Using a Gaussian Mixture Clustering Algorithm to Inspect z = 6-8 Galaxies from JWST CEERS

Lipika Chatur[1] , Marissa Perry[1], and Urvi Thakurdesai[1]
[1]The University of Texas at Austin

The University of Texas at Austin
Department of Astronomy
College of Natural Sciences

CEERS GEVIP NSF

Introduction

The epoch of reionization was a period that transformed the universe from a dark, dense fog of gas to a developed cosmic field. This remarkable transformation led to the emergence of the first ever stars in the universe and well as the formation of the astrophysical objects we see in our universe today, such as galaxies, planets, etc. Thus, studying the reionization era is valuable for scientists as they can gain insight into the assembly of the early universe, the formation of early stars and galaxies, and their evolution to present time. In our research, we focused specifically on z = 6 - 8 galaxies within the JWST (James Webb Space Telescope) and CEERS (The Cosmic Evolution Early Release Science) survey. We define a spurious source as a false detection of a galaxy, thus unusable in a dataset. The NIRCam on JWST CEERS captures a source in multiple filters (wavelengths) in order to more clearly determine the source's distance from us.

Research Goal

Filter z = 6 - 8 galaxy candidates from a sample of 1816 JWST CEERS using various methods such as dimensional reduction of our dataset and visual inspection to detect whether sources were potentially good or spurious and further cluster them separately. Criteria for notably bad sources included diffraction spikes, image edge, disturbance by nearby source, etc.

Methods

First, we visually inspected the NIRCam photometry of each of the 1816 sources and labeled them as either a real or spurious source. We found 3% of our dataset contained spurious sources, where a majority of these were reasoned as diffraction spike sources (Fig 1). Next, we used Python and the Jupyter Notebook IDE to import this labeling and image data into a dataframe. Each source contained an identification number, label, reason (if spurious), and multi-wavelength filter image data. Before applying a data mining task on our dataset, we took a preprocessing step to reduce the size of our dataset, thereby reducing the computational cost of our task. For this, we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique. Lastly, we implemented a clustering task to our dimensionally reduced dataset. For this, we used the Gaussian Mixture Model (GMM) clustering algorithm. Our final task was to choose a reasonable number of clusters for the algorithm to split up our dataset. We found that ~30 clusters yielded the best model of our dataset (Fig. 4).

Figures and Results

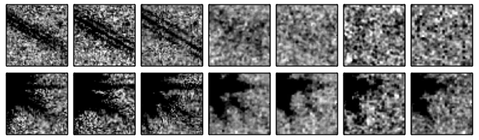

Fig 1a,1b. Examples of a spurious source in our dataset. Detection of a diffraction spike, a stellar imaging artifact (top). External source being too close to our target galaxy (bottom).

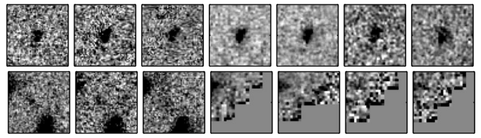

Fig 2a, 2b. Example of a good source in our dataset (top). Example of spurious source with edge data and noise (bottom).

