

# Classifying Galaxies using a data-driven approach

Vasileios Stampoulis  
Supervisor : Prof. David van Dyk

Department of Mathematics  
Imperial College London

London, April 2015



# Outline

- 1 The Classification Problem
  - The Scientific Problem
  - The Dataset
  - Existing Classification methods
  - Inefficiency of Existing Methods
- 2 Clustering Methodology
  - Introduction
  - Finite Mixture Models
  - Stylized Facts of the Dataset
- 3 Application to the SDSS dataset
  - Initial Approach
  - Current Approach
  - Comparing the Classification Schemes
- 4 Further Research Directions

# Outline

- 1 The Classification Problem
  - The Scientific Problem
  - The Dataset
  - Existing Classification methods
  - Inefficiency of Existing Methods
- 2 Clustering Methodology
  - Introduction
  - Finite Mixture Models
  - Stylized Facts of the Dataset
- 3 Application to the SDSS dataset
  - Initial Approach
  - Current Approach
  - Comparing the Classification Schemes
- 4 Further Research Directions

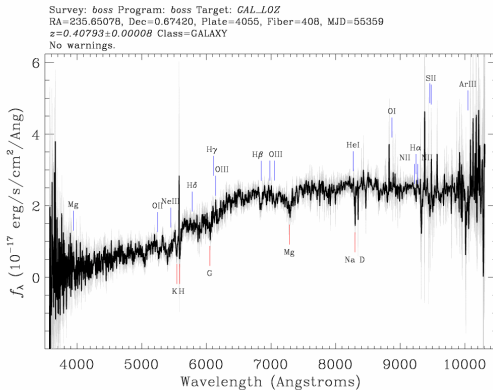
# The Scientific Problem

**Spectroscopy** is the study of the measurement of radiation intensity as a function of wavelength.

- When electrons orbiting atoms fall from energized orbits to rest orbits they emit light photons with wavelengths resulting from the excess energy given up by the electrons.
- The atoms and molecules have unique spectra.
- These spectra can be used to detect, identify and quantify information about the atoms and molecules.

Spectroscopy has also been utilised in identifying the **main power source** in active galaxies.

# The Scientific Problem



**Figure :** A selected spectrum from the dr10 BOSS data, showing absorption (red) and emission (blue) lines (<http://www.sdss.org>).

# The Scientific Problem

Based on the different mechanisms that excite the gas inside the galaxies (and which, as a result of those mechanisms, glows in different wavelengths), the Galaxies may be separated, according to Kewley et al (2001):

- **HII** region-like galaxies (star forming galaxies).
- **AGN** (Active Galactic Nuclei) which can be divided further to **LINER** and **Seyferts**.

# The Dataset

- Sloan Digital Sky Survey (SDSS).
- The SDSS is a major multi-filter imaging and spectroscopic redshift survey.
- The latest Data Release (DR 10) the from SDSS-IIIs Baryon Oscillation Spectroscopic Survey (BOSS) contains 1,848,851 Optical Galaxy spectra.

## Existing Classification methods

- From early on (see Alloin et al. (1978)), intensity ratios, such as  $OIII/H\beta$ , have been utilized as means to classify the emission line spectra of galactic HII region.
- Baldwin et al. (1981) proposed three diagnostic diagrams based on four emissions intensities ratios;  $\log(NII/H\alpha)$ ,  $\log(SII/H\alpha)$ ,  $\log(OI/H\alpha)$  and  $\log(OIII/H\beta)$ .
- Those three diagrams, known as Baldwin-Phillips-Terlevich (BPT) diagrams, were introduced in order to classify the dominant energy source in emission-line galaxies.



