text{SED}}[Bin(\lambda_{rest})]$
end
`// Similarity: use squared error to calculate similarity`
$\mathbf{b} \leftarrow zeros(N, N)$
for $\mathbf{f}_i$ *in* $\mathbf{F}$ **do**
  for $\mathbf{f}_j$ *in* $\mathbf{F}$ *if* $j > i$ **do**
    `// calculate scale factor; ∘: Hadamard (elementwise) product`
    $a_{12} \leftarrow sum(\mathbf{f}_i \circ \mathbf{f}_j)/sum(\mathbf{f}_j^2)$
    `// calculate similarity`
    $\mathbf{b}_{ij} \leftarrow (sum((\mathbf{f}_i - a_{12}\mathbf{f}_j)^2)/sum(\mathbf{f}_i^2))^{1/2}$
  end
end
`// Get the analogues from samples with b < 0.05`
ind $\leftarrow argsort(sum(\mathbf{b} < 0.05).axis(1))$
$Composite_{\text{SED}} \leftarrow []$
for i *in* ind **do**
  `// find analogues`
  $i_a \leftarrow find(\text{b}_i < 0.05)$
  if $length(i_a + i) < 19$ **then**
    break
  end
  $Composite_{\text{SED}}$.append($[i_a + i]$)
end

Algorithm 1: Grouping galaxies together based on SED shape

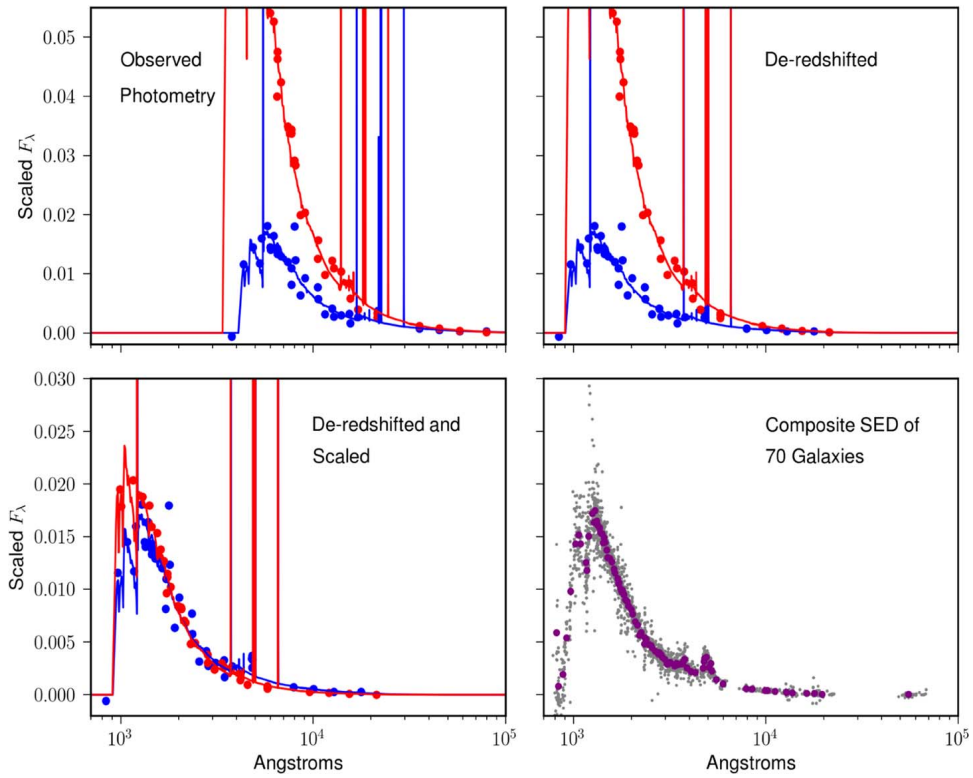

Figure 3

Grouping method for composite SED in Forrest et al. (2018). First, we take the observed multi-wavelength photometry. Second, de-redshift the wavelengths to rest-frame. Next, scale the flux using a_{12} (see Algorithm 1). In the final panel, we get the composite SED based on the similarity metric $b_{12} < 0.05$.

