Imperial College London

Creating a Data-driven Model of Galaxy Spectra

Katherine Sephton **Annabel Hii**

Project Code: ASTR-LEISTEDT-1 **Supervisor: Boris Leistedt** **Astrophysics Group**

1. Introduction

Hundreds of terabytes of astronomical data from hundreds of millions of sources across the electromagnetic spectrum have been collected in sky surveys, such as the Sloan Digital Sky Survey (SDSS), over the last decade alone. These surveys provide information on the interactions, evolutions and properties of the objects surveyed.

An un-biased model of galaxy spectra that is not influenced by pre-existing physical models can be created purely from the SDSS data.

2. Objective

The objective of this project is to produce a data-driven model that can reconstruct spectra accurately by accounting for heteroskedastic noise, expanding on the work of Portillo et al. (2020) [1].

3. Method

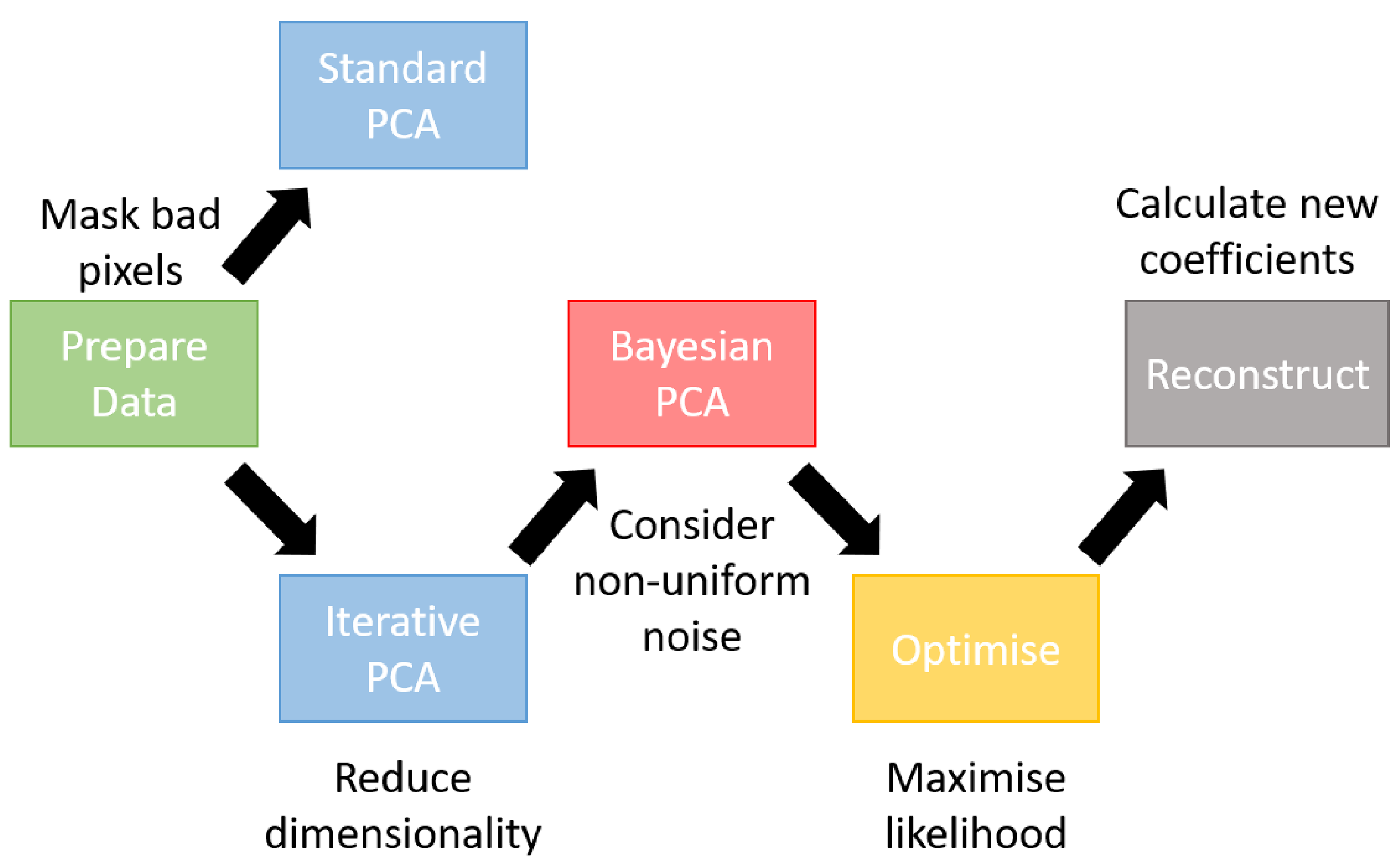

Figure 1: An overview of the five step process utilising principal component analysis (PCA) to build a model. Bayesian PCA includes a prior distribution and considers the noise on the values, improving on standard PCA [2] .

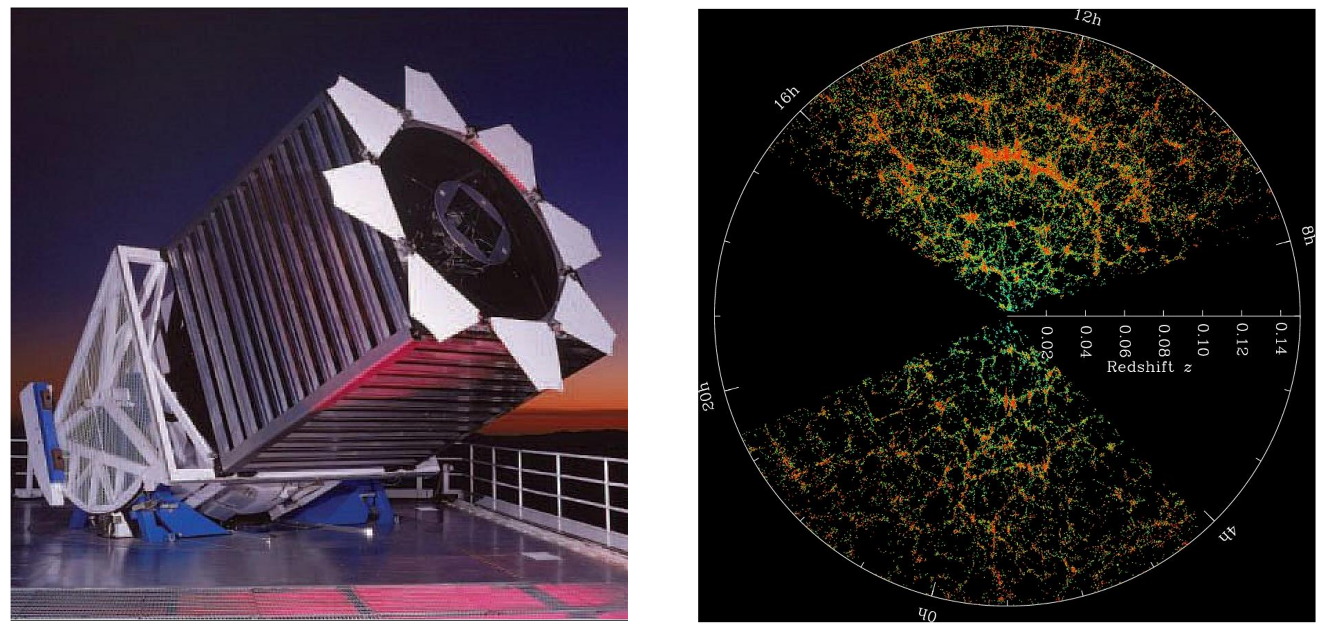

Figure 2 : (Left) An image of the 2.5 m telescope used for the SDSS [3]. (Right) The SDSS map of the Universe [4].

