Decoding Behavior from Neural Data

Berkan Ottlik and Quentin Chu

1 Abstract

Decoding neural data involves extracting information from neural signals, offering insights into information processing, memory encoding, and the generation of thoughts and behaviors. This study focuses on decoding neural activity using the International Brain Lab (IBL) dataset (Laboratory, 2023). The IBL's "Brain-wide Map" project, utilizing Neuropixels probes, captures the neural activity across diverse brain regions of dozens of mice during decision-making tasks. However, the sparsity of neural spikes and inconsistency of mouse physiology and probe placements make a meaningful analysis challenging.

We concentrate on a subset of the IBL dataset, specifically decoding wheel speed from neuron spikes in the posterior complex of the thalamus—a region associated with relaying sensory and motor signals. To address data challenges related to sparsity and inconsistency, we propose four distinct models: Reduced Rank Model, Neural Network Model, Continuous Decoder Model and ARD Regression Model.

2 Introduction

Decoding neural data refers to the process of extracting and interpreting meaningful information or patterns from neural signals, which consist of measurements of electrical or biochemical signals generated by neurons in the brain. Decoding neuron activity aims to understand these signals and gain insights into how the brain processes information, encodes memories, and generates thoughts and behaviors. This understanding is crucial for answering key scientific questions such as "what information is stored in different brain regions?", "how much of neural variability is stochastic?", and "how much can single neurons tell us?"

In this project, we perform a decoding task using data from the International Brain Lab (IBL), a collaborative research initiative involving 22 neuroscience laboratories from around the world, with the shared mission of advancing our understanding of the neural systems and circuits that underlie behavior. IBL's "Brain-wide Map" project provides neural activity data during a decision-making task from nearly all major brain areas with Neuropixels probes. Despite the extensive available data, there are significant challenges to utilizing the data. Specifically, since each lab uses a different mouse and probes are inserted in slightly different locations in each mouse, the recordings generally pick up different numbers of neurons. As a result, there is little correspondence and consistency across recording sessions. Moreover, the neuron spikes across different trials and time bins tend to be sparse, which could lead to overfitting and increased computational complexity due to high

dimensionality.

Using a subset of the IBL dataset, we aim to decode the wheel speed of a trial from neuron spikes in the posterior complex of the thalamus, the region of the brain that is associated with relaying sensory and motor signals. To address the aforementioned challenges related to the data, we propose four different models, two of which (Reduced Rank Model and Neural Network Model) first use separate linear transformations to project the each recording into a shared latent space and then learn a single model, while the other two (Continuous Decoder Model and ARD Regression Model) perform regularized regression on each recording individually with an inductive bias towards sparsity. We had expected that the Reduced Rank and Neural Network models would effectively capture information across recording sessions and different mice, and would demonstrate superior performance to the Continuous Decoder and ARD Regression Models. However, our results were contrary to our expectations.

3 Data

In the IBL each of the labs gathered neural recordings from over 100 head-mounted mice using Neuropixel arrays. The mice face a screen that provides a visual stimulus in each trial, and are fed sweet juice through a faucet as reward when they spin the wheel in front of them in the correct direction (Figure 1). To further explore the IBL data, visit this interactive IBL website.

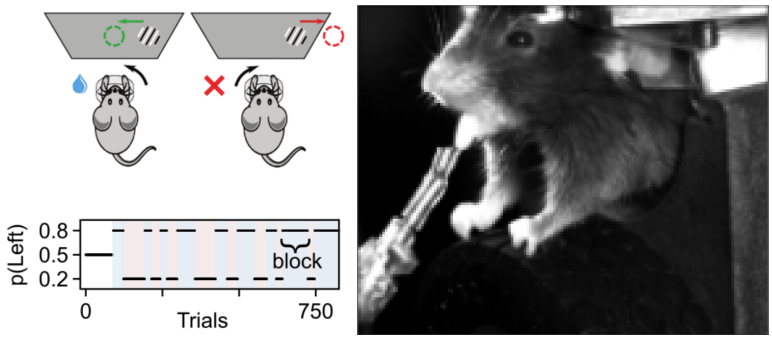

Figure 1: Left: Illustration of mouse performing task. Right: Head-mounted mouse with faucet (Laboratory, 2023).

