A Variational Autoencoder-Based Method to Investigate Degeneracy in the Neural Correlates of Psychological Concepts

Kieran McVeigh[1], Ashutosh Singh[2], Deniz Erdogmus[2], Lisa Feldman Barrett[1,3,4] & Ajay B. Satpute[1]

[1]Department of Psychology, Northeastern University,[2]College of Engineering, Northeastern University, Boston, Massachusetts, [3]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Massachusetts [4]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Massachusetts

Introduction

- Degeneracy or multi-realizability (i.e., one-to-many mappings between neural correlates and behaviors), is a ubiquitous phenomenon in nervous systems [1- 6] yet few methods exist to map degenerate neural correlates of behaviors [c.f. 7].
- In this work we take a brain state estimation approach [2,7] which assumes there is an unobserved (or latent) brain state that generates both physiological measures of CNS activity (e.g., from fMRI or EEG) and behavioral measures (e.g., ratings of affect). Brain states should therefore be estimated with both sets of measurements
- We present a novel brain state estimation method that allows one to learn multimodal brain states estimates, while still recovering neural correlates for many to one brain behavior mappings.

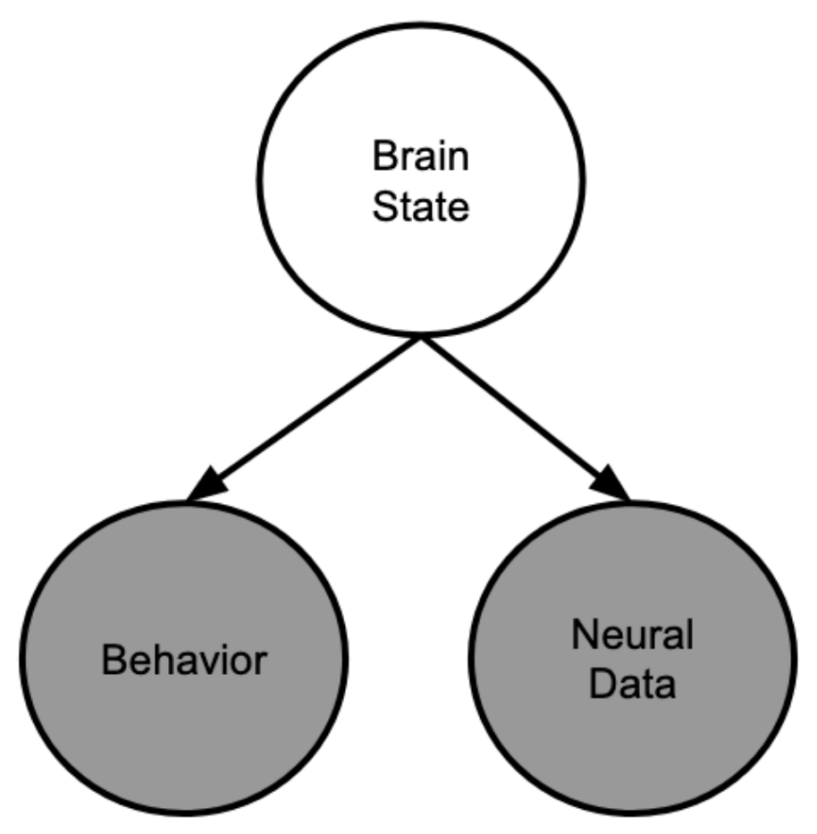


Figure 1. A probabilistic graphical model (PGM) depicting our modeling assumption that a latent brain state generates both observable behavioral and neural data

Data Generation

- To validate our method we simulated a set of data with a known ground truth
- 1000 samples were generated from four gaussian distributions in two dimensional latent "brain state space"
- We assigned samples from each gaussian to one of two class label such that classes were not linearly separable, consistent with the concept of degeneracy
- Samples were then projected from the two dimensional latent space to a 100 dimensional neural observation space via random linear projections