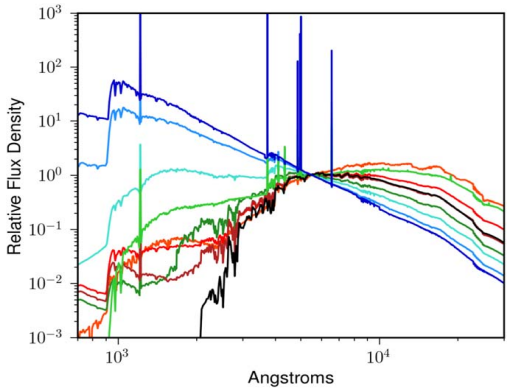

Figure 4

EAZY (Brammer, van Dokkum & Coppi 2008) templates were used in fitting the SED of galaxies in Forrest et al. (2018).

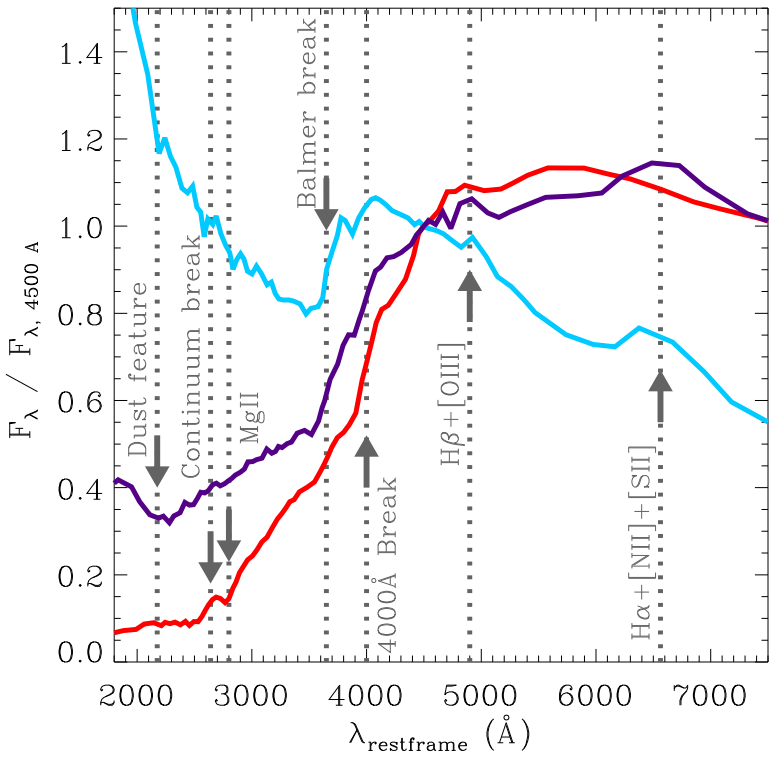

Figure 5

Three types of galaxies (red: quiescent; purple: dusty star-forming; blue: star-forming) identified in Kriek et al. (2011) using composite SED. Several spectroscopic features are labeled in the plot.

2.2.2. Classification of Galaxy Types.

3. BAYESIAN MACHINE LEARNING FOR CLUSTERING AND MODELING SPECTRAL DATA

Bayesian machine learning is a branch of machine learning which aims to solve machine learning problems in a Bayesian perspective. Instead of optimizing the parameters of interest from data using an empirical loss function (e.g., a least-squared function), Bayesian methods build generative models to randomly sample data from parameters and try to maximize the likelihood between observed data and hidden parameters (Barber 2012).

The difference between Bayesian statistics and Bayesian "machine learning" is that Bayesian "machine learning" is trying to approximate *non-linear* functions (Bishop & Tip-