# Existing Classification methods

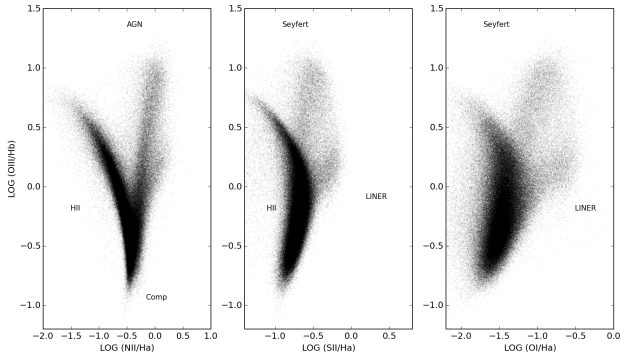


Figure : Example diagnostic (BPT) diagram based on observational data from the DR10 BOSS data.

## Existing Classification methods

In the **first diagnostic**:

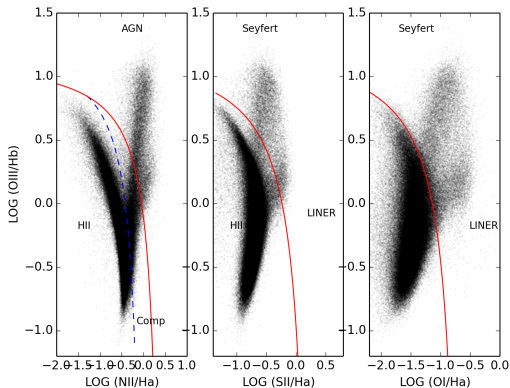
- There are two different loci of sources; a stream moving from the bottom middle to the top left and another fuzzier stream moving from the bottom middle to the top right.
- The first stream corresponds to the star-forming galaxies and the latter to the AGN.

The **second and third diagnostics** verify that the AGN are made up of two groups, the Seyfert and the LINER.

## Existing Classification methods

- Kewley et al. (2001) introduced a maximum 'starburst' line on the BPT diagrams.
- This line (**red line** in the next slide) defines the upper limit of the star-forming galaxies and was developed using theoretical models.
  
- Kauffmann et al. (2003) presented an empirical line to distinguish the pure HII galaxies from Seyfert- HII **composite objects** (**cyan dashed line** in the next slide).
- The spectra of those composite galaxies contain significant contributions from both AGN and star-formation according to Kewley et al. (2006).

# Existing Classification methods



**Figure :** Diagnostic (BPT) diagram based on observational data from the DR10 BOSS data with the maximum 'starburst' line (red) and composite line (cyan) plotted.

# Inefficiency of Existing Methods

- The theoretical model of Kewley et al. (2001) fails to distinguish the AGN from the star-forming galaxies.
- The use of the three diagnostics independently of each other often gives contradicting classification for the same source.
- In Kewley et al. (2006), 8 percent of the galaxies in their sample are characterized as ambiguous galaxies.

# Outline

- 1 The Classification Problem
  - The Scientific Problem
  - The Dataset
  - Existing Classification methods
  - Inefficiency of Existing Methods
- 2 Clustering Methodology
  - Introduction
  - Finite Mixture Models
  - Stylized Facts of the Dataset
- 3 Application to the SDSS dataset
  - Initial Approach
  - Current Approach
  - Comparing the Classification Schemes
- 4 Further Research Directions

# Introduction

- The goal of **Cluster Analysis** is to partition the dataset into subgroups so that the members assigned to the same cluster are more similar (according to a certain criterion) compared to those in other clusters.
- The Probabilistic (model-based) algorithms assume that the data is an i.i.d. sample from a population described by a density function, which is taken to be a mixture of component density functions.
- It is common to assume that the mixture components are all from the same parametric family, such as the normal.

# Finite Mixture Models

Given data  $x$  ( $n \times p$ ) with independent multivariate observations  $x_1, \dots, x_n$ , the general framework for a Finite Mixture model with  $K$  components:

$$p(x|\theta, \pi) = \sum_{k=1}^K \pi_k f_k(x|\theta_k),$$

where  $f_k$  and  $\theta_k$  are the density and parameters of the component  $k$  in the mixture model and  $\pi_k$  is the probability that an observation comes from the component  $k$ . Furthermore,  $\pi_k \geq 0$  and  $\sum_{i=1}^K \pi_k = 1$ .