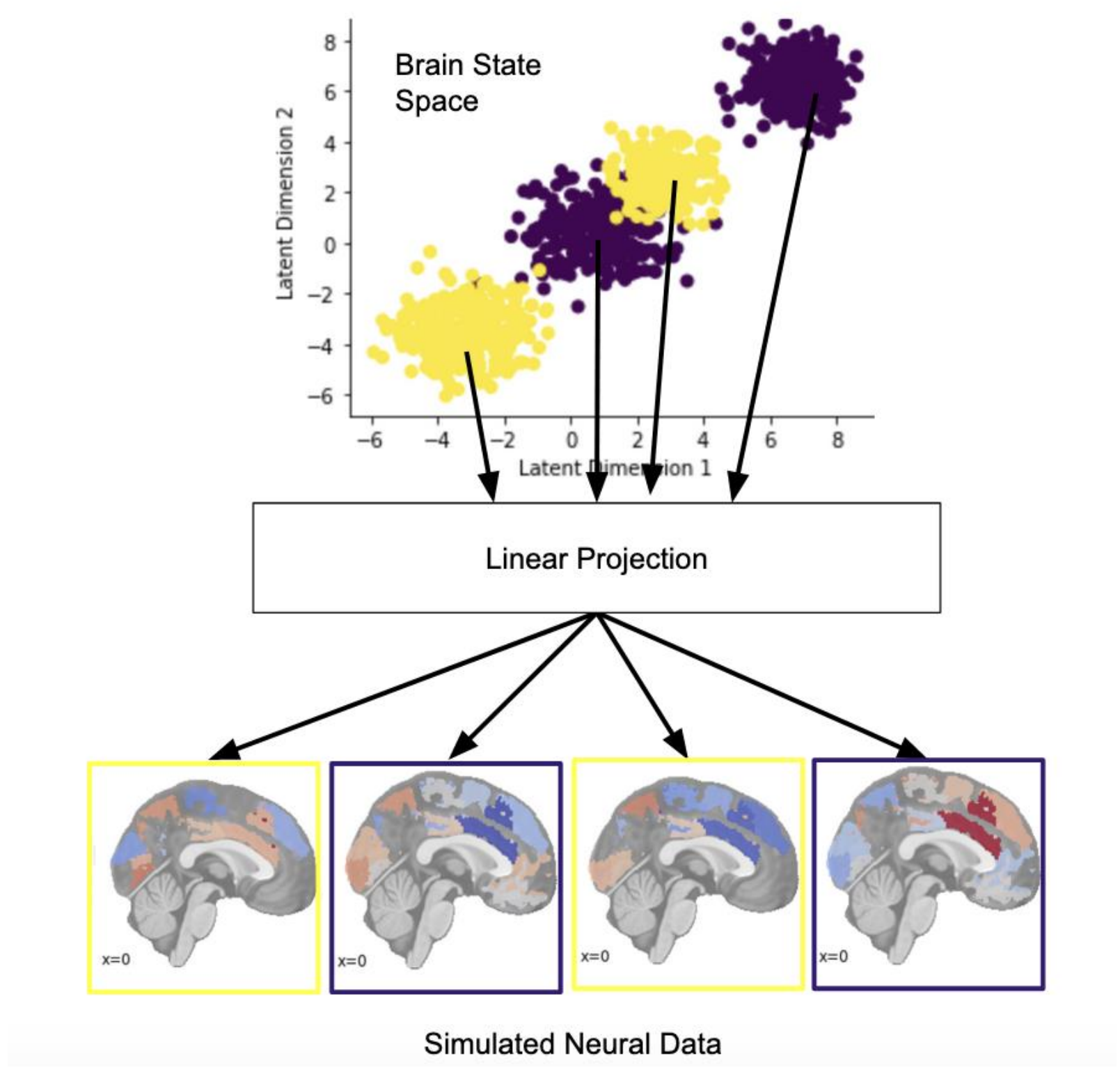
