Reconstructing Hidden Neural Dynamics: Autoencoder Applications Under Partial Observability

Rong Ding
s1067006

1 Introduction

System identification presents a key challenge in understanding complex dynamical systems like brain signals. While neuroscience benefits from high-resolution data, efforts to uncover neural dynamics are often constrained by partial observability—the inability to fully access all system components. Non-invasive techniques like EEG and fMRI offer valuable insights but suffer from noise, distortion, and limited access to latent states, highlighting the need for methods to reconstruct hidden dynamics from partial and noisy observations.

Machine learning, particularly encoder-decoder architectures, has shown promise in addressing partial observability. Autoencoders, for instance, have revealed latent dynamics in physics and engineering domains (1) and have even been used to identify governing equations of physical systems (3). These advances suggest that machine learning models could help uncover neural dynamics in the absence of full observability.

Traditional neuroscience methods like multi-variate pattern analysis (MVPA) often rely on external stimulus-driven designs (4), despite evidence that intrinsic dynamics significantly shape brain activity (2). This highlights the need for tools that reveal latent processes beyond stimulus-driven paradigms.

In this study, we use the Jansen-Rit Neural Mass Model (5) as a test case to evaluate an autoencoder's ability to reconstruct latent states under partial observability. By systematically reducing observable dimensions, we assess how well the model can infer hidden states and uncover system dynamics, addressing the question: How effectively can an autoencoder reconstruct latent states and reveal system dynamics under varying degrees of partial observability?

2 Simulation of Jansen-Rit Neural Mass Model (NMM)

We simulated the Jansen-Rit neural mass model to generate biologically plausible cortical oscillations for autoencoder-based system identification. The model captures the activity of three neural populations: pyramidal neurons, excitatory interneurons, and inhibitory interneurons. The following six state variables describe the system dynamics:

$$\dot{y}_0(t) = y_3(t), \tag{1}$$
$$\dot{y}_3(t) = AaS(y_1(t) - y_2(t)) - 2ay_3(t) - a^2 y_0(t), \tag{2}$$
$$\dot{y}_1(t) = y_4(t), \tag{3}$$
$$\dot{y}_4(t) = Aa\left[p(t) + J_2 S(J_2 y_0(t))\right] - 2ay_4(t) - a^2 y_1(t), \tag{4}$$
$$\dot{y}_2(t) = y_5(t), \tag{5}$$
$$\dot{y}_5(t) = BbS(J_4 y_0(t)) - 2by_5(t) - b^2 y_2(t). \tag{6}$$

Here:

Preprint. Under review.

- $y_0(t)$: Pyramidal neuron membrane potential,
- $y_1(t)$: Excitatory interneuron membrane potential,
- $y_2(t)$: Inhibitory interneuron membrane potential,
- $S(V) = \frac{2e_0}{1+\exp[r(V_0-V)]}$: Sigmoid activation function.

For the model simulation, all the state variables were initialized to 0, and the external input $p(t)$ was modelled as uniformly distributed noise in the range $[120, 320]$. Model parameters (A, B, a, b, etc.) were taken from previous literature, where these values were shown to be biologically plausible (REF). The system of equations was numerically integrated using JAX for acceleration. A time step of $\Delta t = 0.001\,\mathrm{s}$ was used, simulating $10.0\,\mathrm{s}$ of activity (i.e., 10000 timesteps). Activity of all the state variables was recorded over time to train the autoencoder.

3 Autoencoder-Based System Identification

In this project, we implemented an autoencoder architecture to examine the system identification of the NMM under differing degrees of observability. The autoencoder consists of an encoder, which compresses the observable input states into a lower-dimensional latent representation, and a decoder, which reconstructs all system states from the latent representation.

3.1 Autoencoder Architecture

1. Input Layer: The input layer received the observable states. To investigate to what extent the autoencoder can handle partial observability, the observable varied from all six system states (full observability) to a single observable state (extreme partial observability). Observed states included neuronal potentials and their changing rates of pyramidal, excitatory, and inhibitory neurons.

2. Hidden Layers: The encoder and decoder respectively included 2 fully connected hidden layers: one hidden layer contained 128 neurons, and the other 64 neurons. Each hidden layer used the Rectified Linear Unit (ReLU) as the activation function. The hidden layer structure of the decoder mirrored that of the encoder.

3. Latent Space: The encoder compressed the input into a low-dimensional latent space composed of three neurons, which was then used by the decoder to reconstruct all six system states. The rationale behind the choice of three neurons was the intuitive presumption that a column of the NMM consists of three types of neurons.

