A Machine's Perspective on Galactic Environments: Dissecting the Galaxy Distributions with Variational Autoencoders

Eun Taek Gim[1,2], Jun-Sung Moon[3], Suk-Jin Yoon[1,2]

[1]Deparment of Astronomy, Yonsei University, Seoul, 03722, Republic of Korea

[2]Center for Galaxy Evolution Research, Yonsei University, Seoul, 03722, Republic of Korea

[3]Astronomy Program, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Republic of Korea

Abstract

The environment where galaxies reside is one of the most influential factors that affect galaxy evolution in various ways. Previous studies have found that galaxy properties, including morphology and star formation activity, vary depending on the galactic environments, which are often characterized by quantities, such as the projected surface densities or the distances to the nearby neighbors. Here, we apply an unsupervised machine learning approach to ~67,000 galaxies in a volume-limited sample from the Sloan Digital Sky Survey to extract the key features describing the local environments around the galaxies. Specifically, we use a type of neural network model called the variational autoencoder (VAE), and the model is trained using a set of galaxy distribution maps centered on each target. We find that the galactic environments can be mapped onto the low-dimensional latent space, and the original input distributions can be reconstructed by using only 128 latent variables corresponding to distinct environmental features. While the most informative latent variable is related to the overall number density, other latent variables represent various anisotropic and asymmetric structures oriented along different directions. We also examine the correlations between latent variables and galaxy properties using Spearman's correlation coefficient, mutual information, and Kolmogorov–Smirnov statistics. Interestingly, some latent variables correlates, albeit mildly, with specific galaxy properties, indicating that each environmental feature plays a distinctive role in shaping the galaxy properties. Our results suggest that the VAE can be useful for recognizing and distinguishing meaningful environmental features that influence galaxy properties.

Data

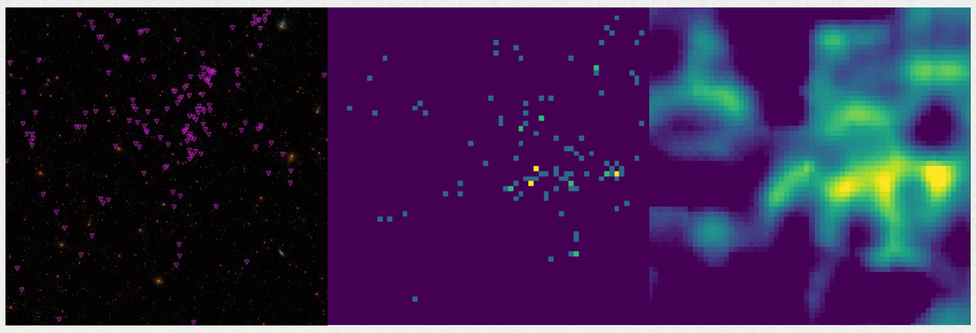

Fig 1. An example of the input data. The left panel shows the galaxy distribution in the sky, and the middle panel shows the converted input image. The right panel shows the reconstructed image after the training is over.

- Target galaxies are selected from a volume-limited sample of the SDSS DR7 (Abazajian et al. 2009) in the range $0.02 < z < 0.07$ and $Mr - 5 \log h < -18.95$.
- Each distribution map covers an area of 10 h^{-1}Mpc x 10 h^{-1}Mpc centered on a target galaxy and contains neighboring galaxies with a radial velocity difference of 1000km/s.
- The distribution maps are converted into 64-pixel x 64-pixel images. Each pixel counts the number of galaxies in it.
- Targets near the SDSS survey boundary are removed. We have ~67,000 targets in total.