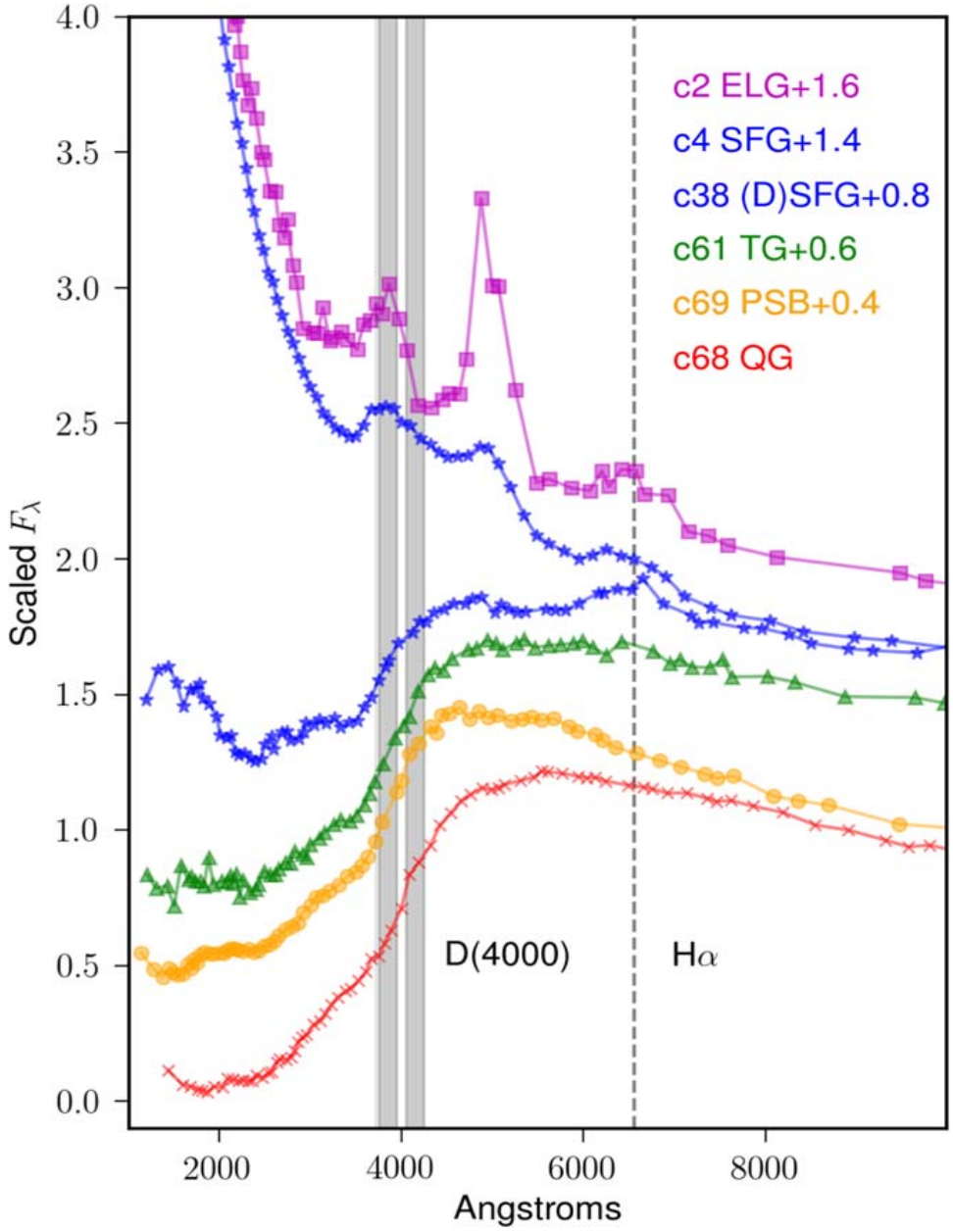

Figure 6

Seven galaxy types identified by visual inspecting on two features: 4 000 break $D(4000)$ (an indicator of age) and H_α equivalent width (a probe of star formation activity).

ping 2003). After the publish of Rasmussen & Williams (2005), learning unknown complicated functions from observed data using *Gaussian processes* (GP) became popular.