# EM Algorithm

Dempster et al. (1977) proposed a framework for computing the maximum likelihood estimators in finite mixture models for a general specification of the component density functions  $f_k$  using the **Expectation-Maximization (EM) algorithm**.

- The EM algorithm is an iterative method for posterior mode finding.
- Given a statistical model consisting of a set  $x$  of observed data, a set of unobserved latent data or missing values  $z$ , and a vector of unknown parameters  $\theta$ , the EM alternates between performing an expectation (E) step and a maximization (M) step.

# EM Algorithm

**E-step:** Compute  $Q(\theta|\theta^{(t)}) = E[\log p(\theta|x, z)|x, \theta^{(t)}]$ ,

**M-step:** Set  $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta|\theta^{(t)})$ ,

where the subscript  $t$  indexes the iteration.

The EM algorithm enjoys stable **convergence properties**:

- The likelihood  $p(x|\theta)$  increases in each iteration.
- The algorithm converges to a stationary point of  $p(x|\theta)$ .

## Finite Mixture Models cont.

In the context of **Finite Mixture models** Dempster et al. (1977):

- Introduced an unobserved vector  $z$  ( $n \times K$ ), where  $z_i$  is the indicator vector of length  $K$  with  $z_{ik} = 1$  if the observation  $x_i$  belongs to cluster  $k$  and 0 otherwise.
- The  $z$ 's are latent variables that specify from which component each observation is drawn.

## Finite Mixture Models cont.

The joint distribution  $p(x, z|\theta, \pi) = p(z|\theta, \pi)p(x|z, \theta, \pi)$ .

- $p(z|\theta, \pi)$  is multinomial distribution  $p(z|\theta, \pi) = \prod_{k=1}^K \pi_k^{z_k}$ ,  
 $\theta = (\theta_1, \dots, \theta_K)$  and  $\pi = (\pi_1, \dots, \pi_K)$ .
- Conditional on  $z_{ik} = 1$ ,  $x_i \sim f_k(x_i|z_{ik} = 1, \theta_k)$ .

The resulting complete-data log-likelihood (the likelihood of both the observed and missing data) is:

$$\ell(\theta, \pi|x, z) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log[\pi_k f_k(x_i|\theta_k)]$$

## Finite Mixture Models cont.

**E-step:** estimate the components of the  $z_i$ 's given the observed data  $x$  and the current (at step  $t$ ) fitted parameters  $(\theta^{(t)}, \pi^{(t)})$ .

This represents the current conditional probabilities of  $x_i$  coming from each component:

$$E[z_{ik} | \theta^{(t)}, x] = \frac{\pi_k^{(t)} f_k(x_i | \theta_k^{(t)})}{\sum_{k=1}^K \pi_k^{(t)} f_k(x_i | \theta_k^{(t)})} \quad (1)$$

## Finite Mixture Models cont.

**M-step:** maximize the complete data log-likelihood with respect to  $\pi$  and  $\theta$ .

The derivatives of the complete-data log-likelihood for component density functions  $f_k$  that belong to an Exponential Family have an attractive representation, since they are functions of the sufficient statistics.

# Multivariate Gaussian Mixture Model

- The Multivariate Gaussian mixture model has been used with success for many clustering problems.
- Multivariate normal mixtures can be used for data with varying structures due to the flexibility in how we define the covariance matrices.

The Multivariate Normal distribution is defined as:

$$f_x(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

# Multivariate Gaussian Mixture Model

- The EM formulation for multivariate Gaussian mixture is presented in detail in Dempster et al. (1977).
- The E-step has the same formulation as in Equation 1.
- In the M-step, the estimates of the means, the prior probabilities  $\pi_k$  and the covariances (eg. full covariance matrix) have closed form solutions.



