University of Waterloo

STAT 840 — FINAL PAPER

Unsupervised Clustering of Simulated Galaxies from Illustris-TNG

Dario Greco (Primary Author)
Scott Holtshousen

April 2025

Abstract

In this paper, we explore a variety of unsupervised clustering techniques on simulated galaxy formation data from the Illustris-TNG project [6]. For each galaxy, physical properties such as stellar mass, dark matter mass, gas mass, and star formation rate are retrieved and used for clustering. Hierarchical clustering and Gaussian mixture models (GMMs) are initially applied to galaxy snapshot data at a fixed redshift ($z = 0$). We then reduce the dimensionality of the dataset using principal component analysis (PCA) and reapply the clustering methods.

Next, we consider the temporal evolution of individual galaxies over the full simulation period. To capture these changes, we apply techniques from Functional Data Analysis (FDA) to properties such as star formation rate, treating them as growth curves. In particular, we cluster the smoothed galaxy curves using K-Means, and subsequently cluster again after applying functional principal component analysis (FPCA).

1 Introduction

The study of how galaxies form and evolve has long been a central focus in astrophysics research. Simulations offer a simple and insightful way of investigating the formation and evolution of galaxies. The Illustris-TNG project [6] offers high-quality galaxy formation simulations data, spanning from the Big-Bang to the present day. For this study, we were granted access to the Illustris-TNG JupyterLab, which then gave us access to the code to begin exploring these simulations, particularly the TNG-50 data [5] (see Figure 11 in the Appendix for a look at a full simulation).

While statistical clustering of galaxy simulation data has previously been explored by Cohn and van de Voort [2], we wanted to develop a similarly accessible overview of basic clustering techniques applied to the TNG data. In our study, we viewed galaxy properties in two different, but related, manners. The first was to look at the galaxies at a fixed redshift ($z = 0$ or "present-day") and apply hierarchical clustering and Gaussian mixture models (GMMs) to a subset of the data.

The second approach was to apply Functional Data Analysis (FDA) to analyze time-series data of star formation rate (sfr) over time. We then clustered the "smoothed" galaxies curves over time, and measured how much the clustering changed when we would first apply Functional Principal Component Analysis (FPCA) to our data. FDA has been applied to astronomical data before by [4], but as far as we could tell it has not been applied to the Illustris-TNG data. Thus, the application of FDA to this data is both novel and provides another much needed example to the growing FDA literature in scientific domains.

From an astronomical perspective, applying these unsupervised clustering techniques to simulated galaxy data can help identify distinct groupings of galaxies that may have otherwise been missed or distorted by traditional classification methods. For example, Du et al. applied their 'auto-GMM" algorithm to Illustris-TNG data in an effort to remove human bias in identifying the number of structures in a galaxy [3]. That is, though there has been methods of classifying galaxies, such as the Hubble sequence which may classify a galaxy as "elliptical" or "irregular" for example, by focusing only on the data we hope to gain some novel insight into the fundamental structure of the galaxies [1]. As a result, by relying solely on the data, we aim to reveal structures that may be more meaningful from a statistical standpoint.

2 Methodology

2.1 Data Access

To begin, we were granted access to the Illustris-TNG JupyterLab, which provided convenient access to the TNG-50 data along with tutorials on how to use it [5]. The TNG website also hosted a wealth of documentation and articles that proved invaluable in getting our analysis started. Various galaxy properties, such as stellar mass, dark matter mass, gas mass, and star formation rate, were then compiled into a data frame for ease of access at redshift $z = 0$ (representing present-day galaxy attributes). For a more detailed analysis of this dataset, we refer the reader to "Statistical Inference of Galaxy Populations" (Scott Holtshousen's paper). To investigate galaxy evolution, via time-series data, we identified the 50 most massive galaxies based on their stellar mass at $z = 0$. For each of these galaxies, we retrieved the relevant attributes at every fifth snapshot throughout the simulation, resulting in a time series with 20 data points per galaxy.

Our primary focus became the star formation rate over time, as we thought it offered an interesting, nuanced view of each galaxy's evolution.

2.2 Unsupervised Learning Techniques

In general, clustering techniques seek to find a structure of "natural groupings" in unlabeled data based on a similarity or distance measure. In contrast to a classification problem, where there is a known number of groups, clustering techniques make no prior assumption about the number of groups. The fact that there is no labels is why such methods are often referred to as "unsupervised learning" techniques. We employed different "clustering" techniques throughout this process to analyze the galaxy data, the main ones we describe below.

2.2.1 Clustering Methods

To compare how different clustering algorithms would work on our dataset, we first employed agglomerative hierarchical clustering. The algorithm initially takes each data point to be its own cluster and at each step "joins" the two points that are closest, or most "similar", to each other based off of some arbitrary distance metric. In our case we used the Ward's distance which optimizes for lower within cluster variance [11]. The result is a "dendrogram", which visualizes how the points were merged at each step, and the corresponding increase in similarity.