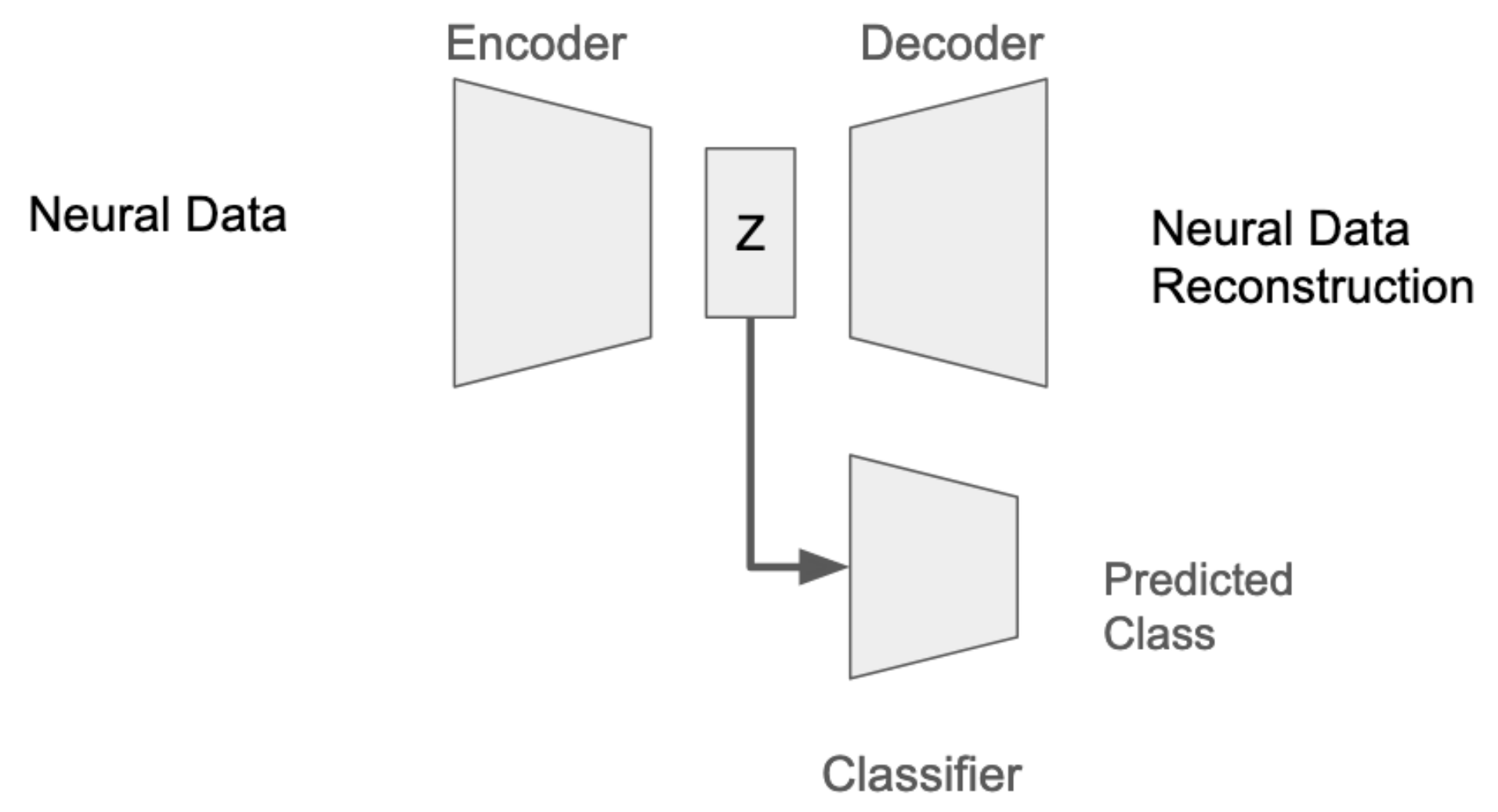