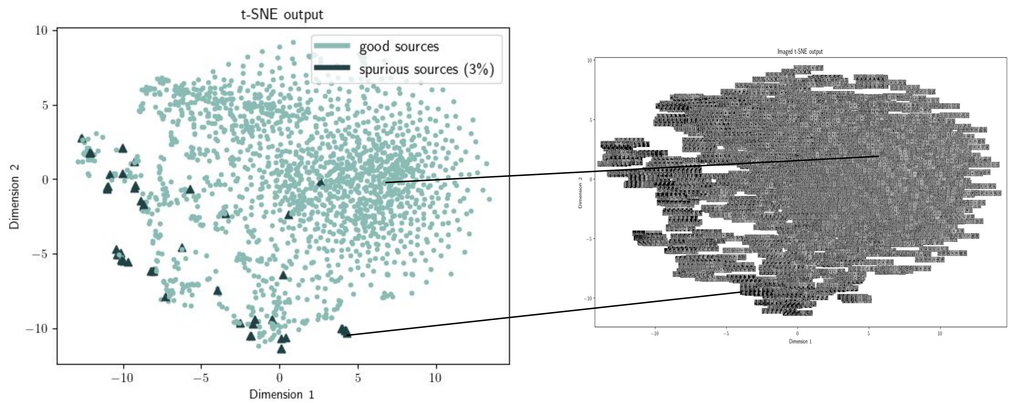

Fig 3. Graph of t-SNE plot with 1816 sources visually inspected and categorized as good or spurious(left). Graph of images of 1816 sources that correspond to each scatter point (right). Note that the x and y dimensional axes are meaningless.

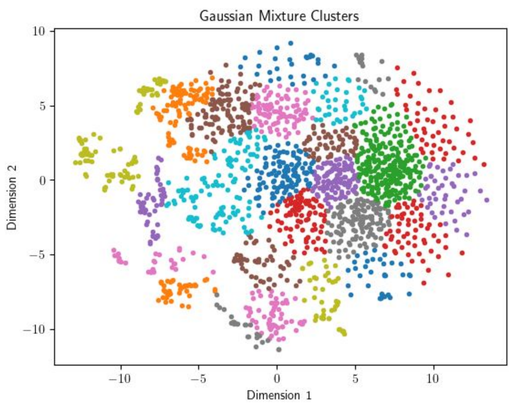

Fig.4 Plot of GMM Clustering model implemented over our t-SNE dimensionality reduced dataset. A total of 30 clusters were produced from this run. Note that the x and y dimensional axes are meaningless.

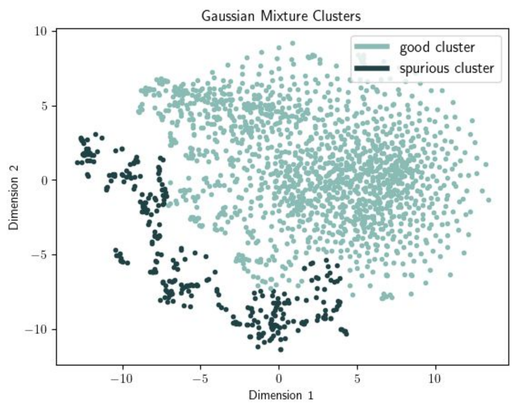

Fig 5. Plot highlighting a significantly spurious region in our dataset. This region contains 7 of the clusters from Fig 4 which exhibited a significant percentage (>5%) of spurious sources.

Future Work

As we can see from the results of our project, we were successfully able to implement a machine learning method of locating spurious sources. Our technique of combining GMM with t-SNE showed that the majority of our spurious sources lie on the edges of the main field, as seen in Fig 5. However, choosing a number of clusters to run the algorithm is subjective. We noticed that implementing 30+ clusters allowed the algorithm to pick out the spurious sources precisely. However, a more methodical approach to choosing this value could be implemented in the future. Another potential advance in our methods would be the inclusion of a training and a test dataset. Feeding the GMM algorithm galaxies which have labels attached to them, we can analyze the accuracy of its predictions. By showing the algorithm the correct labels, we can train it to learn from its false predictions, leading to an increase in performance of the algorithm. Lastly, increasing the overall number of galaxies in our dataset would also result in better performance by creating a stronger Gaussian spread.

Conclusion

Our project aimed at classifying good and bad sources in a sample of the JWST CEERS survey which was populated by galaxies with 6-8 redshift. With visual inspection and t-SNE, we were able to separate the sources and identify whether a galaxy is a reasonable enough source to perform research on. Finally, by applying the GMM clustering algorithm to our dataset, we were able to analyze the clustered dimensionally reduced dataset. This algorithm can further be used by other scientists looking to effectively find good galaxy candidates. In addition to visually inspecting the data, machine learning is a great additional eye when it comes a large amount of sources.

Acknowledgements

We would like to thank Dr. Finkelstein for providing us with our dataset and Gene Leung, Oscar Chavez Ortiz, and Nick Davila for guiding us throughout this project.