Our other two main clustering tools were Gaussian mixture models (GMM's) and K-means as introduced in the EM lecture notes from class. K-means seeks to classify each of the data into K clusters. The objective is to minimize:

$$\min \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik} ||y_i - \mu_k||_2^2$$

where $q_{ik} \in [0, 1]$. The iterative algorithm, that we learned in class, to estimate q_{ik} and μ_k is as follows:

We have data $y_1, \ldots, y_n$ iid from some exponential family distribution; number of clusters K; initial means $\mu_1^{(0)}, \ldots, \mu_K^{(0)}$; parameter $\beta > 0$. The algorithm is as follows:

While not converged:

- **Update probability:**

$$q_{ik}^{(t+1)} = \frac{\exp(-\beta \|y_i - \mu_k^{(t)}\|^2)}{\sum_{j=1}^{K} \exp(-\beta \|y_i - \mu_j^{(t)}\|^2)}$$

- **Update means:**

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t+1)} y_i}{\sum_{i=1}^{n} q_{ik}^{(t+1)}}$$

In the GMM framework, a $K-$component gaussian mixture model is:

$$f(y|\theta) = \sum_{i=1}^{K} \rho_k \cdot g(y|\eta_k)$$

where the mixture weight are $\rho_k \geq 0$, $\sum_{i=1}^{K} \rho_k = 1$, and $g(y|\eta_k)$ is a gaussian density with some parameters defined by η_k. These parameters can be estimated using the Expectation-Maximization algorithm as seen in class, which is a generalization of the K-means algorithm seen above. The algorithm is run until we the parameters of interest converge sufficiently.

2.2.2 Principal Component Analysis

Principal component analysis (PCA) is a common dimension reduction technique used in unsupervised learning settings. Partly inspired by the work of Cohn and van de Voort [2], we thought it would be a good idea to examine how much our clustering results change before and after applying PCA.

In short, PCA seeks to reduce the dimensionality of the data by finding the greatest directions or modes of variation. This is done by first computing the covariance matrix of the data, performing the eigen-decomposition, and taking the first M eigenvectors corresponding to the largest M eigenvalues to be your principal components [14]. The choice of M can be determined by how much of the variation you want to be explained by your components. This is particularly useful in our case as we have many, often correlated, measurements for each galaxy.

2.3 Silhouette Score

As our main concern is clustering, we naturally need some metric to evaluate how well the various algorithms are performing. One widely used metric for this purpose is the silhouette score, introduced by Peter J. Rousseeuw in 1987 [9]. Suppose observation i is assigned to cluster A, then the silhouette coefficient $s(i)$ is given by:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Where $a(i)$ is the average distance between i and all objects in A and $b(i)$ represents the minimum average distance between object i and all the objects in some other cluster that is not A. We can then take the mean silhouette coefficient of all our observations, called the silhouette score, to be a measure of how well a given clustering algorithm is performing. The score can range between -1 and 1, with 1 being the best possible score, -1 the worst, and 0 indicating possible overlapping clusters.

3 Clustering Snapshot Data

We were first interested in finding meaningful galaxies clusters at snapshot z = 0. After some initial data cleaning, the dataset ended up having 7090 galaxies and included attributes such as black hole mass, accretion rate onto the central black hole, various metal fractions, and star formation rate. For a detailed overview of the data please refer to Table 2 in the Appendix.

Given the large number of measured galaxy properties, the application of PCA seemed both natural and useful. The obvious lack of physical interpretation of PCs notwithstanding, we focus for now on the statistical viewpoint. We wanted to first have enough PCs to be able to explain at least 90% of the variance. Table 1 shows how the explained variance increases with each additional principal component.

Table 1: Explained variance ratios for the first seven principal components and their cumulative contribution.

Principal Component	Explained Variance	Cumulative Variance
1	0.43	0.43
2	0.18	0.61
3	0.17	0.78
4	0.05	0.83
5	0.03	0.87
6	0.03	0.90
7	0.03	0.93

We then fit GMMs with varying numbers of components and applied hierarchical clustering to the dimensionally reduced dataset. Figure 1 displays the silhouette scores for the GMMs across different numbers of clusters, along with the hierarchical clustering resulting dendrogram.

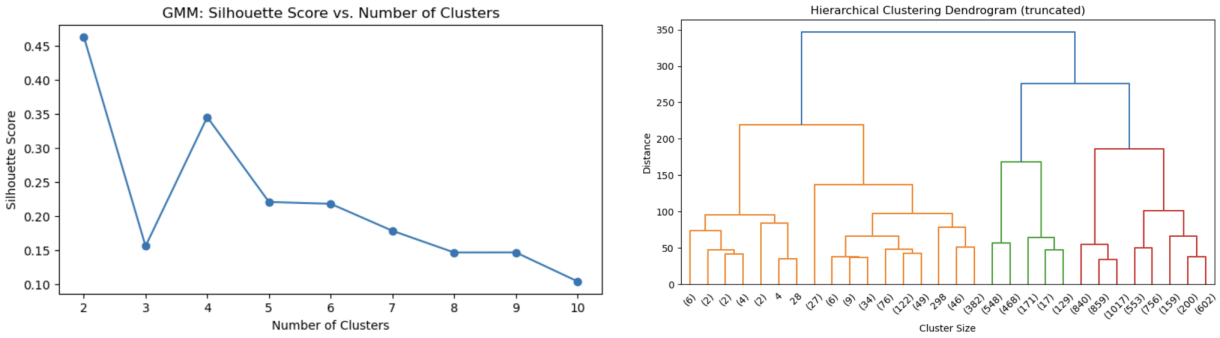


Figure 1: Results of Clustering on Reduced data

The GMM contains with 2 clusters obtains the best silhouette score, 0.46, while the 4-cluster GMM also seems to be doing a reasonable well. There is a clear decreases in cluster purity as we increase the number of specified clusters. The dendrogram sees its largest increase in within node similarity after splitting the data into two groups, though 4 groups also appears reasonable. Figure 2 shows the results of clustering the data into two groups based off of both the GMM and hierarchical clustering algorithm.