Figure 2. Above is a visual depiction of the simulated 2d latent space, and projection to the neural observation space. Different classes are indicated by different colors.

Models and Training

- Trained Variational Auto Encoder Classifiers (VAE-C) [8] to model latent brain states, and their neural and behavioral correlates.
- We systematically varied the architectures of the encoder – decoder and classifier architectures, in each variation these portions of the network were either only able to learn linear functions (i.e. no hidden layers) or were capable of learning non-linear functions.
- We also trained models with several different weights on the components of the loss functions, to examine how changes to amount of supervision influenced model performance.
- All VAE-Cs were trained for 1500 epochs and tested on 20% held out test data [per standard procedures; 9]



$$\text{Loss function} = -E_{q\phi(z|x)}[\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)\|p(z)) + \lambda(-\textstyle\sum y_i \log f(z_i))$$

$$\text{Loss function} = L_{reconstruction} + L_{KL} + \lambda L_{classification}$$

Figure 3. A diagram of the VAE-C architecture. Z is the set of neurons modeling the latent estimated brain state.

Baseline Model Performance

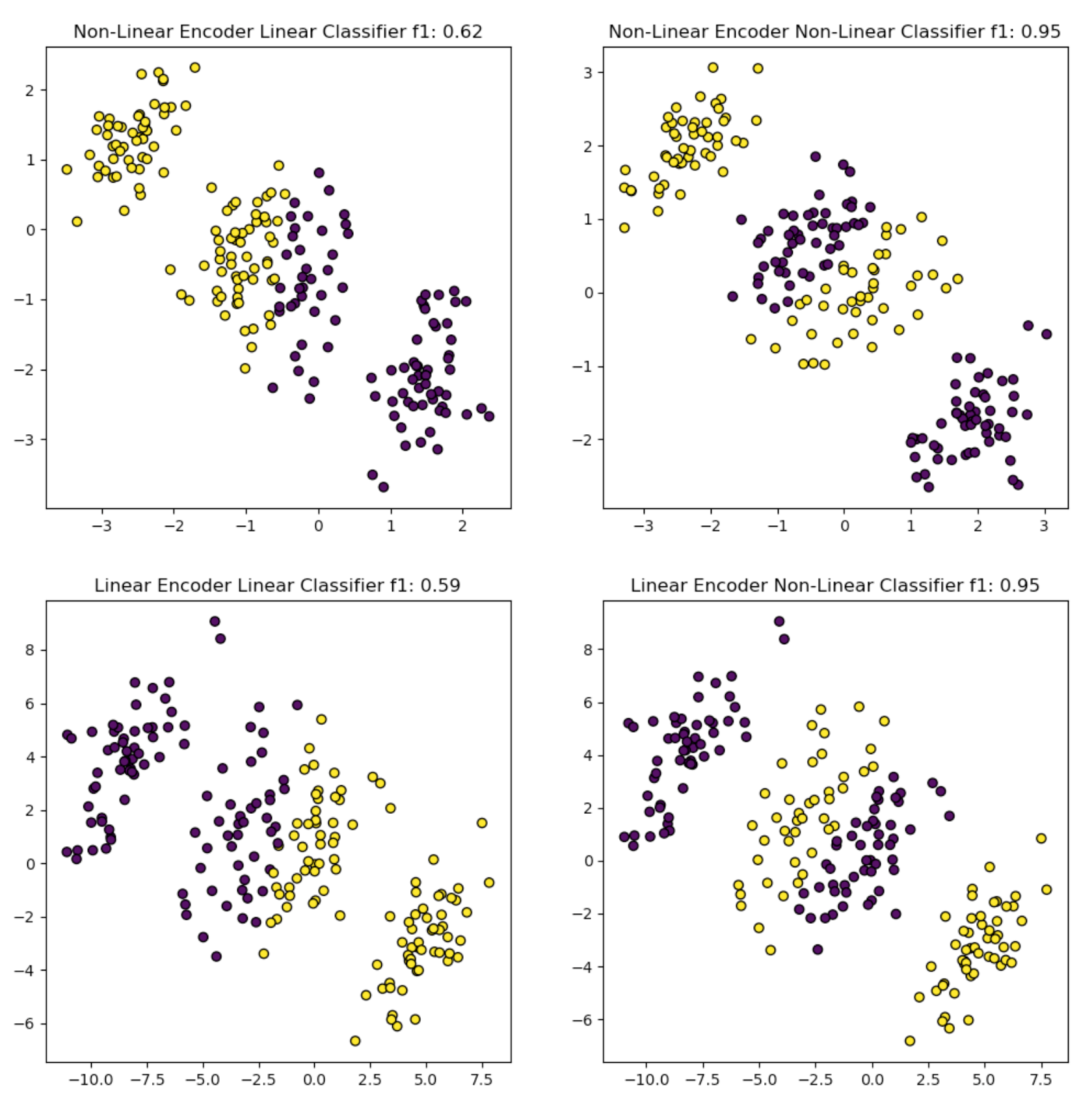


Figure 4. Learned latent spaces for each of the 4 VAE-C architectures. Dots are colored based on the models class prediction. Note latent space rotations do not affect model performance

- Consistent with theoretical predictions all models were able to learn mappings from the latent space to the neural observation space all R^2 ~ 1.00
- Only models with Non-Linear classifiers learned the many to one mapping from the latent space to behavioral class labels.

Varying Classification Weight (λ)