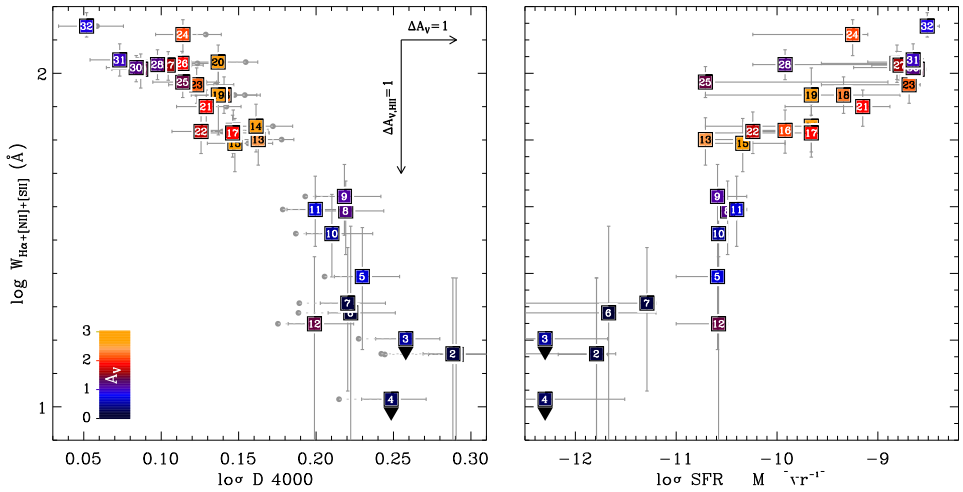

Figure 7

Left panel: correlation between H_α equivalent width and $D(4000)$; **Right panel**: correlation between H_α equivalent width and star formation rate.

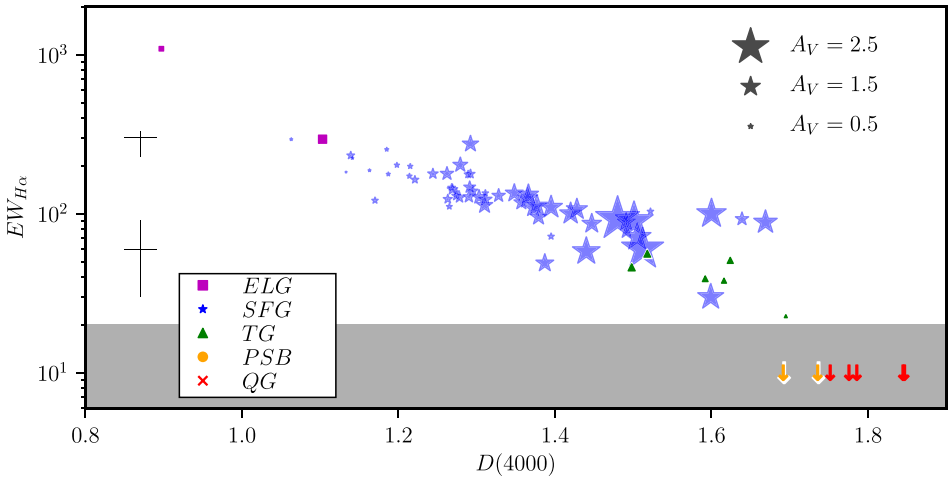

Figure 8

The correlation between H_α equivalent width and $D(4000)$ in Forrest et al. (2018). The special thing in this figure is that it labeled the types of galaxies: magenta for emission line galaxies, blue for star-forming galaxies, green for transitional galaxies (because they are in green valley), yellow for post starburst, and red for quiescent galaxies.