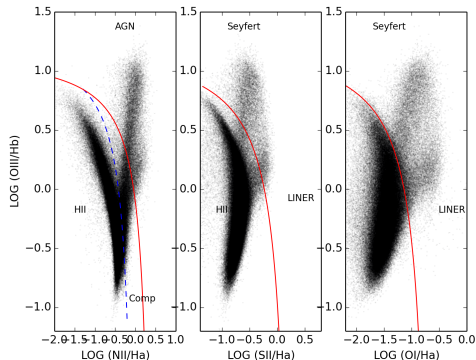
# Limitations of the Multivariate Normal Model

- Fraley et al. (2002) pointed out that data generated by mixtures of multivariate normal distributions are characterized by clusters centered at the means  $\mu_k$  with increased density for points closer to the means.
- The practical use of multivariate mixture models is limited for data that exhibit non-normal features, including asymmetry, multi-modality and heavy tails as in the SDSS dataset.

# Non Gaussian Features of the Dataset

- The distribution of the Star forming Galaxies (the HII-region ) is clearly highly skewed and not-linear (curve-shaped).
- In the LINER and Seyfert clusters, the mass doesn't seem to be concentrated at the centers.
- There seems to be a big overlap between the 4 different clusters.

# Non Gaussian Features of the Dataset



**Figure :** Example diagnostic (BPT) diagram based on observational data from the DR10 BOSS data with the maximum 'starburst' line (red line) and composite line (cyan line) plotted.

# Outline

- 1 The Classification Problem
  - The Scientific Problem
  - The Dataset
  - Existing Classification methods
  - Inefficiency of Existing Methods
- 2 Clustering Methodology
  - Introduction
  - Finite Mixture Models
  - Stylized Facts of the Dataset
- 3 Application to the SDSS dataset
  - Initial Approach
  - Current Approach
  - Comparing the Classification Schemes
- 4 Further Research Directions

# Fitting 3 Multivariate Normal Distributions + Curved Gaussian

- 3 Multivariate Gaussian distributions for the Seyferts, the LINERs and the composites.
- A curved Gaussian distribution for the HII galaxies.

$$p(x_1, x_2, x_3, x_4) = p(x_4)p(x_3|x_4)p(x_2|x_3, x_4)p(x_1|x_2, x_3, x_4)$$

$$p(x_4) \sim N(\mu_4, \sigma_4^2)$$

$$p(x_3|x_4) \sim N(a_0 + a_1x_4 + a_2x_4^2, \sigma_3^2)$$

$$p(x_2|x_3, x_4) \sim N(b_0 + b_1x_4 + b_2x_4^2 + b_3x_3 + b_4x_3^2, \sigma_2^2)$$

$$p(x_1|x_2, x_3, x_4) \sim N(c_0 + c_1x_4 + c_2x_4^2 + c_3x_3 + c_4x_3^2 + c_5x_2 + c_6x_2^2, \sigma_1^2)$$

# Fitting 3 Multivariate Normal Distributions + Curved Gaussian

- We pre-allocated about 50% of the data based on the existing classification scheme.

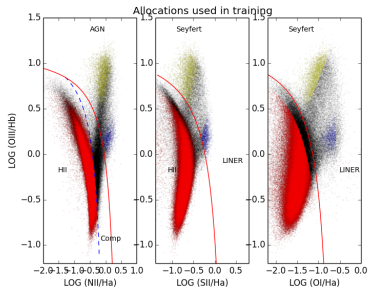
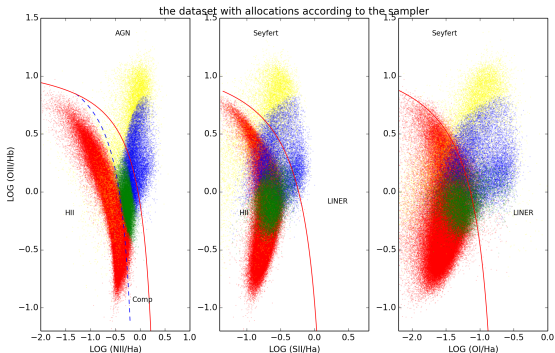


Figure : The datapoints that were pre-allocated.

