Interpretable Encoding of Galaxy Spectra

Yan Liang
Department of Astrophysical Sciences, Princeton University
Princeton NJ 08544, USA
yanliang@princeton.edu

Peter Melchior
Department of Astrophysical Sciences, Princeton University,
Center for Statistics & Machine Learning, Princeton University
Princeton NJ 08544, USA
peter.melchior@princeton.edu

Sicong Lu
Department of Physics and Astronomy, University of Pennsylvania
Philadelphia PA 19104, USA
siconglu@upenn.edu

Abstract

We present a novel loss function to train autoencoder models for galaxy spectra. Our architecture reliably captures intrinsic spectral features regardless of redshift, providing highly realistic reconstructions for SDSS galaxy spectra using as little as two latent parameters. But the interpretation of encoded parameters remains difficult because the decoding process is non-linear and the latent space can be highly degenerate: different latent positions can map to virtually indistinguishable spectra. To resolve this encoding ambiguity, we introduce a new similarity loss, which explicitly links latent-space distances to data-space distances. Minimizing the similarity loss together with the common fidelity loss leads to non-degenerate, highly accurate spectrum models that generalize over variations in noise, masking, and redshift, while providing a latent space distribution with clear separations between common and anomalous data.

1 Introduction

Spectroscopy is a critical tool to probe the physical mechanisms that drive the formation and evolution of present-day galaxies. Despite the vast amount of available galaxy spectra provided by large spectroscopic surveys, extracting physical knowledge from them is still a difficult task. Ideally, one would infer galaxy properties by directly fitting the observed spectrum to a theoretical model, but analytical models are not yet sophisticated enough to reproduce typical individual high S/N galaxy spectra, especially the strong emission lines [1, 2]. The physical processes contributing to the observed spectral features may be still poorly understood, thus using oversimplified models could lead to biased interpretation of the data.

Alternatively, one may construct a fully data-driven model via unsupervised learning. The main challenge in this approach is to properly disentangle the intrinsic physical spectra from redshift (causing a stretching of the spectra), noise level, and artifacts such as telluric contamination. Linear models, i.e. a combination of empirical or theoretical templates [3, 4], are commonly used for redshift

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

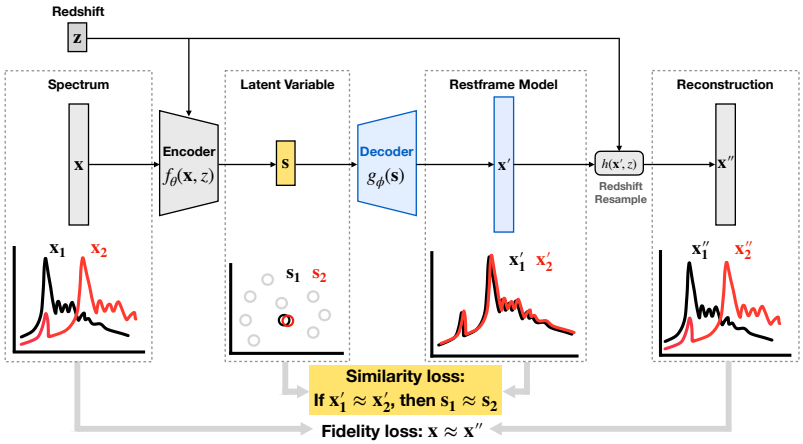

Figure 1: Our autoencoder loss function combines fidelity loss, which compares separately input spectra $(\mathbf{x}_1, \mathbf{x}_2)$ to their redshifted and resampled reconstructions $(\mathbf{x}''_1, \mathbf{x}''_2)$, with a novel similarity loss, which links the distance in latent space $|\mathbf{s}_1 - \mathbf{s}_2|$ between two spectra to their distance in restframe $|\mathbf{x}'_1 - \mathbf{x}'_2|$. Spectra of two physically similar galaxies observed at different redshifts (z_1, z_2), for which the underlying restframe models are very similar, will thus yield similar latent vectors.

estimation and spectral classification [5, 6]. The reconstruction power of linear models is limited by template quality, and often requires many components to achieve a good fit.

