Dimensionality Reduction for Predicting COVID-19 Dynamics

Nyambura Njenga [1] Asim Suhail [2] Emily Zhao [1]
Georgia Institute of Technology [1] University of Texas at Austin [2]

Georgia Institute of Technology

Georgia Tech College of Sciences
School of Mathematics

Introduction

Dimensionality reduction is the idea of **compressing high-dimensional data into low-dimensional features** that still retain meaningful properties. Then, rather than working in the high-dimensional space, computations are performed on the low-dimensional representation to save resources.

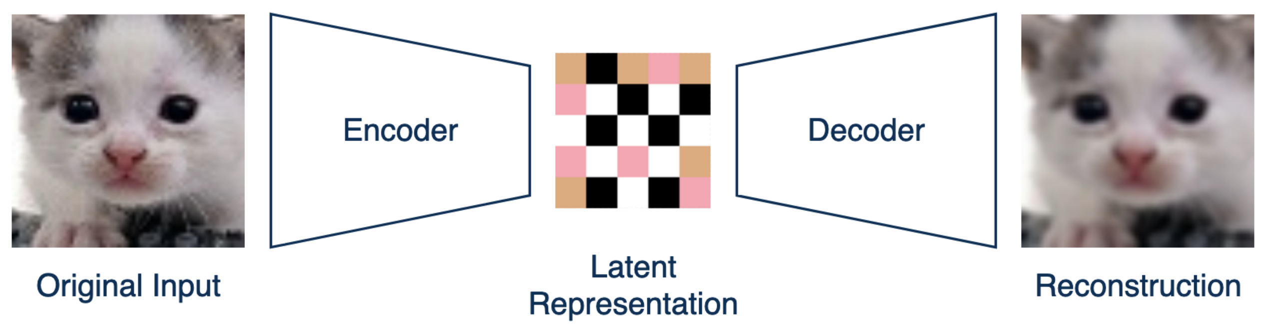

Figure 1. A visualization of data compression and reconstruction by an autoencoder.

Our work uses a neural-network dimensionality reduction technique known as **autoencoding** on a COVID-19 dataset. We then examine three frameworks for predictive modeling in the reduced latent space:

1. Long Short-Term Memory Neural Network (LSTM)
2. Transformer Neural Network
3. Parametric Latent Space Dynamics Identification (LaSDI)

Encoding Process

The **Intrinsic Dimension (ID)** of a dataset is the minimal number of variables necessary to express the characteristics of the dataset. This provides a **lower bound on the number of latent variables** we we may attempt to compress into.

- To approximate this value, we use a k-nearest neighbors algorithm with maximum likelihood estimation (MLE).
- Our results from testing up to 100 nearest neighbors show that **the ID is less than 3**.

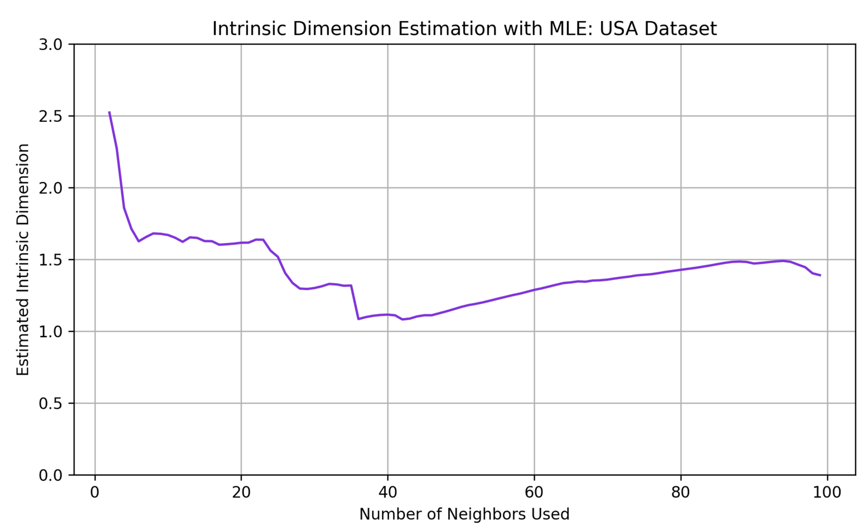

Figure 2. Number of nearest neighbors vs MLE intrinsic dimension

We record the relative mean squared error between the original data and reconstructed outputs from:

1. An autoencoder
2. Principal Component Analysis (PCA), a linear dimensionality reduction method.

For this dataset, autoencoders incur less error than PCA for encoding into computationally tractable dimensions.