# Fitting 3 Multivariate Normal Distributions + Curved Gaussian



**Figure :** The SDSS dataset in which each datapoint is plotted with different color according to the most probable allocation.

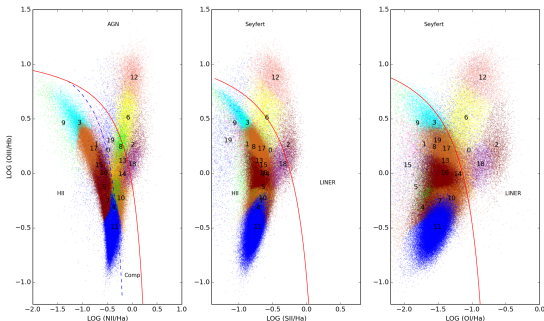
# Fitting 3 Multivariate Normal Distributions + Curved Gaussian

- We are using too much prior information.
- When we reduce the percentage of the pre-allocated data, the resulting classification scheme is far from expected.
- This is mainly because of the skewness in the HII galaxies.



# Fitting 20 Multivariate Normal Distributions

**Approach:** Fit 20 Multivariate Normal distributions and then try to combine them.



**Figure :** Fitting 20 multivariate Normals in the SDSS dataset and plotting each datapoint with different color according to the most probable allocation.

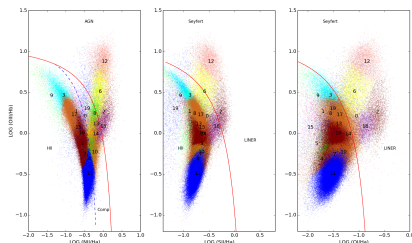
# Combining the 20 Multivariate Normal Distributions

HII : 1, 3, 4, 5, 7, 9, 11, 15, 16, 17

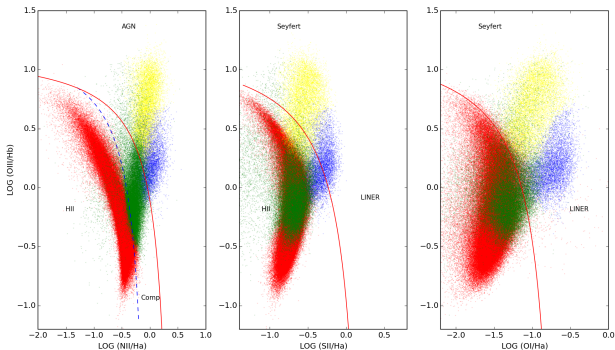
Seyfert : 6, 12

LINER : 2, 18

Composites : 0, 8, 10, 13, 14, 19

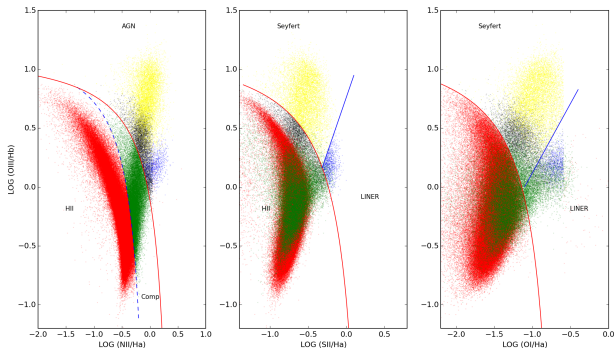


# Combining the 20 Multivariate Normal Distributions



**Figure :** Combining the 20 multivariate Normals in the SDSS dataset as explained above and plotting each datapoint with different color according to the most probable allocation.