Variational Autoencoder

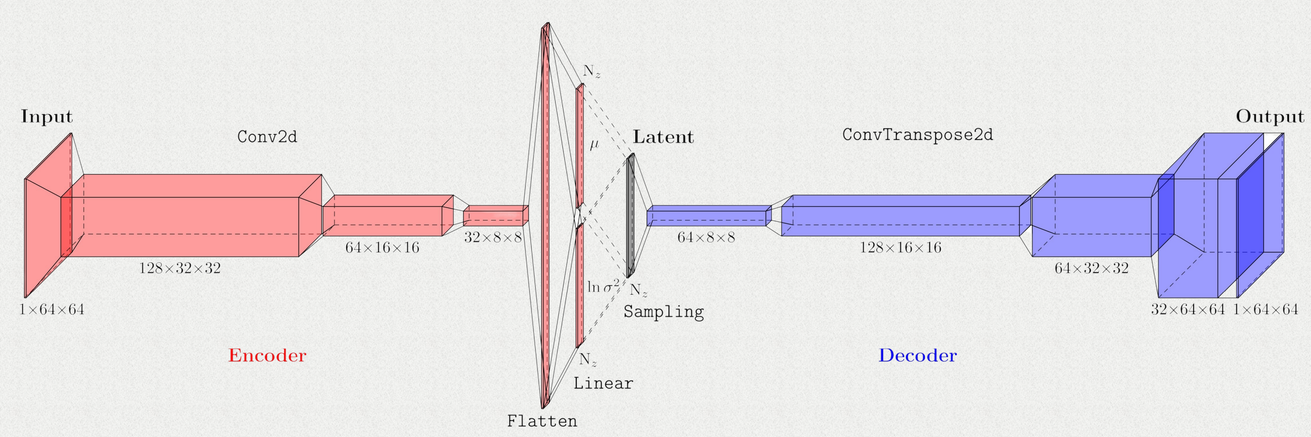

Fig 2. The VAE structure used in this study. It takes (64,64) shaped pixel images converted from the SDSS galaxy distribution. As the input data passes across layers inside the model, the VAE compresses the data into a low-dimensional latent space (Encoder). From the compressed data, the model reconstructs the input image (Decoder).

- The Variational Autoencoder (VAE) is an efficient algorithm to identify meaningful features from the input data by reducing the data dimension (e.g., Kingma & Welling 2013; Burgess et al. 2018; Portillo et al. 2020; Wei et al. 2020; Sedaghat et al. 2021).
- The encoder compresses the input image, and the extracted features are stored in a low-dimensional latent space at the end of the encoder.
- The decoder reconstructs the original image from the information in the latent space, and the reconstructed image is used to evaluate the training performance.

Results

Loss Function

- Loss = Reconstruction Loss + Regularization Loss
- Reconstruction loss evaluates the similarity between the input and reconstructed maps. We employ the Poisson loss.

$$\text{Reconstruction loss} = \hat{y} - y \log \hat{y}$$

- Regularization loss estimates the similarity between the distribution of latent variables and the normal distribution. The Kullback-Leibler Divergence (KLD) is used.

$$\mathrm{KLD}(p||q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

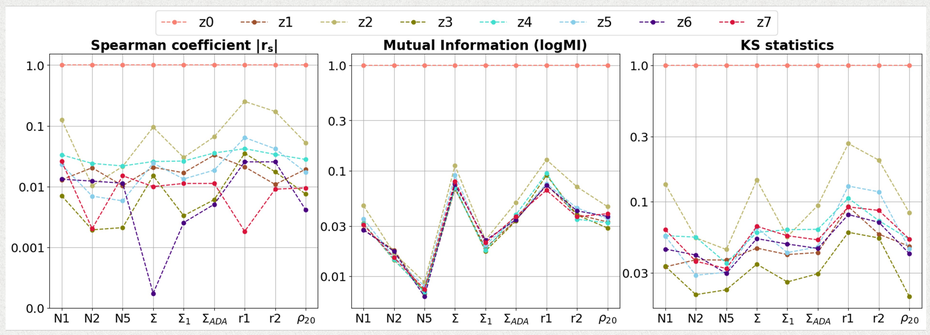

Fig 3. Spearman correlation coefficient, Mutual Information, and the KS statistics between eight most informative latent variables and the environmental parameters. All these coefficients are normalized by the value of the most informative latent variable (z0).

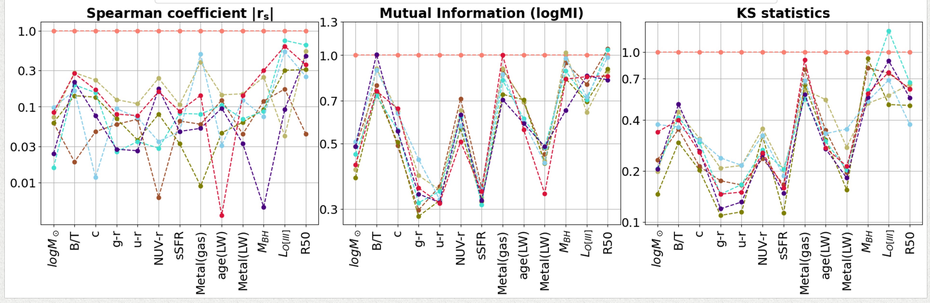

Fig 4. Same as the Fig 3, but x-axis data are galaxy properties.