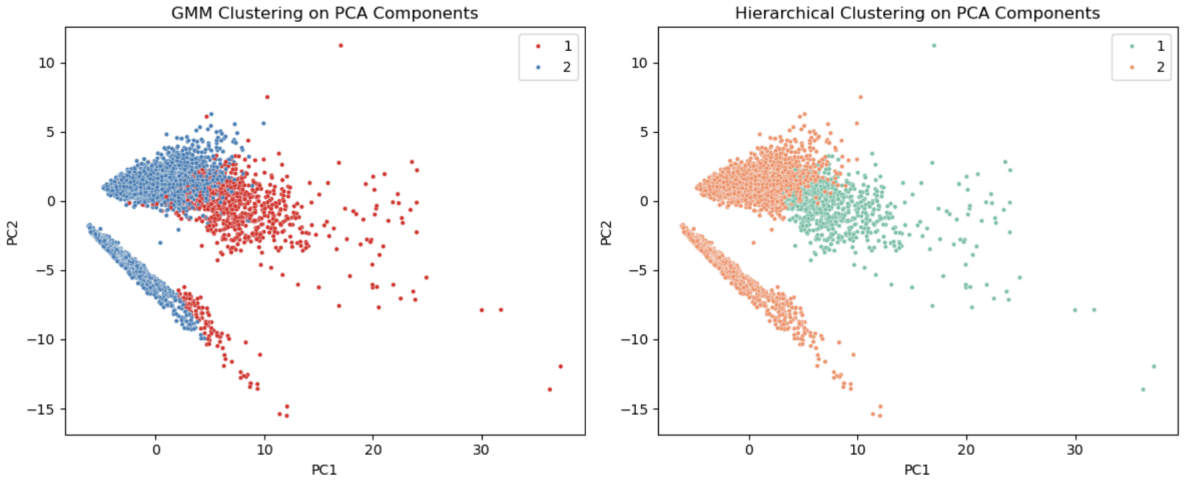


Figure 2: Results of 2-Component Clustering on Reduced data

The results are hard to interpret. PC1 explains far more variation in the data than does PC2, so the change in clustering as we move along PC1 appears reasonable. However, another natural grouping appears to be between the "disk" like shape in the bottom right of the plot and the "bulge" like formation more in the center. However, by reading too much into the data we risk committing the "bias" we set out to avoid. It appears that the two algorithms are overall in general agreement.

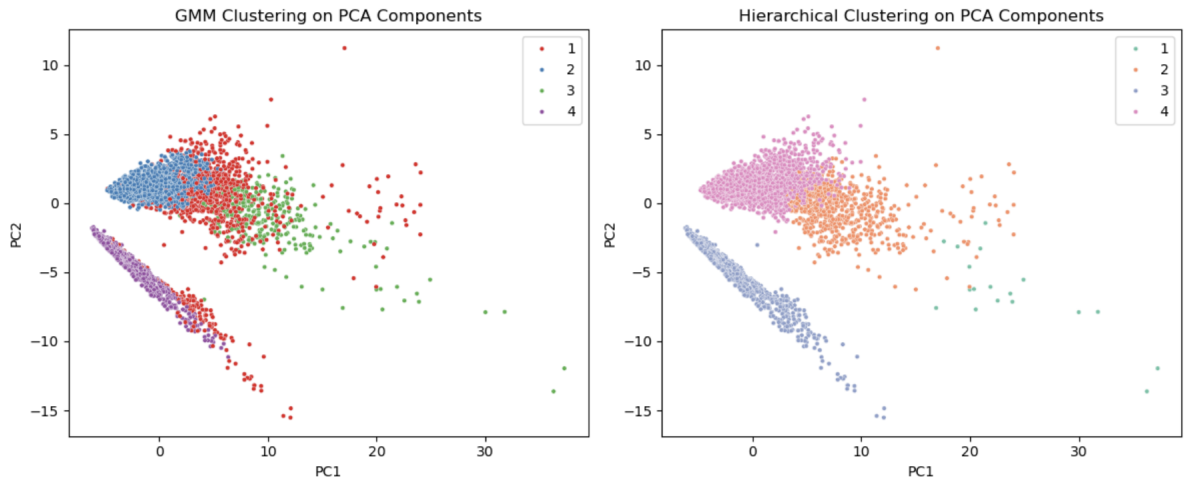


Figure 3: Results of 4-Component Clustering on Reduced data

Figure 3 shows the results of clustering the data into 4 groups with both the GMM and hierarchical clustering algorithm. We once again see that PC1 dominates the cluster labeling, as expected, but both algorithms appear to agree with the "natural" groupings discussed above. In fact, the hierarchical clustering does a great job of clustering the disk and bulge shape mentioned earlier, with the remaining "outlier" like points in a group of their own.