4. Output Layer: The decoder reconstructed all six system states, including both observed and unobserved states.

3.2 Loss Function

To handle the variable input-output dimensionality due to partial observability, a customised Mean Squared Error (MSE)-based loss function was used. This function calculates the reconstruction error for the observed states and ensures the loss only accounts for the dimensions available at each time step. Specifically, we applied the following loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\|\hat{y}_i - y_i\|^2 \right)$$

where $\hat{y}_i$ and y_i are the predicted and true values of the system states, respectively. For cases with partial observability, the loss was computed only over the observable dimensions at each time step. This approach thus allowed the model to handle varying and mismatching numbers of input-output dimensions without errors.

3.3 Training

The training of the autoencoder was conducted in PyTorch using the Adam optimizer, with a learning rate of 0.001 to ensure stable convergence. Training was conducted over 50 epochs with a batch size of 32.

3.4 Partial Observability Testing (Experiments)

To evaluate the model's ability to infer unobserved states under partial observability, we systematically reduced the input dimensions. Specifically, we used the full observability model as the baseline, in which all six system states (membrane potentials of pyramidal, excitatory, and inhibitory neurons and their change rates) were available for reconstruction.

Then, we progressively reduced the number of observable states to test how well the autoencoder could reconstruct the unobserved dynamics. In one scenario, we restricted the input to only three membrane potential states—pyramidal, excitatory, and inhibitory neurons. In the other scenario, we further limited the input to only the membrane potential of pyramidal neurons, representing a case of very limited observability. This setup allowed us to assess the model's performance under increasingly constrained conditions, examining its ability to recover the hidden states and reconstruct the full dynamics of the system.

To quantify the accuracy of the reconstructed states, we used the Pearson correlation coefficient between the true and reconstructed states over the entire simulated/reconstructed time series. We computed this correlation for each state per model. If the correlation coefficient was below zero (negative), the accuracy value was set to zero, as we think that a negative correlation indicates differences from the actual data and thus poorer reconstruction.

We also visually inspected the latent space of each trained autoencoder model ([100,600] timestep), so that we could analyze how the reduced input dimensions affected the structure of the encoded representations, hence providing insight into the model's ability to capture the underlying system dynamics.

4 Results

As shown in Figure 1, the trained autoencoder successfully reconstructed all observable state variables under full observability. Notably, the autoencoder reconstructed the amplitude and rate of inhibitory neurons more accurately in general than those of excitatory neurons.

Under partial observability, when only the membrane potentials of the three cell populations were provided, the reconstruction accuracies for the rate variables were notably lower (Pearson's $r < 0.35$). Interestingly, the inhibitory interneuron rate achieved higher accuracy compared to the rates of pyramidal neurons and excitatory interneurons, which were poorly reconstructed.

Under extreme partial observability, where only the pyramidal neuron membrane potential was available, distinct reconstruction patterns emerged across neuronal types. The pyramidal neuron rate was reconstructed poorly (Pearson's $r = 0.0442$). For excitatory interneurons, the membrane potential reconstruction (Pearson's $r = 0.4622$) outperformed that of its rate (Pearson's $r = 0.3127$). Conversely, for inhibitory interneurons, the rate reconstruction (Pearson's $r = 0.8512$) was far more accurate than the potential (Pearson's $r = 0.1312$). Note, though, that while Pearson's correlation captures trends in data similarity, the absolute values of the reconstructed signals deviated significantly from the ground truth (see Figure 1A).

The latent space representations of the trained autoencoders are shown in Figure 1C. Under full observability, the latent space exhibited chaotic clustering, reflecting the complex dynamics of the NMM. With only membrane potential inputs, the latent space captured oscillatory relationships between dimensions, corresponding to the rhythmic activity of the system. When limited to the pyramidal membrane potential alone, the latent space displayed more linear relationships.

5 Discussion and Conclusion

In this project, we explored the ability of an autoencoder to reconstruct neural oscillatory dynamics under partial observability using the Jansen-Rit model. We evaluated how well the autoencoder inferred unobserved states and represented latent dynamics with progressively restricted input dimensions. The results showed that reconstruction accuracy declined with reduced observability, with inhibitory interneurons being more robustly reconstructed than excitatory neurons, even when inhibitory activity was unobserved. This robustness may stem from the opposing relationship between inhibitory and pyramidal dynamics, which is then more easily disentangled. Conversely, the