4a. Results

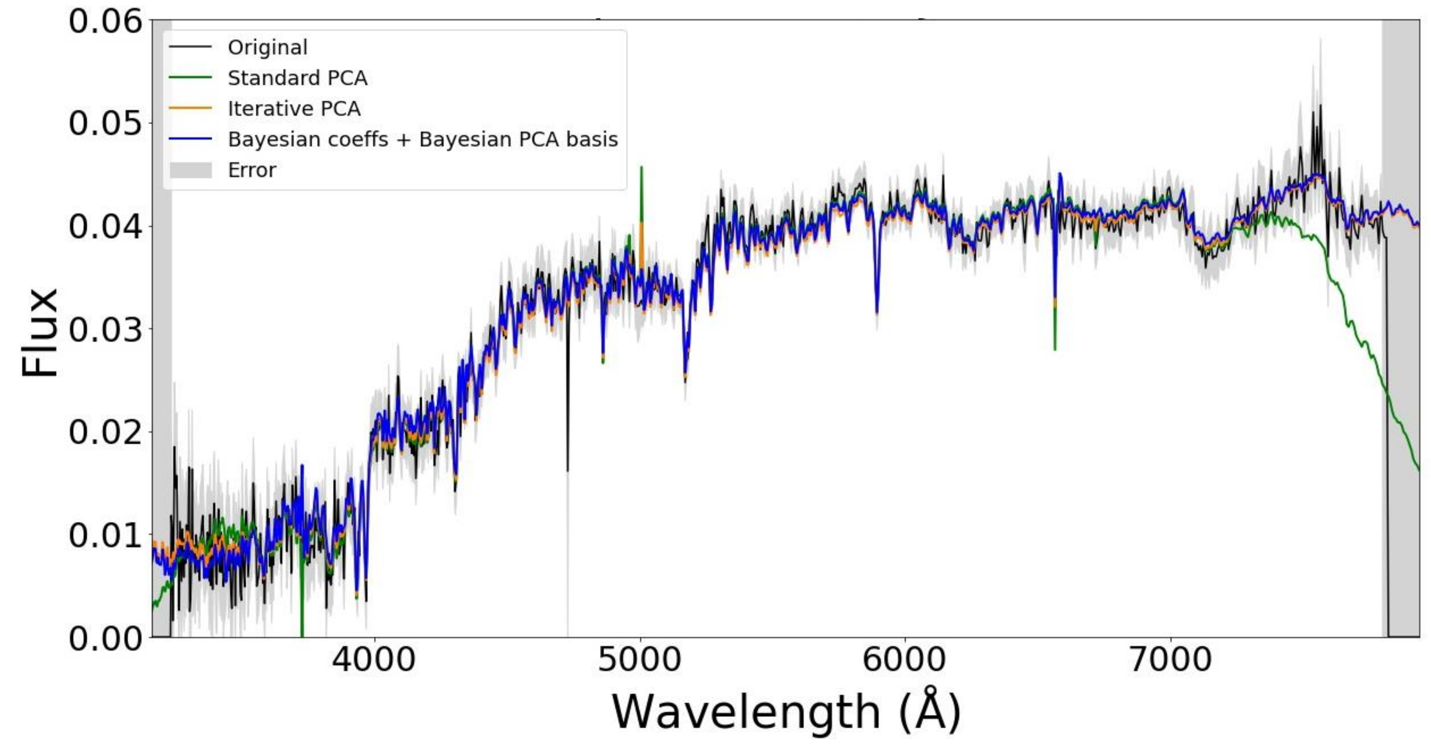

Figure 3: A reconstructed spectrum and the original data (plate: 1733, MJD: 53047, fiber: 439). Bayesian PCA reconstructs better than standard and iterative PCA at the beginning and the end of the spectrum, where the error in the data is greater.