# The theoretical Classification



**Figure :** The SDSS dataset in which each datapoint is plotted with different color according to the allocation from the theoretical classification scheme.

## Comparing the Classification Schemes

The following Table depicts the allocations of the galaxies from the SDSS dataset to the different clusters based on the existing theoretical models (Kewley et al. (2006)) and our Data-driven approach.

	HII	Seyferts	LINER	Composites	Ambiguous	Total
Theoretical	157515	8501	1024	29714	13037	209791
Data-driven	161098	11134	6805	30754	NA	209791

# Outline

- 1 The Classification Problem
  - The Scientific Problem
  - The Dataset
  - Existing Classification methods
  - Inefficiency of Existing Methods
- 2 Clustering Methodology
  - Introduction
  - Finite Mixture Models
  - Stylized Facts of the Dataset
- 3 Application to the SDSS dataset
  - Initial Approach
  - Current Approach
  - Comparing the Classification Schemes
- 4 Further Research Directions

## Merging the 20 Multivariate Normal distributions

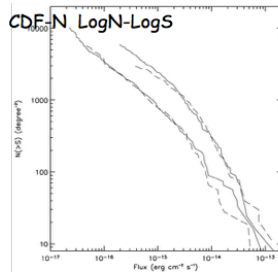
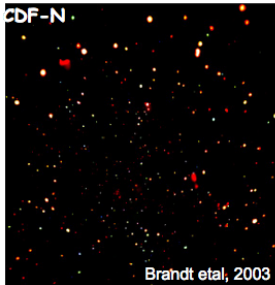
- Merging the 20 Multivariate Normal distributions by-eye give us a good description of the 4 clusters.
- Could we use a more objective way of merging the Normals?

**Hierarchical Clustering:** merge clusters that are closer to each other by using a criterion (or criteria).

- Distance between the means of the clusters.
- Distance from a curve and 2 straight lines.

# LogN - LogS

**Definition:** Cumulative distribution of number of sources per unit intensity; S is the flux (observed source intensity).



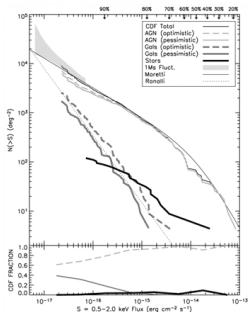
Bauer et al 2006

- (a) CHANDRA Deep Field North. (b) LogN-LogS diagram for CHANDRA Deep Field North.



# LogN - LogS

**Goal:** Use the classification scheme in order to create LogN-LogS diagrams for AGNs and star-forming Galaxies and check the total contribution of each class.








**Figure :** LogN-LogS diagram for CHANDRA Deep Field North for all sources and for AGNs and Stars.

# Thank you!



Many Thanks to:

David van Dyk  
Andreas Zezas

# Bibliography

-  Alloin, D., Bergeron, J., & Pelat, D. (1978). Properties of a Sample of Irregular Galaxies. *Astronomy and Astrophysics*, 70, 141.
-  Baldwin, J. A., Phillips, M. M., & Terlevich, R. (1981). Classification parameters for the emission-line spectra of extragalactic objects. *Publications of the Astronomical Society of the Pacific*, 5-19.
-  Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.
-  Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.
-  Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D., ... & Schneider, D. P. (2003). The host galaxies of active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 346(4), 1055-1077.

# Bibliography

-  Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., and Trevena, J. (2001). Theoretical modeling of starburst galaxies. *The Astrophysical Journal*, 556(1), 121.
-  Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. (2006). The host galaxies and classification of active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 372(3), 961-976.
-  Kewley, L. J., Heisler, C. A., Dopita, M. A., & Lumsden, S. (2001b). Optical classification of southern warm infrared galaxies. *The Astrophysical Journal Supplement Series*, 132(1), 37.