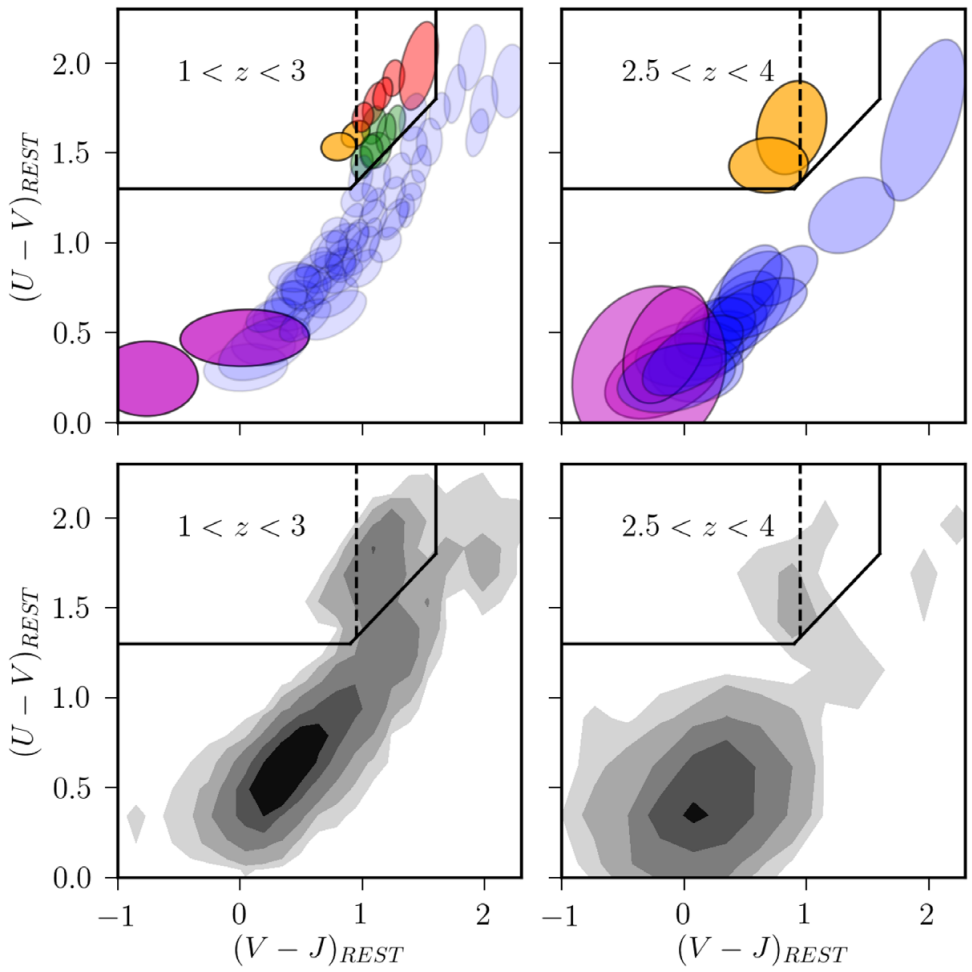

Figure 9

Color-color diagram in Forrest et al. (2018). **Top**: Different colors represent different galaxy types classified by composite SEDs. A special feature of classification using composite SED is that we can track the population of marginal (or rare) types of galaxies on the diagram, e.g., the transitional galaxies (green) and post-starbursts (yellow). **Bottom**: The distribution of number of analogue galaxies.

3.1. Modeling Spectral Data using Gaussian Processes (gp)

A *Gaussian process* is a bunch of random variables, and any finite subset of these random variables is a joint Gaussian distribution (Rasmussen & Williams 2005). GP could be a powerful tool to model any kind of functional data (continuous data) in a non-parametric way. By non-parametric, it actually means we use infinite many parameters to describe our function (Gelman et al. 2014). GP could be treated as a random function (or a stochastic process) which draws samples from the n-dimensional distribution,

$$\mu(x_1), ..., \mu(x_n) \sim Normal((m(x_1), ..., m(x_n)), K(x_1, ..., x_n)). \quad 1.$$

The construction of a GP could be considered as finding the mean function ($m(\vec{x})$) and a suitable covariance function ($K(\vec{x}, \vec{x}')$). In normal cases, a zero mean is usually used as a prior for GP regressions. For $K(\vec{x}, \vec{x}')$, pre-defined covariance functions (e.g., squared potential function $\exp(\frac{-r^2}{2\ell^2})$) are often been implemented. However, the usage of GP in modeling functional data will also be restricted by the intrinsic properties of the covariance functions. Learning a suitable covariance function is the most crucial part of machine learning in GP.

Finding a suitable choice of covariance often reflects our interpretations of the characteristics of our data (Rasmussen & Williams 2005). For example, the usage of the squared potential function $\exp(\frac{-r^2}{2\ell^2})$ implies the assumption that we believe each point on the function would have less impact to each other if they are far away on the functional space. Therefore, we need a special kind of covariance function to suit our purpose of modeling spctral data.