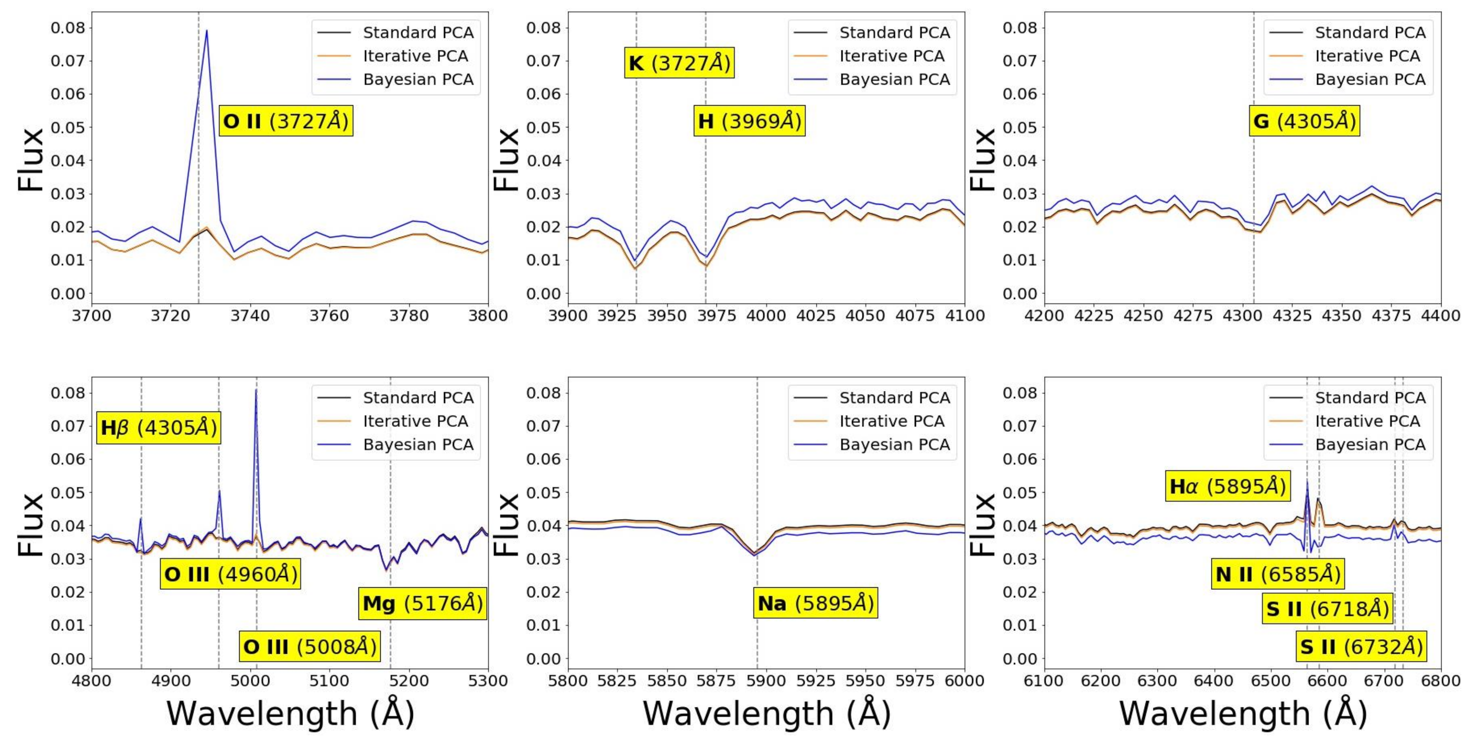

Figure 4: Absorption and emission lines from a mean spectrum reconstruction of a normal galaxy, with reference wavelengths (dashed lines) [5].