4 Clustering Time Series Data

4.1 Functional Data Analysis

To explore how galaxies develop over time, we turn to Functional Data Analysis. Functional data analysis (FDA) can be summarized as the study of information available in curves, surfaces, or anything else that can be thought of as a "function". It can be thought of as a generalization of multivariate statistics, where the phenomenon in question is intrinsically "infinite dimensional", hence the use of "functional". Ramsay states, "functional data analysis is to think of observed data functions as single entities, rather than merely as a sequence of individual observations" [8].

Though we often revert back to common multivariate methods to study the data, FDA makes a conceptual claim that is useful when working with certain data types. For example, in our case, we can think of the increasing mass of the galaxy over time as a realization of a "random curve", rather than simply a $N-$dimensional data point. There is no "discreteness" in the growth of galaxy, and the process can naturally be thought of as a continuous occurrence. Further, if we were to simply take our N time measurements for a given galaxy as a data point, all the measurements could be highly correlated and classical multivariate techniques become unsuitable. Moving forward, we rely heavily on the techniques and functions available in the *scikit-fda* package[7].

For our example, we take 47 of the largest galaxies available in our dataset and extract the stellar mass of them at every 5th step. That is, we have 47 "functional" data points, corresponding to the stellar growth curves of each galaxy. To make our data "functional", we need some way to make it continuous and therefore measurable at any arbitrary point. To do this we interpolate, or "smooth" our data $y_{i1}, \ldots, y_{in}$, from the i^{th} function $X_i(t)$ using some arbitrary amount of basis function.

In general, we can assume our function can be stated as:

$$X(t) = \sum_{i=1}^{\infty} a_i \phi_i(t)$$

for some basis function ϕ and coefficients a. What we do to begin the FDA process is to pick some arbitrary K to smooth the curves such that:

$$\hat{X}(t) = \sum_{i=1}^{K} a_i \phi_i(t)$$

which still does a good job of capturing the essence of the curve. In our case, we decide to use the computationally efficient B-splines as the basis function, and found that 10 terms did a reasonable job of smoothing the curve.

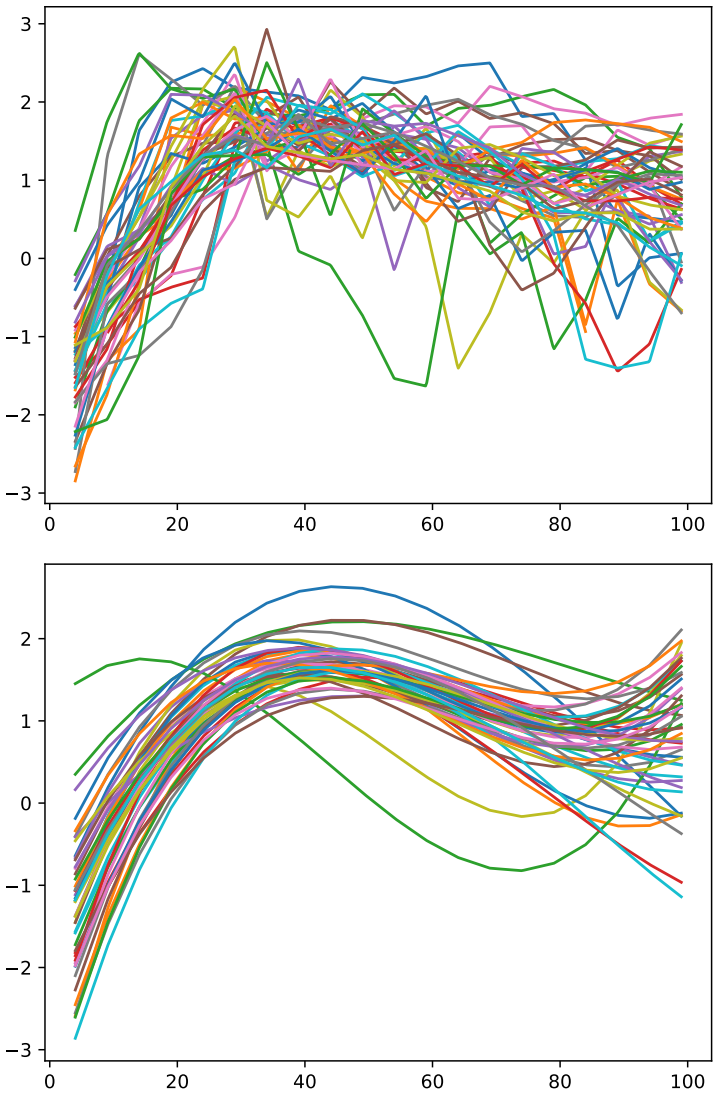


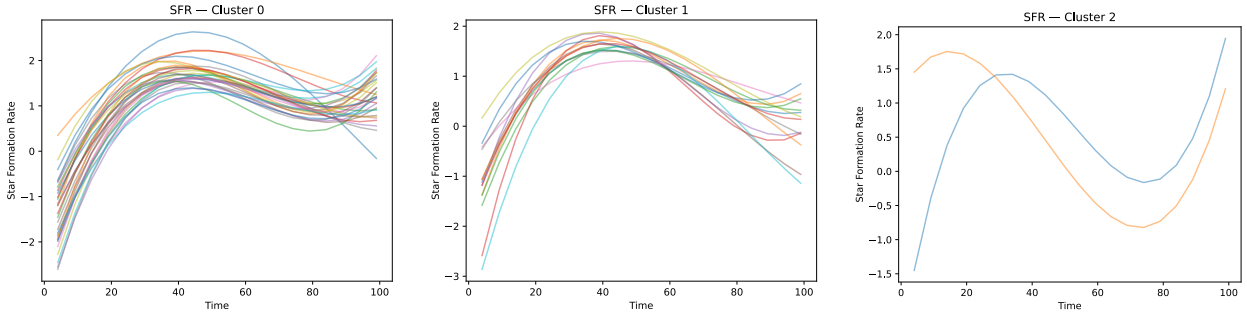
Figure 4: Raw and Smoothed Growth Curves

Figure 4 shows the raw star formation growth rate curves and shows the smooth curve from 10 B-spline basis functions.