Autoencoders (AE) can yield models with good fidelity and small latent dimensionality [7]. But for conventional AEs, all galaxy spectra needed to be de-redshifted and resampled to a common restframe, restricting either redshift or wavelength ranges that can be probed. [8] showed that by explicitly adding a redshift transformation to the decoder path, one can utilize the entire spectrum for galaxies at all redshifts. However, the authors noted that redshift-invariant encoding was achieved only over a limited redshift range, where a set of important "consensus features" was observable in the each spectrum [9]. We now introduce a new loss term that solves this problem by relating latent space distances to distances between reconstructed restframe models. Our approach resolves degeneracies in the decoding process and establishes an encoding that is robustly invariant to changes in redshift. At no reduction in fidelity, the latent space also becomes directly interpretable—the spectra of physically similar galaxies cluster, and latents in the same neighborhood reconstruct similar-looking spectra.

2 Data

We obtain 500,000 galaxy spectra spanning redshift $z \in [0, 0.5]$ from Sloan Digital Sky Surveys Data Release 16 [10]. Our sample includes all optical spectra that are classified as galaxies and has redshift error $z_{\rm err} < 10^{-4}$. Approximately 70% of the samples are used for training, 15% for validation, and 15% is held for test. All spectra are normalized by the median flux and zero-padded to a homogeneous wavelength $\lambda_{\rm obs} = 3784...9332$Å. We mask out telluric contamination by assigning zero weights to within 5Å of the top $\sim$100 telluric lines, amounting to 12% of the data vectors.

3 Method

Our model architecture is taken from [8] (see Figure 1 for an overview and the aforementioned paper for details). Let $\mathbf{x} \in \mathbb{R}^M$ denote an input spectrum with $M = 3921$ elements. It gets encoded, by a modified version of the CNN encoder from [11], into a low-dimensional latent respresentation, $\mathbf{s} \in \mathbb{R}^2$. A standard MLP with (256, 512, 1024) nodes and a leaky ReLU activation generates a restframe spectrum $\mathbf{x}'$, whose 7000 spectral elements are chosen to create a mildly super-resolved representation with an extended wavelength coverage ($\lambda_{\rm rest} = 2359 \ldots 9332$ Å). The reconstructed spectrum $\mathbf{x}''$ is then redshifted and linearly interpolated from $\mathbf{x}'$ to the same wavelengths as $\mathbf{x}$.

This architecture allows for a redshift-invariant encoding as the reshift transformation is explicitly performed in the generator.

Our extended loss function

$$L_{\rm total} = L_{\rm fid} + L_{\rm sim} + L_{\rm c} \tag{1}$$

operates on all four of the stages shown in Figure 1. The fidelity loss term $L_{\rm fid}$ quantifies the reconstruction quality, assuming normally distributed noise. It measures the mean log-likelihood of the reconstruction of spectral elements averaged over batches of N spectra with spectral size M:

$$L_{\rm fid} = \frac{1}{2NM} \sum_{i}^{N} \mathbf{w}_i \odot (\mathbf{x}_i - \mathbf{x}_i'')^2, \tag{2}$$

where $\mathbf{x}_i$ is the i-th input spectrum, $\mathbf{w}_i$ its inverse variance, and $\odot$ the element-wise multiplication.

We define a similarity loss term that, unlike the fidelity loss, operates on the two intermediate stages. Let $\mathbf{s} = f_\theta(\mathbf{x}, z)$ be the encoded latent vector and $\mathbf{x}' = g_\phi(\mathbf{s})$ be the restframe model, where f_θ, g_ϕ are parameterized encoder and decoder functions. Ideally, if two restframe models, $\mathbf{x}_i'$ and $\mathbf{x}_j'$ are similar, their latent positions $\mathbf{s}_i$ and $\mathbf{s}_j$ should be similar as well; otherwise, the mapping from latent space to spectrum space becomes difficult to interpret. On the other hand, distinctively different models should have well-separated latent positions. While one could intuitively expect an autoencoder to establish such a relation in its latent space, we find empirically that this is often not the case. However, the desired relation can be satisfied naturally if the latent distance is proportional to a spectral distance: $|\mathbf{s}_1 - \mathbf{s}_2|^2 \propto |\mathbf{x}_1' - \mathbf{x}_2'|^2$, which motivates us to set up a loss term as follows:

$$S_{\rm ij} = \frac{1}{2}|\mathbf{s}_i - \mathbf{s}_j|^2 - \frac{1}{M}|\mathbf{w}' \odot (\mathbf{x}_i' - \mathbf{x}_j')|^2 \tag{3}$$

$$L_{\rm sim}(k_0, k_1) = \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} \mathrm{sigmoid}(\mathrm{k_1 S_{ij}} - \mathrm{k_0}) + \mathrm{sigmoid}(-\mathrm{k_1 S_{ij}} - \mathrm{k_0}) \tag{4}$$

where $\mathbf{w}'$ is the inverse-variance weight defined at restframe wavelengths, and k_0, k_1 are adjustable hyper-parameters that control the steepness of the slope. $L_{\rm sim}$ encourages pairwise latent space distances proportional to the spectra (dis)similarity defined in Equation 3. The double sigmoid function in Equation 4 serves two purposes. It limits the effect of the similarity loss by setting $L_{\rm sim} \leq 1$, such that it is only important when the fidelity loss is low $L_{\rm fid} \lesssim 1$. And it provides relatively smooth gradients (compared to e.g. $S_{\rm ij}^2$), improving the trainability of the model. It is crucial to measure the similarity between restframe pairs rather than input pairs, because the former provides a stable measure independent of redshift and observational window. In addition, restframe models usually contain less noise than the raw data (see Figure 2, bottom left panel).

Inspired by [12], we add a the consistency loss as a more guided form of the similarity loss:

$$L_{\rm c} = \frac{1}{N} \sum_{i}^{N} \mathrm{sigmoid}\left[\frac{1}{2\sigma_s^2}|\mathbf{s}_i - \mathbf{s}_{{\rm aug},i}|^2\right] - 0.5, \quad \sigma_s = 0.1, \tag{5}$$

where $\mathbf{s}_i$ and $\mathbf{s}_{{\rm aug},i}$ are the latent positions of the original and its corresponding augment spectrum, that has been redshifted by a random $z_{\rm new}$ from a uniform distribution between $[0, 0.5]$. Minimizing $L_{\rm c}$ reduces the latent distances between galaxy spectra and their augments, improving the encoding stability against redshift. The consistency loss reinforces that physically similar spectra pairs are pulled together in latent space, while dis-similar pairs are moved apart by the (dis)-similarity loss term Equation 3.

4 Results

We implement the autoencoder architecture using `pytorch`. To evaluate the impact of different loss terms, we trained three models using different strategies: In the first training, we optimize the model with $L_{\rm fid}$ alone to serve as the baseline. In the second model, we optimize $L_{\rm fid} + L_{\rm sim}$ simultaneously, with an "inverse-annealing" cycle to gradually increase the slope k_1, while k_0 is held constant. For the third model, we add the consistency regularization and optimize the full loss from Equation 1.

We train each model on a NVIDIA A100 GPU using the same training and validation data set for 700 epoch and observe convergence. The results are summarized in Table 1. The best performance

Training objective	$L_{\rm fid}$	$L_{\rm sim}$	L_c	$\lvert \mathbf{s}_{\rm i} - \mathbf{s}_{\rm aug,i} \rvert$
fidelity	0.476	-	-	2.496
fidelity + similarity	0.473	0.158	-	0.450
fidelity + similarity + consistency	0.470	0.155	0.001	0.007

Table 1: Training performance using different strategies, evaluated on the test set. $L_{\rm sim}$ is evaluated assuming $k_0 = 2.5, k_1 = 1.0$. Note that when the latent distance and spectral distance is perfectly aligned, the minimum similarity loss $L_{\rm sim}$ is 0.151.