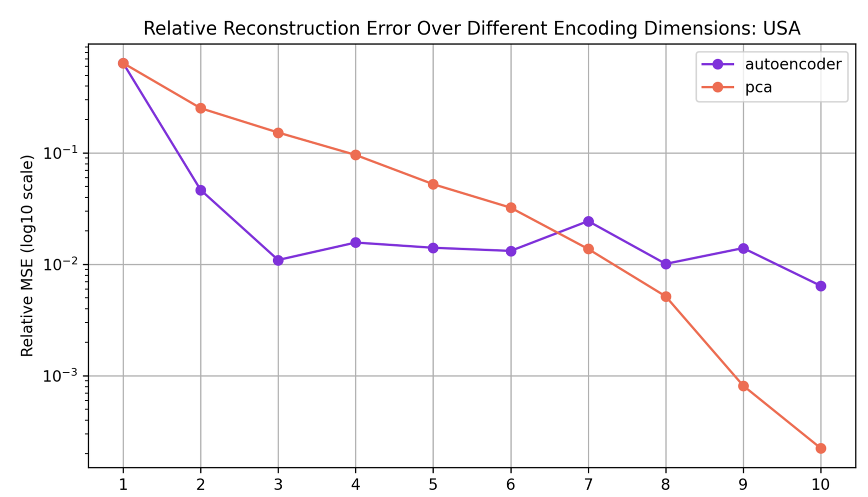

Figure 3. Encoding Dimension vs Relative Reconstruction Error for Autoencoder and PCA

Dynamics Prediction Methods

Long Short-Term Memory Network (LSTM)

LSTMs are a type of recurrent neural network (RNN) designed to better **capture long-term dependencies** in sequence prediction tasks.

- Traditional RNNs can only remember and use the most recently predicted value of the sequence, so they struggle to predict extended sequences.
- In contrast, LSTMs implement memory cell structures that store a range of previously predicted values. This increases interaction between past and current predictions.

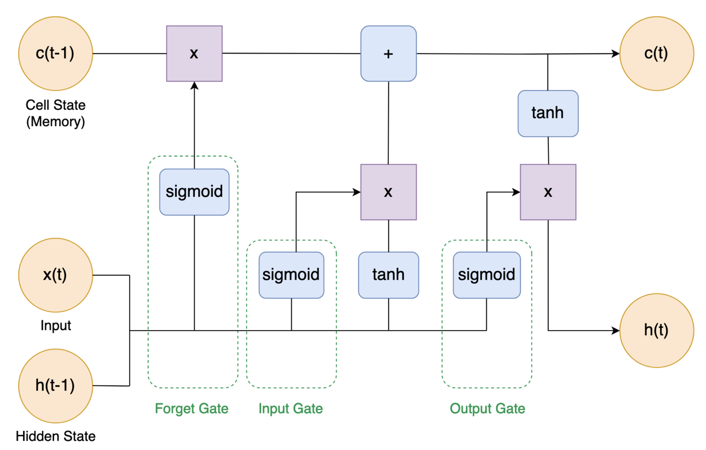

Figure 4. Typical LSTM memory cell

Transformer Neural Network

The transformer architecture allows **full interaction** between all parts of the sequence.

This is achieved through **positional encoding**, which assigns a unique value to each position in the sequence. We used the following formulas for positional encoding:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

where pos is the position and i is the dimension.

Then, an **attention layer** identifies previous positions that are most pertinent to the current prediction.

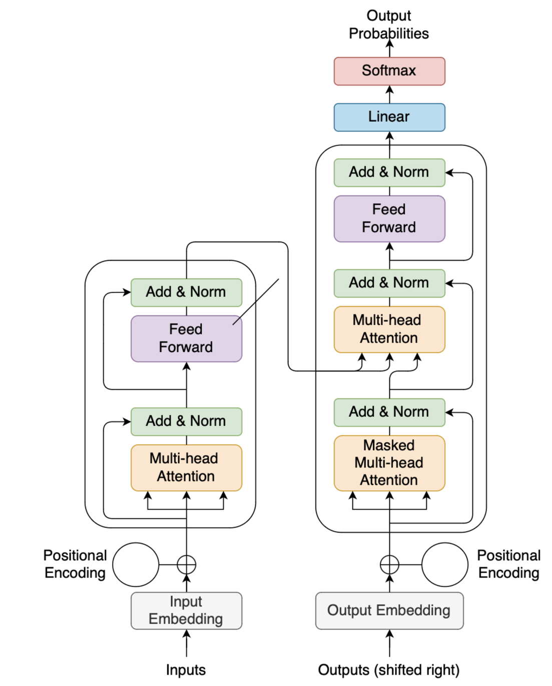

Figure 5. Transformer architecture

Parametric LaSDI

LaSDI finds a dynamical system $\dot{\hat{\boldsymbol{u}}}(t) = f(\hat{\boldsymbol{u}}(t))$ whose solution best matches the latent space trajectory data.