We can now apply the K-means algorithm to our functional data. Functional K-Means in scikit-fda extends the classic algorithm to curves by iteratively assigning each function to the nearest centroid, using the functionally relevant L^2 norm:

$$\int |\phi(x)|^2 dx$$

for some function $\phi(x)$ [12].



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

Figure 5: Smoothed functional data for each cluster.

We apply 3-means clustering to the functional data and get the results in Figure 5. The clustering appears to be doing a reasonable job. Galaxies with high star-formation growth rates are concentrated in cluster 2; two outliers stand alone, while the remaining galaxies form their own distinct cluster 1. The distinction becomes more clear when we plot the curves on top of each other as seen in Figure 6.

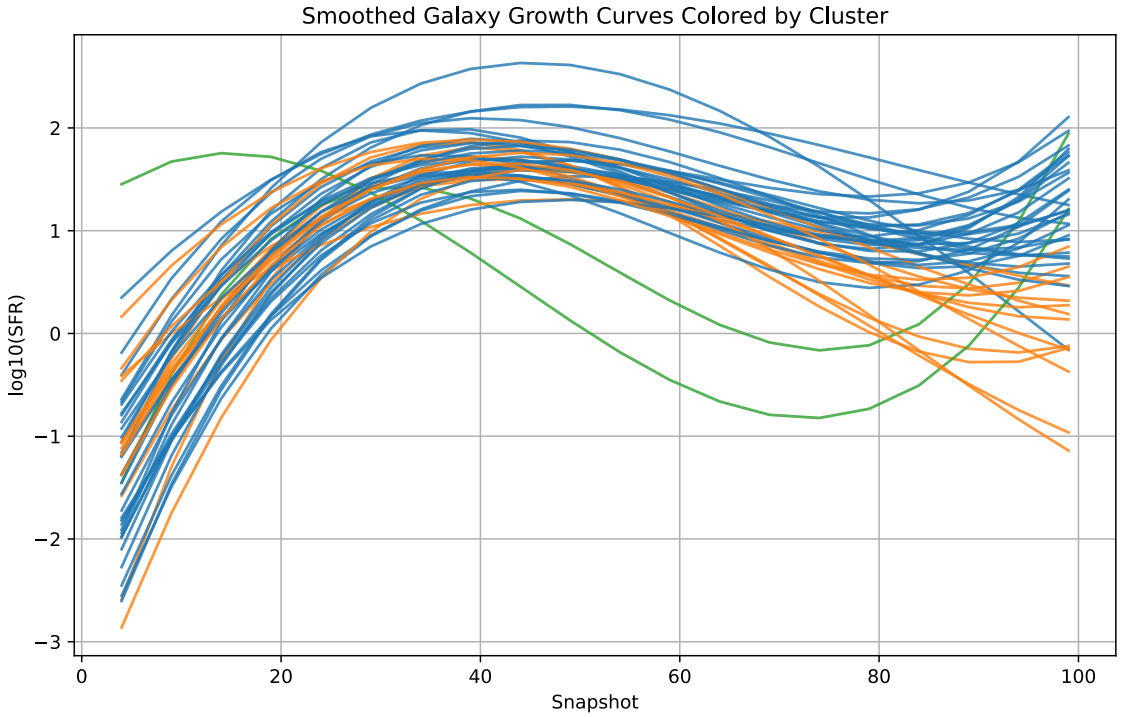


Figure 6: Growth Curve Clusters

4.2 Functional Principal Component Analysis

Functional principal component analysis works much in the same way as the classical principal component analysis implemented above. A functional principal component is associated with an eigenfunction of the auto covariance function, representing a dominant mode of variation among the observed random functions. In short, we hope to write some function $X_i(t)$, in a reduced way as:

$$X_{iK} = \mu(t) + \sum_{k=1}^{K} A_{ik}\phi_k(t) \tag{1}$$

where $\phi(t)$ is an eigenfunction and A_{ik} is the functional principal component score. Please see the Appendix for more details on this notation, derivation, and for a more detailed cited reference please see [10] or [13] . Truncating to the first K functional principal components reduces infinite-dimensional functional observations to a low-dimensional set of scores while retaining most information. In our case we want to reduce the 10 basis functions to something more manageable such as 2, and see if there is an improvement in our ability to cluster.