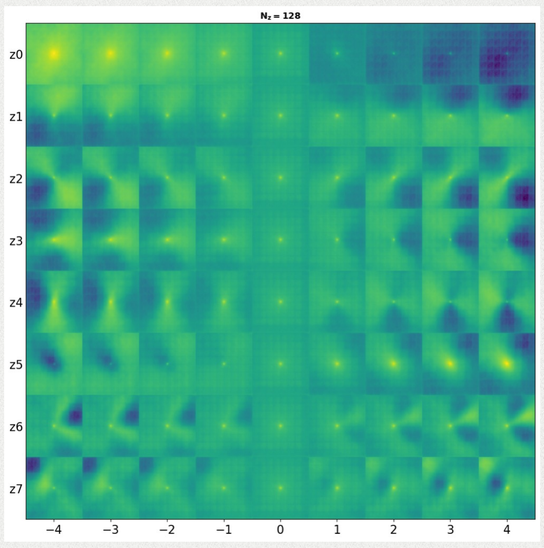

Fig 5. Extracted features using VAE with latent dimension (N_z) of 128. The latent variables are sorted in the order of informativeness. The most informative latent variable (z0) represents a feature related to number density.

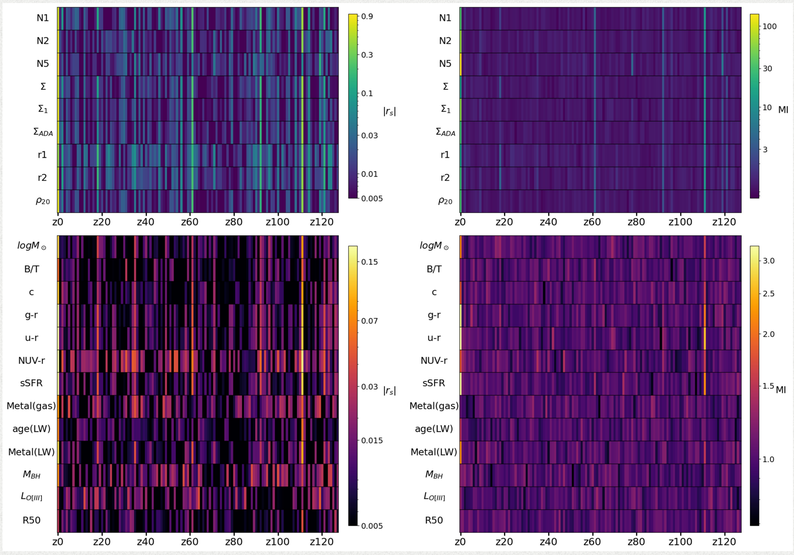

Fig 6. This figure shows the Spearman correlation coefficient and Mutual Information between latent variables and environmental parameters or galaxy properties by colors. The MI values are normalized by normalization factor.

- Mutual Information (MI, Shannon 1948; Kinney & Atwal 2014):

$$\mathrm{I}(X, Y) \equiv \mathrm{H}(X) - \mathrm{H}(X|Y)$$

where H(X) is Shannon entropy of random variable X, which defined by

$$\mathrm{H}(X) = -\sum_{x \in X} p(x) \log p(x)$$

- MI evaluates the correlation between X and Y by measuring how much "uncertainty" of X decreases when we know about Y.
- We compute the MIs from 50 randomly shuffled data and use their average value as the normalization factor (MI_{sh}) for MI.
- By normalizing the MI, we can compare how much information is contained relative to the case in which parameters are completely uncorrelated and discuss how strongly they are correlated.

Conclusion

- Defining a galaxy environment using densities or distances have a limitation in representing the distribution of galaxies within specific regions. However, using VAE to decompose the galaxy environment can reveal various components of galaxy distribution, including asymmetric structures, which were challenging to quantify with the conventional approaches. This result suggests that the machine learning techniques have significant potential in the field of galaxy environment studies.

- We find that the latent variables mildly correlate with a few galaxy properties, including star formation rates and optical/UV colors. Our result is based on the limited input data, which consists solely of the positions of neighboring galaxies and their numbers. We expect that the prediction of galaxy properties can be more accurate with complemented data in future studies.

Reference

- Abazajian, K., N. et al., 2009, ApJS, 182, 543
- Burgess, C. P. et al. 2018, arXiv: 1804.03599
- Kingma, D.P., Welling, M., 2013, arXiv: 1312.6114
- Kinney, J., B., Atwal, G., S., 2014, PANS, 111, 3354
- Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J., 2020, AJ, 160, 45
- Shannon C. E., 1948, Bell System Technical Journal, 27, 379-423, 623-656
- Sedaghat, N., Romaniello, M., Carrick, J. E., & Pineau, F.-X., 2021, MNRAS, 501, 6026
- Wei, R., Garcia et al., 2020, IEEE Access, 8, 153651