GP: : *Gaussian Processes*

$\sim$ Normal: : Sampling from a Gaussian (normal) distribution

Bayesian ML: : Bayesian Machine Learning

K: : Covariance function (matrix)

μ: : Mean function

Garnett et al. (2017) took a machine learning approach to learn the covariance function, with a wavelength range from Ly_∞ to Lyα, from training data (quasar spectra). The optimization choice was to firstly decompose covariance matrix with (Garnett, Ho & Schneider 2015),

$$\mathbf{K} = \mathbf{M}\mathbf{M}^T, \tag{2.}$$

and then use the first 10 principle components of the flux of quasar spectra, $\mathbf{Y}$, to constitute the matrix $\mathbf{M}$. The optimization was done by maximizing the log likelihood, $\mathcal{L}$, of the data by given $\mathbf{M}$ and absorption noise ω,

$$\mathcal{L}(\mathbf{M}, \omega) = \log p(\mathbf{Y} \mid \lambda, \mu, \mathbf{M}, \mathbf{N}, \omega, z_{\text{qso}}, Model). \tag{3.}$$

The goal of optimizing above function is to find optimal covariance matrix, $\mathbf{M}$, and absorption parameter, ω, with some given conditions. Those conditions are: a given mean vector μ, the noise on the spectra $\mathbf{N}$, the redshift of the QSO, and with a given model. In a perspective of generative modeling, optimizing the data likelihood implies we are trying to find a covariance matrix to better generate our spectral data.

The covariance matrix built in Garnett et al. (2017) with a wavelength range from Ly_∞ to Lyα is in Fig 10. The scale in Fig10 represents the strength of correlations between pairs of rest-frame wavelengths on the QSO spectra. The features of Lyman series are distinct. The off-diagonal term demonstrates the correlations of pairs of corresponding emission lines.

The mean function of GP modeling in Garnett et al. (2017) can be simply obtained by stacking the training spectra,

$$\mu_j = \text{median}(y_{ij}), \tag{4.}$$

where y_{ij} are the fluxes for spectrum. Fig 11 shows the mean function with a range from Ly_∞ to Lyα. The features emission line of Lyman series are also visible in the figure.

Generally, the GP model for spectral data could be described as

$$p(\mathbf{y} \mid \lambda, \mathbf{v}, \omega, z, Model) = Normal(\mathbf{y}; \mu, \mathbf{K} + \mathbf{\Omega} + \mathbf{V}), \tag{5.}$$

where $\mathbf{y}$ is the observed flux of the spectrum, λ is the spectroscopic grids we chose to bin the flux, $\mathbf{v}$ is the instrumental noise given by the observed data (it is SDSS QSO catalogue

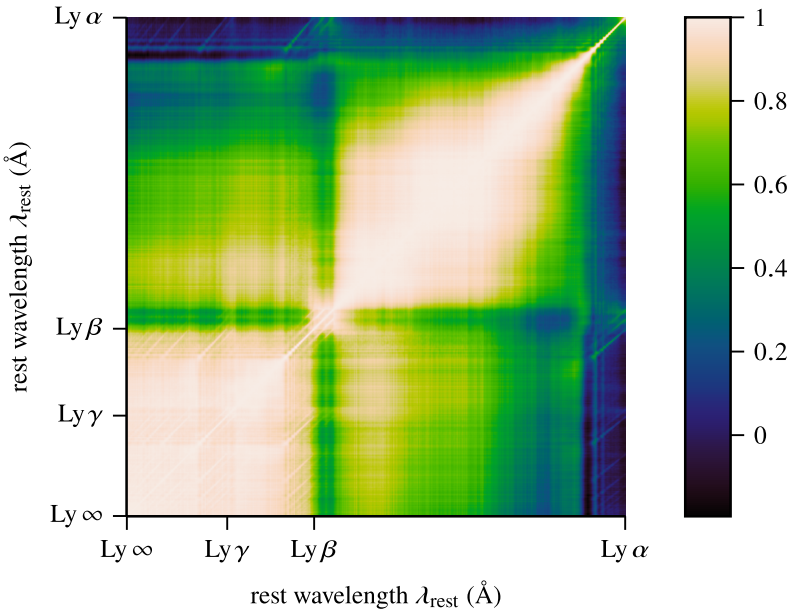

Figure 10

Covariance function for quasar spectra in Garnett et al. (2017)

(Pâris, I. et al. 2012) in Garnett et al. (2017)), z is the redshift dependence of the GP model, and ω is the absorption redshift dependence (it was used to model Lyα forest in Garnett et al. (2017)).