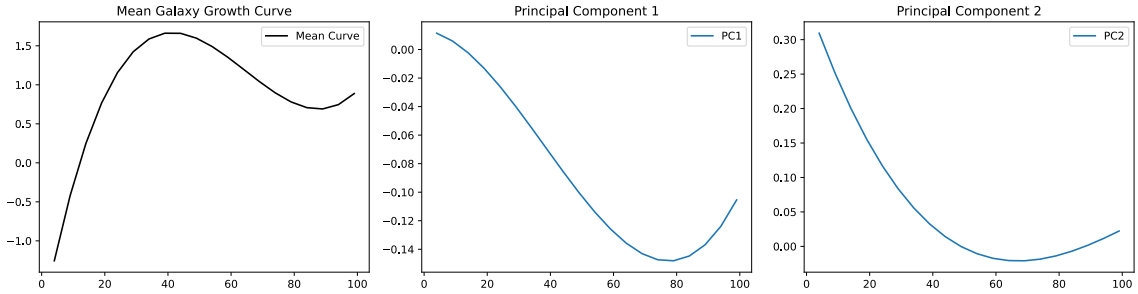


Figure 7: FPCA decomposition

In general, FPCA allows us to explain most of the variation in the function of interest with some K fixed basis functions. As can be seen from Figure 8, only 2 principal components was needed to explain 0.765 percent of the variance.

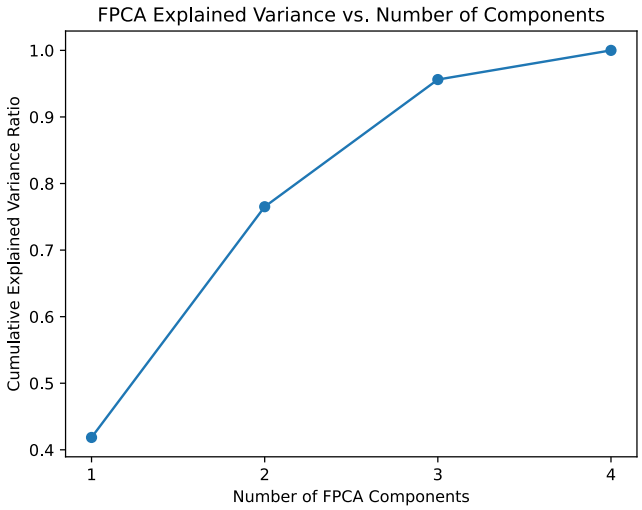


Figure 8: Explained Variation for FPCA

Once we run the functional principal component algorithm, the original smoothed growth curves can be reconstructed in a much more parsimonious manner. Figure 9 shows the reconstruction of two of the star formation growth curves using Equation 2 in [10]. It is clear that the curve reconstructions do a fairly reasonable job of capturing the original curvature, though both struggle towards at the ends of the curve.

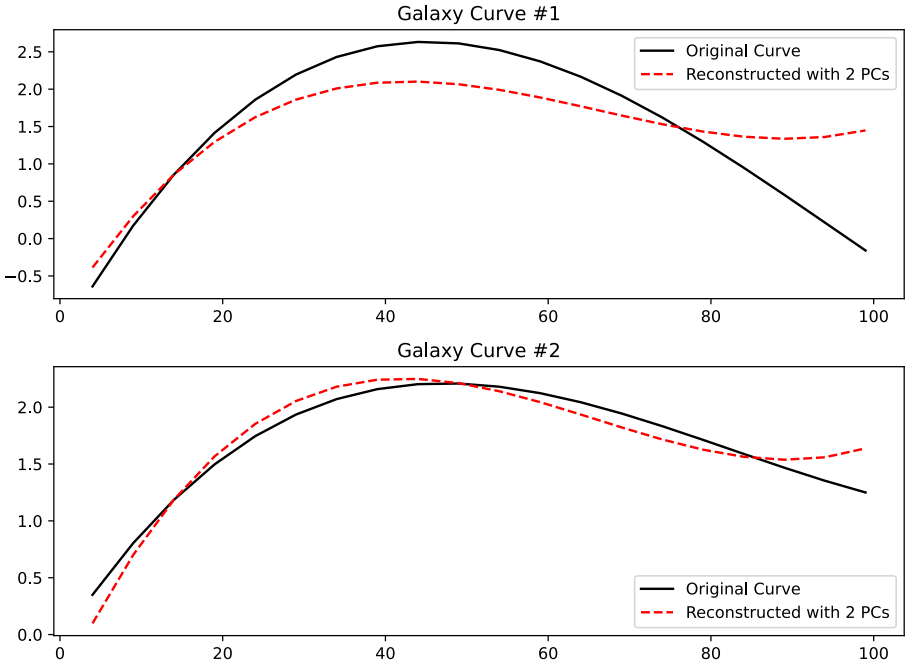


Figure 9: Curve Reconstructions from FPCA

And we return to our original task of clustering the galaxies. We now apply 3-means to the 2 FPC components curve that were constructed above, using the same algorithm that gave us Figure 6.