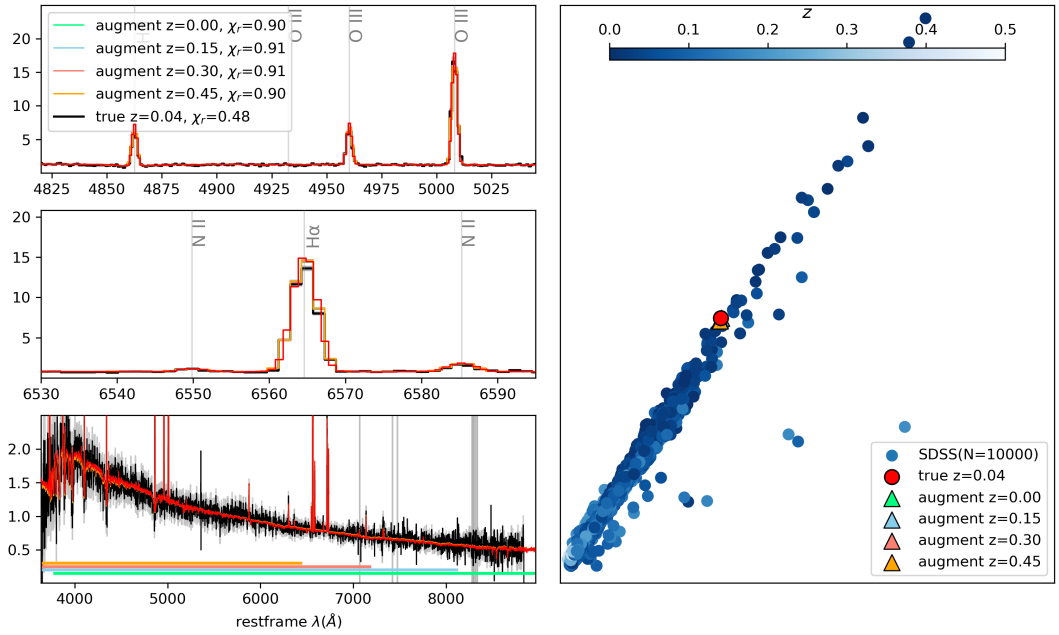

Figure 2: Left: Observed spectrum (at $z = 0.04$, black) from the test data, its reconstruction (red), and reconstructions of augmented spectra with artificially altered redshifts (color-coded). Zoomed-in versions are shown on the top. The colored-bar in the bottom-left shows the observed wave-length range of the augments. Right: 2D latent space distribution of 10,000 SDSS spectra randomly selected from the test set. The red circle marks the example spectrum, and the triangles mark its augments (essentially indistinguishable in the plot).

is achieved with all three losses combined. In particular, our best model reduces the average latent distance of augmented redshifted spectra, as a measure of redshift invariance in latent space, by a factor of 350 comparing to the fidelity-only model. This improvement has been achieved without any decrease in fidelity. On the contrary, minor gains in fidelity could indicate that the model benefits from a non-degenerate latent space, where each latent position maps to more spectra from a wider redshift range. We suggest that the consistency loss should not be used without the similarity loss because it seeks to collapse the latent distribution at the origin of the latent space. Instead, it is best added once a model trained with fidelty and similarity losses is available.

With the best model, we achieve a desirable redshift-invariant behavior. Figure 2 shows the reconstructed restframe models and encoded latent positions of an example galaxy. The latent positions of the original and redshift-augmented spectra are stably grouped together (colored markers in right panel), even for the $z = 0.45$ augment (colored in orange) where the dominant Hα emission line is outside of the observable wavelength range, i.e. unavailable to the encoder. The reconstruction quality is evidently robust to such missing features.

5 Conclusion

We introduced a novel method to establish an interpretable, redshift-invariant encoding of galaxy spectra. Our model generates highly realistic spectra for the entirety of the SDSS spectroscopic galaxy sample with only two latent parameters, making it a powerful dimensionality reduction method. We define a new loss term to relate distances in latent space to distances in restframe space. It discourages latent-space degeneracies and pulls intrinsically similar galaxies together, regardless of redshift. The decoder can thus learn a more complete underlying restframe model over the entire redshift range. In addition, the new loss term encourages an encoding that is well-suited for anomaly detection: Outliers will be pushed away from more common samples, who themselves will tend to cluster.

6 Limitations

For simplicity, we have fixed the latent space dimension to 2. This choice allows for direct visualization and does not noticeably affect the quality of the reconstructed spectra. Even though the detailed performance may differ, the usefulness of our proposed similarity loss will transparently translate to higher-dimensional latent spaces.