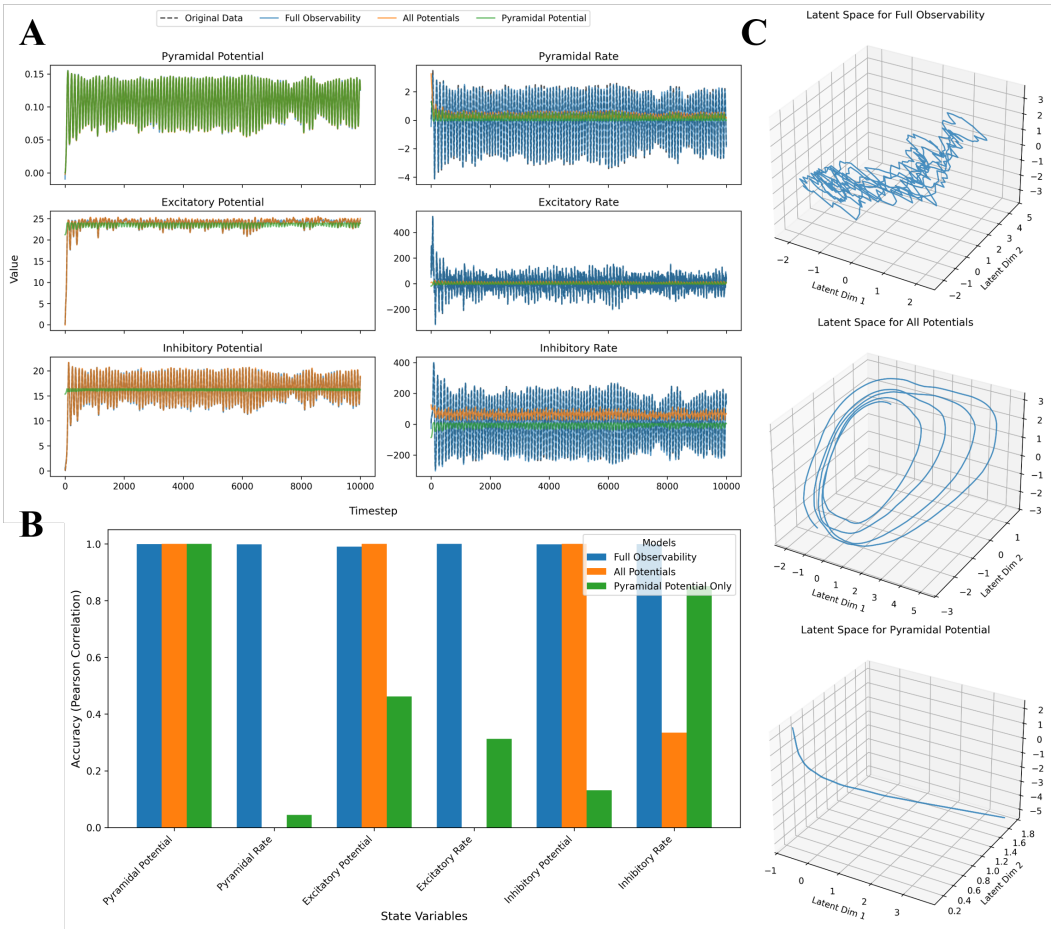

Figure 1: (A) Simulated data of the NMM and reconstructed activity of the autoencoder under three different degrees of observability. (B) Accuracy of reconstruction of each state variable per trained autoencoder. (C) Latent space per autoencoder model.

reconstruction of excitatory interneurons was poorer, possibly due to stronger non-linear interactions; this was evidenced as under very limited observability the linear latent space revealed a stronger slope between dimensions reflecting opposing activity (e.g., inhibitory and pyramidal) compared to those with positively correlated dynamics.

When observability was limited to membrane potentials, the autoencoder struggled to reconstruct rate variables, possibly reflecting lower contributions of rates to observable dynamics or strong correlations with potentials. The latent space reflected this limitation, capturing the cyclic nature of the neural dynamics but losing the chaotic features present under full observability. This aligns with the role of external input, which primarily influences the excitatory rate in the Jansen-Rit equations through which chaotic dynamics are introduced into the system. Under partial observability, the autoencoder also exhibited a clear transition from chaotic to linear latent space representations, reflecting significant information loss. This finding thus suggests the importance of capturing sufficient dimensions of the observable state space to preserve the complexity of latent dynamics.

Pearson's correlation, although useful for assessing trend similarity, may not have been the optimal measure of reconstruction accuracy, as it failed to account for discrepancies in absolute signal values between reconstructed and actual data despite high correlation values.

This study employed a single, simple autoencoder structure and does not fully represent the potential of autoencoders in identification of hidden dynamics. Nonetheless, our findings highlight the promise of these models for disentangling complex neural dynamics under partial observability. Future work should explore more advanced architectures and metrics to better address these challenges.

References

[1] Brunton, S.L., Proctor, J.L., & Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), 3932–3937.

[2] Buzsáki, G. (2019). *The Brain from Inside Out*. Oxford University Press.

[3] Champion, K., Lusch, B., Kutz, J.N., & Brunton, S.L. (2019). Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45), 22445–22451.

[4] Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, 87(2), 257–270.

[5] Jansen, B.H., & Rit, V.G. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics*, 73(4), 357–366. https://doi.org/10.1007/BF00199471