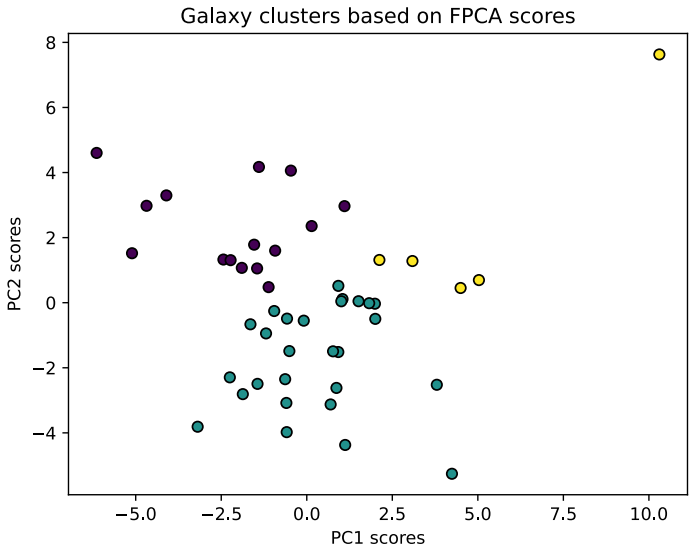


Figure 10: Clustering Results using FPCA

K-means before and after applying FPCA agreed on approximately 0.60 percent of the labels. The most obvious curves that they agreed on were the "outlier" curves that you can see in Figure 6, though the K-means on the FPCA data added a few more observations to that cluster. Furthermore, when we measure the silhouette score of each clustering procedure, the raw 3-means gets a score of 0.231, while the FPCA clustering gets an improved score of 0.380.

5 Discussion

We have given a broad overview of unsupervised clustering techniques applied to both static and temporal galaxy simulation data from Illustris-TNG. First we apply hierarchical clustering methods and GMMs to static data at z = 0, and study the effects of dimension reduction, through PCA, on the data. Next we attempt to study galaxy star formation rate through time, using methods from FDA. We generalize many of the methods used in the static, finite dimensional case, to the functional case. For example, generalizing PCA by FPCA. These FDA examples provide a unique way to study galaxy growth curves that has not been fully utilized in the literature. Further, it provides another case study in the applications of FDA, a field that has tended to recycle the same examples, to the current literature.

While our analysis may have been statistically interesting, a number of important limitations remain. There are two main drawbacks to the work conducted above. One is the lack of physical interpretation to the clusters. Though we have found statistical interesting ways to cluster the data, it is unclear if this is useful to the broader astrophysical community. In future work, it would be interesting to see how the clusters assigned to each galaxy in our work compares with how the galaxy would be clustered using the Hubble sequence for example. Furthermore, though all the individual properties of the galaxies can be understood on their own, once PCA is applied, it becomes increasingly confusing how the galaxies were actually clustered in regards to their physical attributes.

The other issue, or rather concern, is the not overly impressive silhouette scores. Though we did see some dramatic improvements, from 0.231 to 0.380, after applying FPCA to the functional data above, a silhouette score below 0.5 is generally considered "weak" [15]. The same applies to the clustering done on the static data above which achieved a score of 0.46. Though these results are slightly underwhelming, we should keep in mind that for both the static and time series data we relied on a comparatively small amount of available data. Perhaps in future work, by harnessing more modern clustering techniques such as neural networks for example, more impressive results can be obtained.

6 Author Contribution

This project was done by Scott Holtshousen and I. Scott came up with the original idea of working on astronomical data and found the Illustris-TNG dataset. He provided invaluable work in getting the Jupyter notebooks for us, and cleaning the data into a manageable format. Not to mention the necessary physics information he was able to provide. After that, he was mainly focused on the snapshot results of this paper and we collaborated frequently throughout the process. I have provided some of my interpretation of his results above, but for a more exhaustive discussion please see his papers. I mainly focused on what statistical applications could be applied to the dataset, and the interpretation of the results. Clustering seemed natural for both the snapshot and time series data, not to mention the obvious astronomical significance. I thought the application of FDA could be interesting as a way to study how different attributes of a galaxy change with time. It is also relevant for my own research, so I was eager to apply it.

Data Availability

For extremely well-documented details on the dataset used in this report we refer the reader to the IllustrisTNG homepage and give credit to Dylan Nelson for the excellent work that he has done making the data accessible.

References

[1] James Binney and Michael Merrifield. *Galactic Astronomy*. Princeton, NJ: Princeton University Press, 1998. ISBN: 9780691004028.

[2] J. D. Cohn and Freeke van de Voort. "Characterizing simulated galaxy stellar mass histories". In: *Monthly Notices of the Royal Astronomical Society* 446.4 (2015), pp. 3253–3263. DOI: 10.1093/mnras/stu2293. URL: https://academic.oup.com/mnras/article/446/4/3253/2892469.

[3] Min Du et al. "Identifying Kinematic Structures in Simulated Galaxies Using Unsupervised Machine Learning". In: *The Astrophysical Journal* 880.2 (2019), p. 121. DOI: 10.3847/1538-4357/ab2f73. URL: https://arxiv.org/abs/1909.06063.

[4] Thomas Loredo et al. "Bayesian Functional Data Analysis in Astronomy". In: *arXiv preprint arXiv:2408.14466* (2024). URL: https://arxiv.org/abs/2408.14466.

[5] D. Nelson et al. "First results from the TNG50 simulation: galactic outflows driven by supernovae and black hole feedback". In: *Monthly Notices of the Royal Astronomical Society* 490.3 (2019), pp. 3234–3261. DOI: 10.1093/mnras/stz2306.

[6] D. Nelson et al. "The IllustrisTNG simulations: Public data release". In: *Computational Astrophysics and Cosmology* 6.1 (2019), p. 2. DOI: 10.1186/s40668-019-0028-x.