The beauty of generative modeling the spectrum using GP framework is that we are able to fully control the modeling of instrumental noise and redshift dependence uncertainties. In addition, the whole framework is transparent and flexible, which implies it is interpretable and future improvements are achievable.

SUMMARY POINTS

1. *Gaussian Processes.* A flexible Bayesian non-parametric framework which allows us to model any kind of function.

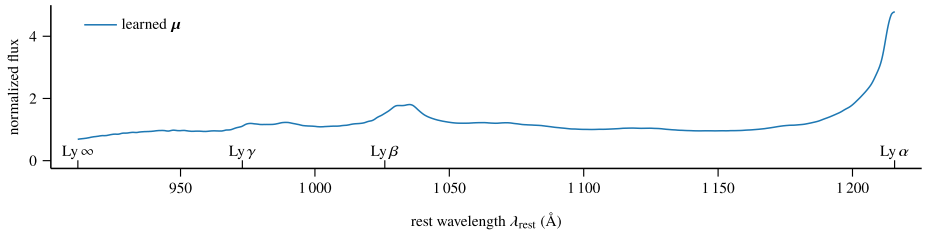

Figure 11

Mean function for quasar spectra in Garnett et al. (2017)

2. Learned covariance matrix $\mathbf{K}$. To model spectral data, we can optimize the covariance matrix using training data to acquire customized covariance function of any given range of wavelengths.

3.2. Dirichlet Processes

The *Dirichlet process* (Teh et al. 2006) is a Bayesian non-parametric method to model infinite mixture models and also could be used for clustering. We haven't seen any application of *Dirichlet processes* (DP) in spctral data clustering. A recent astronomical application of DP is building a mixture model for binary neutron stars in gravitational-wave data (Del Pozzo et al. 2018). Since we expect the application of DP in the clustering of spectral data is achievable in future, we decided to roughly review the basic concept of DP here.

The DP is a stochastic process which is often used to model mixture models. Each random sample in DP is itself a distribution, which means we can treat random samples of a DP as cluster centers of a mixture model. Similar to GP, a finite subset of a DP could be described by Dirichlet distributions.

To model the data v using the Dirichlet mixture model, the clustering memberships could be described by the following conditional probability (Barber 2012),

$$p(v^{1:N} \mid \theta) = \sum_{z^{1:N}} p(v^{1:N} \mid z^{1:N}, \theta) p(z^{1:N}), \tag{6.}$$

where $p(z^{1:N})$ gives the priors over each cluster and θ describe the parameters of the cluster model. The choice of priors over clusters ($p(z^{1:N})$) is crucial

3.2.1. Chinese Restaurant Process. The clustering property of DP is hard to understand simply via mathematical forms above. However, there's an intuitive metaphor described in Teh et al. (2006) to mimic the process of drawing samples from DP using a real-life example: *Chinese Restaurant process* (CRP).

The name of CRP was developed in 1980's. CRP is a process to describe the distribution over partitions.

Now, imagine this: there are infinite number of round tables in a Chinese restaurant, and there are also infinite number of seats in each round table. Each customer will come to

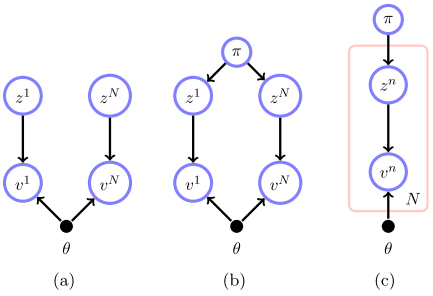

Figure 20.12: **(a)**: A generic mixture model for data $v^{1:N}$. Each z^n indicates the cluster of each datapoint. θ is a set of parameters and $z^n = k$ selects parameter θ^k for datapoint v^n. **(b)**: For a potentially large number of clusters one way to control complexity is to constrain the joint indicator distribution. **(c)**: Plate notation of (b).