Because of the magnitude limit of SDSS, galaxies with $z > 0.3$ galaxies are underrepresented in our dataset. This issue can be addressed by training the autoencoder jointly with additional datasets such as the Baryon Oscillation Spectroscopic Survey (BOSS)[13].

7 Broader Impact

The similarity loss term is agnostic to the specific design of the autoencoder. We expect it to establish interpretable, non-degenerate, outlier-sensitive latent space distributions for other types of data.

References

[1] Rita Tojeiro, Alan F Heavens, Raul Jimenez, and Ben Panter. Recovering galaxy star formation and metallicity histories from spectra using vespa. *Monthly Notices of the Royal Astronomical Society*, 381(3):1252–1266, 2007.

[2] Rita Tojeiro, Will J Percival, Alan F Heavens, and Raul Jimenez. The stellar evolution of luminous red galaxies, and its dependence on colour, redshift, luminosity and modelling. *Monthly Notices of the Royal Astronomical Society*, 413(1):434–460, 2011.

[3] Michael JI Brown, John Moustakas, J-DT Smith, Elisabete Da Cunha, TH Jarrett, Masatoshi Imanishi, Lee Armus, Bernhard R Brandl, and Josh EG Peek. An atlas of galaxy spectral energy distributions from the ultraviolet to the mid-infrared. *The Astrophysical Journal Supplement Series*, 212(2):18, 2014.

[4] M Polletta, M Tajer, L Maraschi, G Trinchieri, CJ Lonsdale, L Chiappetti, S Andreon, M Pierre, O Le Fevre, G Zamorani, et al. Spectral energy distributions of hard x-ray selected active galactic nuclei in the xmm-newton medium deep survey. *The Astrophysical Journal*, 663(1):81, 2007.

[5] Ashley J Ross, Julian Bautista, Rita Tojeiro, Shadab Alam, Stephen Bailey, Etienne Burtin, Johan Comparat, Kyle S Dawson, Arnaud De Mattia, Hélion du Mas des Bourboux, et al. The completed sdss-iv extended baryon oscillation spectroscopic survey: Large-scale structure catalogues for cosmological analysis. *Monthly Notices of the Royal Astronomical Society*, 498(2):2354–2371, 2020.

[6] Zachary J Pace, Christy Tremonti, Yanmei Chen, Adam L Schaefer, Matthew A Bershady, Kyle B Westfall, Médéric Boquien, Kate Rowlands, Brett Andrews, Joel R Brownstein, et al. Resolved and integrated stellar masses in the sdss-iv/manga survey. ii. applications of pca-based stellar mass estimates. *The Astrophysical Journal*, 883(1):83, 2019.

[7] Stephen KN Portillo, John K Parejko, Jorge R Vergara, and Andrew J Connolly. Dimensionality reduction of sdss spectra with variational autoencoders. *The Astronomical Journal*, 160(1):45, 2020.

[8] Peter Melchior, Yan Liang, ChangHoon Hahn, and Andy Goulding. Autoencoding Galaxy Spectra I: Architecture. *arXiv e-prints*, page arXiv:2211.07890, November 2022.

[9] Peter Melchior, ChangHoon Hahn, and Yan Liang. Autoencoding galaxy spectra. In *ICML Workshop Machine Learning for Astrophysics*, 2022.

[10] Romina Ahumada, Carlos Allende Prieto, Andrés Almeida, Friedrich Anders, Scott F Anderson, Brett H Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, et al. The 16th data release of the sloan digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020.

[11] Joan Serrà, Santiago Pascual, and Alexandros Karatzoglou. Towards a universal neural network encoder for time series. In *CCIA*, pages 120–129, 2018.

[12] Samarth Sinha and Adji B. Dieng. Consistency regularization for variational auto-encoders. *CoRR*, abs/2105.14859, 2021.

[13] Kyle S Dawson, David J Schlegel, Christopher P Ahn, Scott F Anderson, Éric Aubourg, Stephen Bailey, Robert H Barkhouser, Julian E Bautista, Alessandra Beifiori, Andreas A Berlind, et al. The baryon oscillation spectroscopic survey of sdss-iii. *The Astronomical Journal*, 145(1):10, 2012.