[7] Carlos Ramos-Carreño et al. "scikit-fda: A Python Package for Functional Data Analysis". In: *Journal of Statistical Software* 109.2 (May 2024), pp. 1–37. DOI: 10.18637/jss.v109.i02. URL: https://www.jstatsoft.org/article/view/v109i02.

[8] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. 2nd. Springer Series in Statistics. Springer, 2005. ISBN: 9780387400808. URL: https://link.springer.com/book/10.1007/b98888.

[9] Peter J. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20.1 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[10] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. "Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3 (2016), pp. 257–295. DOI: 10.1146/annurev-statistics-041715-033624.

[11] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

[12] Eric W. Weisstein. *$L\hat{2}$-Norm*. From MathWorld–A Wolfram Web Resource. Last updated: April 8, 2025; accessed: April 10, 2025. 2025. URL: https://mathworld.wolfram.com/L2-Norm.html.

[13] Wikipedia contributors. *Functional principal component analysis — Wikipedia, The Free Encyclopedia*. [Online; accessed 11-April-2025; last edited 14 August 2024]. 2024.

[14] Wikipedia contributors. *Principal component analysis*. https://en.wikipedia.org/wiki/Principal_component_analysis. Accessed: 2025-04-12. n.d.

[15] Wikipedia contributors. *Silhouette (clustering)*. Accessed: 2025-04-16. n.d. URL: https://en.wikipedia.org/wiki/Silhouette_(clustering).

7 Appendix

7.1 Snapshot Data description

Table 2 describes all fields used in the snapshot clustering section of the paper.

Table 2: Description of fields included in the extended dataset and their corresponding physical interpretations.

Field Name	Description
SubhaloBHMass	Mass of the central black hole associated with the subhalo.
SubhaloBHMdot	Accretion rate onto the central black hole.
SubhaloFlag	Boolean indicator identifying valid, gravitationally bound subhalos (i.e., real galaxies).
SubhaloGasMetalFractionsHalfRad	Metal fractions (by element) in gas within the subhalo's half-mass radius.
SubhaloHalfmassRad	Radius enclosing half of the total mass of the subhalo.
SubhaloMassInHalfRad	Total mass contained within the subhalo's half-mass radius.
SubhaloMassType	Mass contributions by particle type (gas, stars, dark matter, black holes) in the subhalo.
SubhaloSFRinHalfRad	Star formation rate measured within the subhalo's half-mass radius.
SubhaloSpin	Angular momentum vector of the subhalo, indicating its rotational properties.
SubhaloStarMetalFractionsHalfRad	Metal fractions (by element) in stellar particles within the half-mass radius.
SubhaloStarMetallicityHalfRad	Average stellar metallicity within the subhalo's half-mass radius.
SubhaloStellarPhotometrics	Synthetic photometric magnitudes computed from stellar populations (e.g., SDSS bands).
SubhaloVelDisp	Stellar velocity dispersion within the subhalo.
SubhaloVmax	Maximum circular velocity of the subhalo, often used as a proxy for halo mass.

7.2 Functional principal component derivation

We summarize the FPCA derivation given in Wang [10], in section 2.3. Please reference it for a much more detailed outline.

For a given functions $X_i(t)$ and $i = 1, \dots n$ we have an estimated mean function:

$$\mu(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t)$$

and auto covariance function

$$\Sigma(s,t) = \frac{1}{n} \sum_{i=1}^{n} [X_i(s) - \mu(s)][X_i(t) - \mu(t)]$$

and it can be shown that we can write the covariance of Σ as:

$$\Sigma(s,t) = \sum_{i=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$$

where $\phi(t)$ is an eigenfunction. This now allows us to write some $X_i(t)$ as:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{i,k}\phi_k(t)$$

by Karhuenen 1946 and Loève 1946, where the $A_{i,k}$ are the FPC's of $X_i(t)$ or, simply, the scores.

7.3 Other Simulation Plots

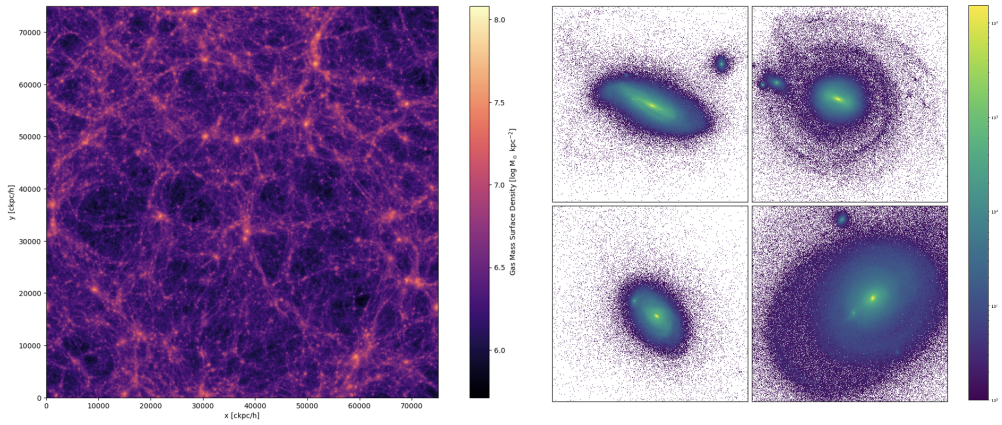


Figure 11: Simulated galaxies from Illustris TNG tutorial

7.4 Code

The notebook we used is appended below