4b. Results

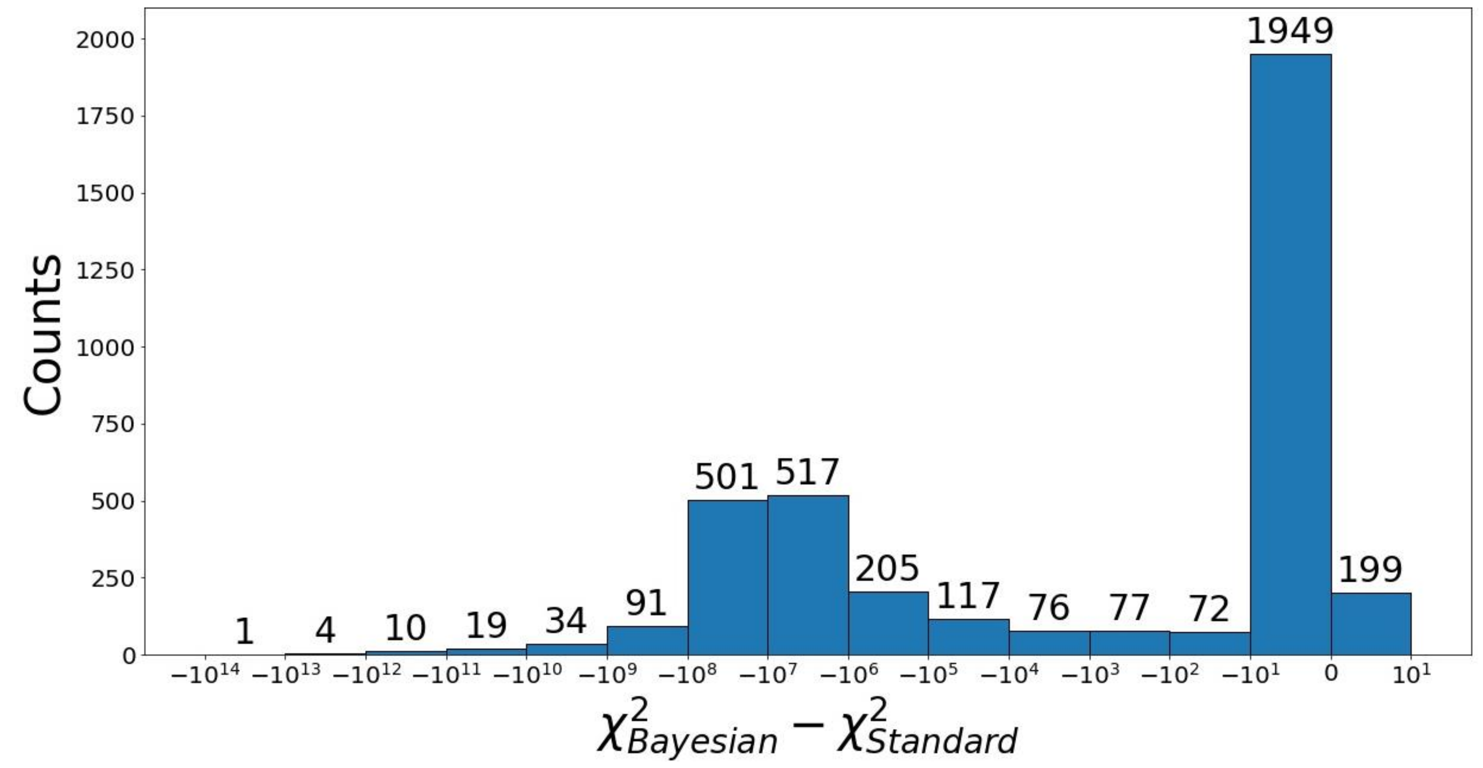

Figure 5: Difference between X^2 values for Bayesian PCA and those for standard PCA. Most spectra show a negative difference, implying that Bayesian PCA gives a better quality of reconstruction than standard PCA.

5. Conclusion

Bayesian PCA is more effective than standard and iterative PCA for galaxy spectra that exhibit heteroskedastic noise.

The quality of reconstruction is especially better for Bayesian PCA at the beginning and the end of the spectra, where the original data contains pixels with missing or negative flux.

Since Bayesian PCA is better suited to accounting for spectra noise, it can be used to create more accurate models for different galaxy types to analyse their features.

As our model uses spectroscopic data, photometric data can be added to create a more detailed model of galaxies.

References

[1] Portillo SKN, Parejko JK, Vergara JR, Connolly AJ. Dimensionality Reduction of SDSS Spectra with Variational Autoencoders. *Astron. J.* 2020;160(1): 45. Available from: doi:10.3847/1538-3881/ab9644

[2] Bishop CM. *Pattern Recognition and Machine Learning.* Springer; 2006.

[3] *SDSS.* Available from: https://directory.eoportal.org/web/eoportal/satellite-missions/s/sdss [Accessed: 1 March 2021]

[4] *The SDSS map of the Universe.* Available from: https://www.sdss.org/science/ [Accessed: 26 February 2021]

[5] SDSS. *Table of Spectral Lines Used in SDSS.* Available from: http://classic.sdss.org/dr6/algorithms/linestable.html [Accessed: 1 March 2021]