First, compressed data is arranged in a matrix $\hat{\boldsymbol{U}}$ where each row represents a time step and each column an encoded feature:

$$\hat{\boldsymbol{U}} = \begin{bmatrix} \hat{u}_0(t_0) & \hat{u}_1(t_0) & \hat{u}_2(t_0) \\ \vdots & \vdots & \vdots \\ \hat{u}_0(t_N) & \hat{u}_1(t_N) & \hat{u}_2(t_N) \end{bmatrix}$$

Then, the time derivative, $\dot{\hat{\boldsymbol{U}}}$ is approximated using a finite time difference method.

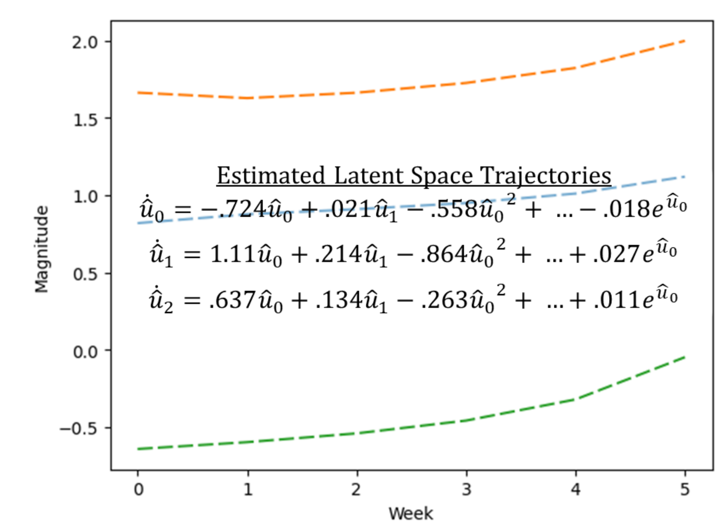

Figure 6. LaSDI trajectory estimations

Next, to estimate $f(\hat{\boldsymbol{u}}(t))$, we define a library of functions $\Theta(\hat{\boldsymbol{U}}(t))$. We include sine, cosine, exp, and polynomial terms up to the second order.

Finally, we find the coefficient matrix, Ξ, which allows us to approximate $f(\hat{\boldsymbol{u}}(t)) = \Theta(\hat{\boldsymbol{U}}(t))\Xi$ by solving the following optimization:

$$\underset{\Xi \in \mathbb{R}^{n_l \times 3}}{\text{argmin}} \left\| \dot{\hat{\boldsymbol{U}}}(t) - \Theta(\hat{\boldsymbol{U}}(t))\Xi \right\|^2$$

Results

For each method, we predicted within the latent space, then decoded the results back into the original variables. We then calculate the relative error for each predicted week.