Figure 12

The probabilistic graphical model of DP in plate notation in Barber (2012). The symbols here: z is the indicator and $p(z^n)$ implies the prior over n^{th} cluster; v is data; θ is used to describe the parameters on the cluster model (the shared model) while z could be treated as the hidden variable used to described component models (the individual models); π here is the parameter of a Dirichlet distribution, which is used to describe distributions.

the restaurant one by one (just like sampling). Let's assume the first customer choose the first table. Which table would next person choose? If every customers are friendly[3], the next customer may intend to choose the round table which has more people there. Therefore, the probability of choose k^{th} table could be proportional to the number of people n_k in that round table. She/he may also have a small amount of probability to choose a new empty table to sit.

The mathematical form of CRP can be described as

$$\theta_{n+1} \mid \theta_1 ..., \theta_n \sim \frac{1}{\alpha + n} \left(\alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*} \right), \tag{7.}$$

where *alpha* is a constant and $\delta_{\theta_k^*} = 1$ only when θ_k^* is selected. The above equation is equivalent to CRP in this way:

1. the first customer ($n = 0$): θ_1 partition would be selected from a smooth distribution H.
2. the second customer ($n = 1$): $\theta_2 \mid \theta_1$ partition would be selected from either $\frac{\alpha H}{\alpha+1}$ (a new table) or $\frac{1}{\alpha+1}$ (the table with customer 1).
3. the $(n+1)^{th}$ customer: $\theta_{n+1} \mid \theta_n, ..., \theta_1$ would be selected from either

[3]There are some literatures discussed about the unfriendly situations.

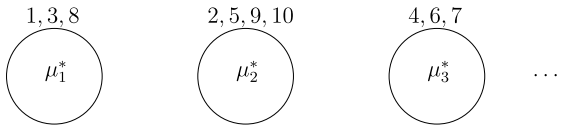

Figure 13

An illustration of CRP in Blei (2007).

SUMMARY POINTS

1. Summary point 1. These should be full sentences.

FUTURE ISSUES

1. Future issue 1. These should be full sentences.

ACKNOWLEDGMENTS

Acknowledgements, general annotations, funding.

LITERATURE CITED

Barber D. 2012. *Bayesian Reasoning and Machine Learning*. New York, NY, USA: Cambridge University Press

Bishop C, Tipping ME. 2003. Bayesian regression and classification. Advances in Learning Theory: Methods, Models and Applications

Blanton MR, Roweis S. 2007. *The Astronomical Journal* 133:734–754

Blei D. 2007. *COS 597C: Bayesian nonparametrics, Lecture 1*

Brammer GB, van Dokkum PG, Coppi P. 2008. *The Astrophysical Journal* 686:1503–1513

Del Pozzo W, Vecchio A, Berry CPL, Ghosh A, Haines TSF, Singer LP. 2018. *Monthly Notices of the Royal Astronomical Society* 479:601–614

Forrest B, Tran KVH, Broussard A, Cohn JH, Robert C. Kennicutt J, et al. 2018. *The Astrophysical Journal* 863:131

Garnett R. 2018. Active search in big data era

Garnett R, Ho S, Bird S, Schneider J. 2017. *Monthly Notices of the Royal Astronomical Society* 472:1850–1865

Garnett R, Ho S, Schneider J. 2015. *Finding Galaxies in the Shadows of Quasars with Gaussian Processes*. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15. JMLR.org

Gelman A, Carlin JB, Stern HS, Rubin DB. 2014. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd ed.

Kriek M, van Dokkum PG, Whitaker KE, Labbé I, Franx M, Brammer GB. 2011. *The Astrophysical Journal* 743:168

Pâris, I., Petitjean, P., Aubourg, É., Bailey, S., Ross, N. P., et al. 2012. *A&A* 548:A66

Rasmussen CE, Williams CKI. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press

Straatman CMS, Spitler LR, Quadri RF, Labbé I, Glazebrook K, et al. 2016. *The Astrophysical Journal* 830:51

Teh YW, Jordan MI, Beal MJ, Blei DM. 2006. *Journal of the American Statistical Association* 101:1566–1581