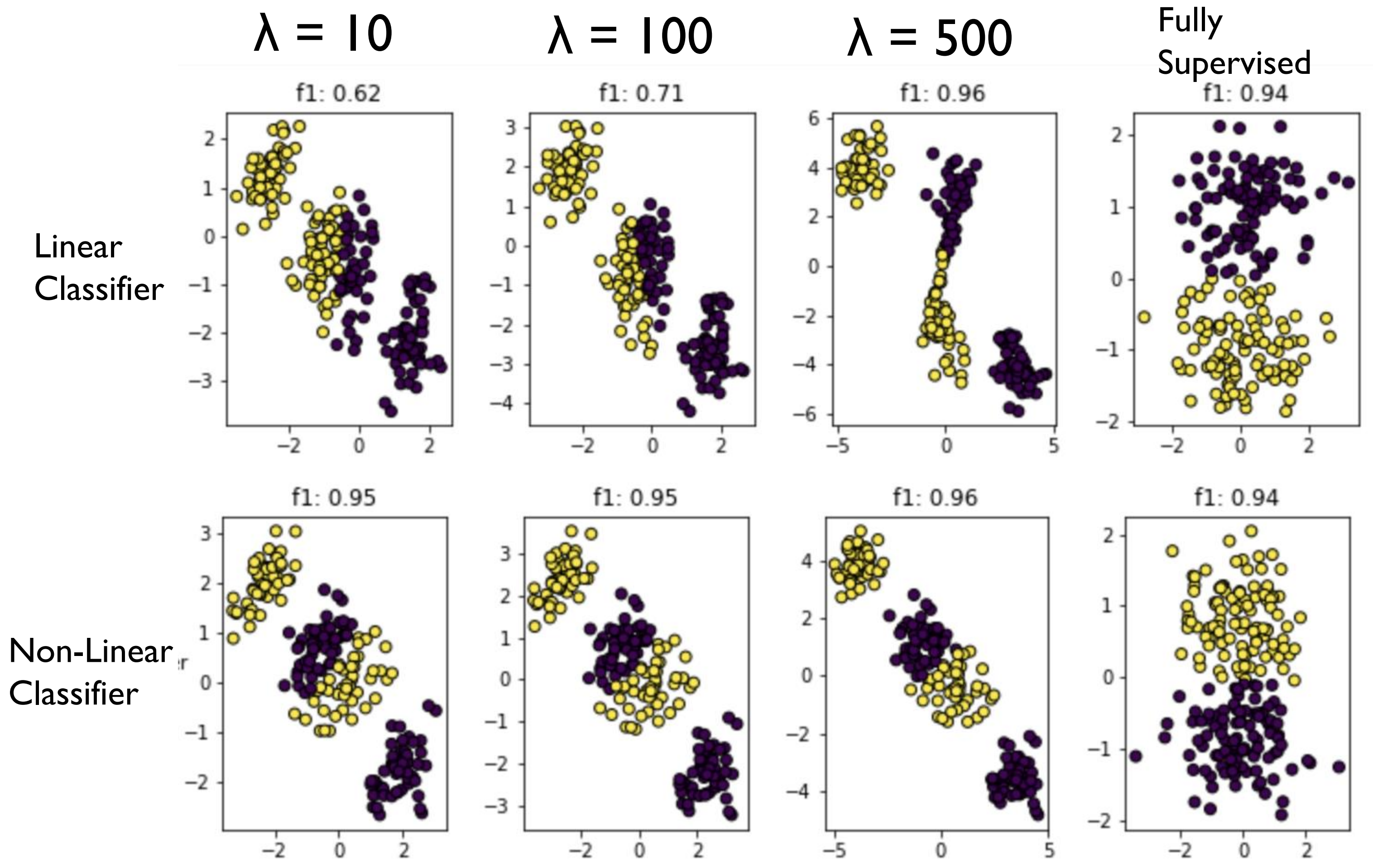


Figure 5. Learned latent spaces for Non-Linear Encoder architecture with Linear and Non-Linear classifiers, for different values of Weight on the classification object (λ) effectively varying the degree of supervision in the VAE-C network.

- The latent space was increasingly influenced by class labels as the weight on the classification portion of the loss function increased.
- In fully supervised analyses degenerate class representations disappear indicating the utility of our VAE-C approach over a non-linear classifier
- R^2 remained ~ 1.00 for all analyses when reconstruction objective was included in the loss function

Conclusions

- Our simulations show the efficacy of our Variational Auto Encoder Classifier method to learn one to many mappings between latent between latent brain states and behavioral class labels, while maintaining reconstruction quality of neural data
- These results suggest the utility of this method for investigating degenerate brain-behavior mappings
- We are currently applying this method to a data set with fMRI and affective ratings.

References

1. Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763-13768.
2. Westlin, C., Theriault, J. E., Katsumi, Y., Nieto-Castanon, A., Kucyi, A., Ruf, S. F., ... & Barrett, L. F. (2023). Improving the study of brain-behavior relationships by revisiting basic assumptions. *Trends in cognitive sciences*, *27*(3), 246-257.
3. Barrett, L. F. (2022). Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, *77*(8), 894.
4. Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in cognitive sciences*, *6*(10), 416-421.
5. Sajid, N., Parr, T., Hope, T. M., Price, C. J., & Friston, K. J. (2020). Degeneracy and redundancy in active inference. *Cerebral Cortex*, *30*(11), 5750-5766.
6. Noppeney, U., Friston, K. J., & Price, C. J. (2004). Degenerate neuronal systems sustaining cognitive functions. *Journal of Anatomy*, *205*(6), 433-442.
7. Khan, Zulqarnain, Yiyu Wang, Eli Sennesh, Jennifer Dy, Sarah Ostadabbas, Jan-Willem van de Meent, J. Benjamin Hutchinson, and Ajay B. Satpute. "A computational neural model for mapping degenerate neural architectures." *Neuroinformatics* 20, no. 4 (2022): 965-979.
8. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, *29*.
9. Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer google schola*, *2*, 5-43.