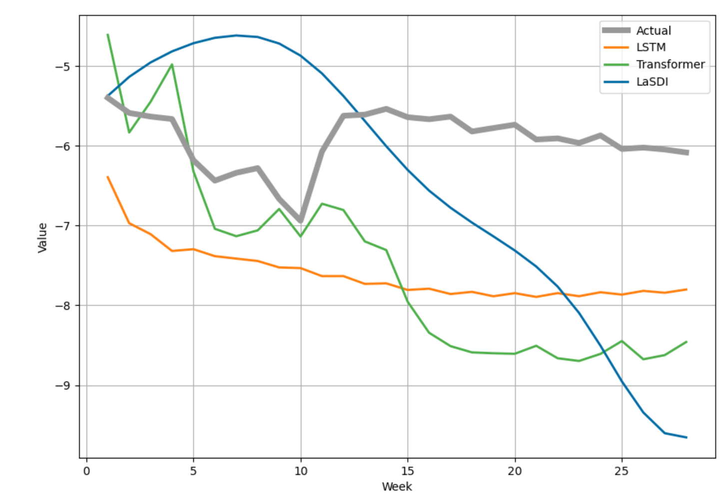

Figure 7. Predictions for the first latent variable

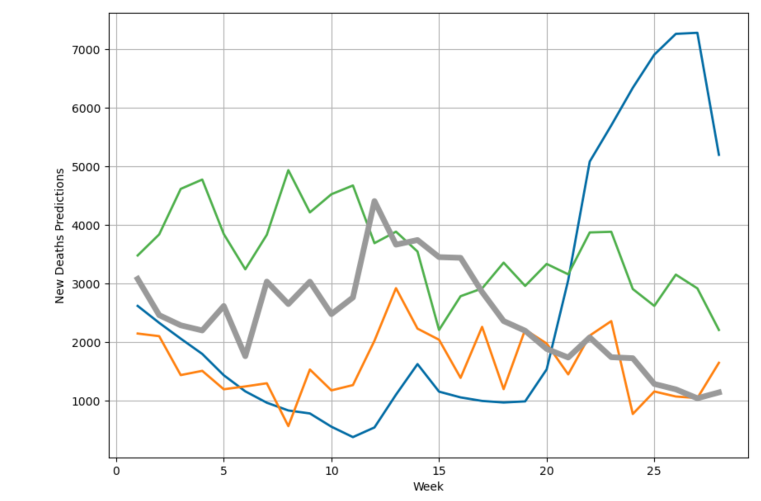

Figure 8. Decoded predictions, new deaths

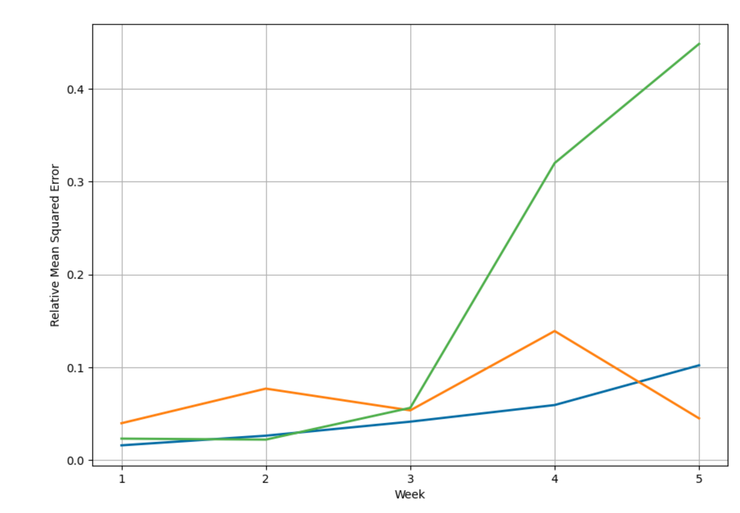

Figure 9. Prediction error for first 5 weeks

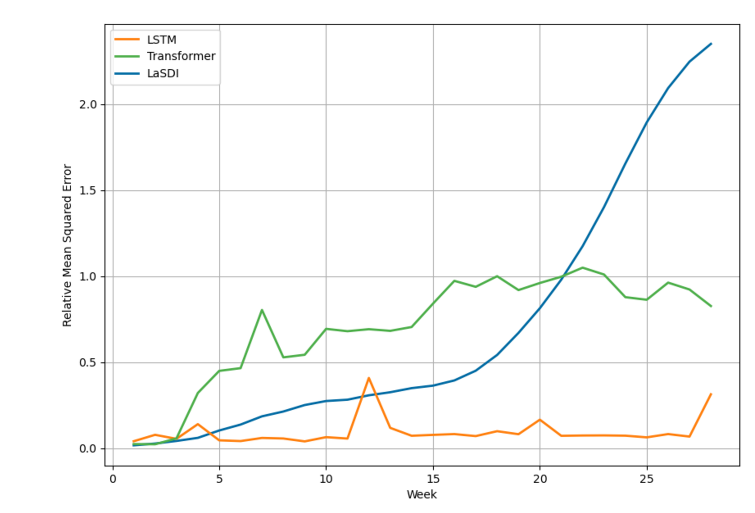

Figure 10. Prediction error for first 28 weeks

Conclusions + Further Research

Our analysis of latent space dynamics prediction for COVID-19 data demonstrates that **LaSDI is a viable method for short-term prediction in "real-world" dynamical systems** while **LSTMs are most suited for predicting long-term trends**.

- LaSDI can provide accurate short-term predictions of disease spread, hospitalization rates, and other critical factors. This can **inform life-saving policies** during a future health crisis.
- Our results highlight **the importance of selecting appropriate models for specific types of systems**. Transformers, while effective in a wide range of applications, may not be suitable for predicting the short-term evolution of continuous, low-dimensional dynamical systems.

Further research in this area includes studying different autoencoder architectures, improving models for long-term predictions, and applying our methods to other real-world dynamical systems.

References

Fries, W. D., He, X., and Choi, Y. (2022). Lasdi: Parametric latent space dynamics identification. *Computer Methods in Applied Mechanics and Engineering*, 399:115436.

Hasell, J., Mathieu, E., Beltekian, D., et al. (2020). A cross-country database of covid-19 testing. *Sci Data 7*, 345.

Levina, E. and Bickel, P. J. (2004). Maximum likelihood estimation of intrinsic dimension. *Department of Statistics, University of California, Berkeley*.

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., et al. (2021). A global database of covid-19 vaccinations. *Nat Hum Behav*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Mentors: Dr. Wenjing Liao and Zhaiming Shen

Funded by NSF grant 2244427 and Georgia Tech College of Sciences

Georgia Tech Math REU 2024