Our data consists of $I = 112$ recording sessions, with each recording performed at a specific lab with a specific mouse. Within the i-th recording, there are $K^{(i)}$ trials, each of which contains data from $C^{(i)}$ neurons across $T = 40$ time bins. Each trial spans 2 seconds, implying that each time bin represents 50 milliseconds. The number of trials across recordings ranges from 340 to 1043, with mean and median of 538.1 and 516.5 respectively. The number of neurons across recordings ranges from 2 to 371, with mean and median of 89.9 and 59.5 respectively. The input tensor $X \in \mathbb{N}^{K^{(i)} \times C^{(i)} \times T}$ represents neuron spikes in the posterior complex of the thalamus, while the output matrix $Y \in \mathbb{R}_+^{K^{(i)} \times T}$ represents absolute wheel speed. Note that both $\mathbb{N}$ and $\mathbb{R}_+$ include 0.

We split our data into training and test sets for each recording session by first shuffling the trials, and then randomly assigning 80% of the trials to the training set and 20% to the test set.

4 Models

As mentioned above, the IBL data present the following challenges: (1) lack of consistency across recording sessions, since each recording involves a different mouse and different probes, and thus detecting a different number of neurons; (2) sparsity of neuron spikes across different trials and time bins, potentially resulting in overfitting and increased computational complexity due to high dimensionality.

To address these challenges, we propose the following models: Reduced Rank Model, Neural Network Model, Continuous Decoder Model, and ARD Regression Model.

Due to sparsity in neuron spikes and large variance in wheel speeds, we applied a Gaussian filter to the data and labels prior to implementing our models.

4.1 Reduced Rank Model

The first model we implemented is the Reduced Rank Model from Zhang et al. (2023). Traditional neural decoding relies on the following *full-rank* model:

$$Y_k^{(i)} = f(B^{(i)\top} X_k^{(i)} + b_k^{(i)})$$

where f represents any arbitrary decoder, and $B^{(i)} \in \mathbb{R}^{C(i) \times T}$ is the coefficient matrix with k and i indexing the trial and recording session respectively. $C^{(i)}$ represents the number of neurons and varies for each session, while T represents the number of time bins and remains constant for each session (Figure 2).

In practice, estimating $B^{(i)}$ can be challenging due to high dimensionality and sparsity of data. To overcome these challenges, we propose the Reduced Rank Model, which imposes a low-rank structure on the coefficients $B^{(i)}$ as follows:

$$Y_k^{(i)} = f((U^{(i)}V)^\top X_k^{(i)} + b_k^{(i)})$$

where $U^{(i)} \in \mathbb{R}^{C^{(i)} \times R}$ and $V \in \mathbb{R}^{R \times T}$, with R being the reduced rank that helps prevent overfitting. The model uses the $U^{(i)}$ matrix as a linear transformation specific to a recording session, and projects all neural recordings with different numbers of neurons into a shared space of dimension R. It then uses the V matrix to capture the temporal activation patterns across all sessions. Since $(U^{(i)}V)^\top X_k^{(i)}$ is a $T \times T$ matrix, we chose a decoder function f that reduces the matrix to a T vector by taking the mean along the first dimension.

4.2 Neural Network Model

The second model we implemented is a Neural Network Model. Similar to the Reduced Rank Model, it uses the $U^{(i)}$ matrix with rank R to capture the sparse neural activation patterns; but

instead of the V matrix, it passes $U^{(i)\top} X_k^{(i)}$ as an input to a fully-connected neural network with 3 hidden layers and 1000 hidden units per layer. The resulting computation graphs between the Reduced Rank and Neural Network Models are identical, except for the parameters that are shared across recording sessions (Figure 2).

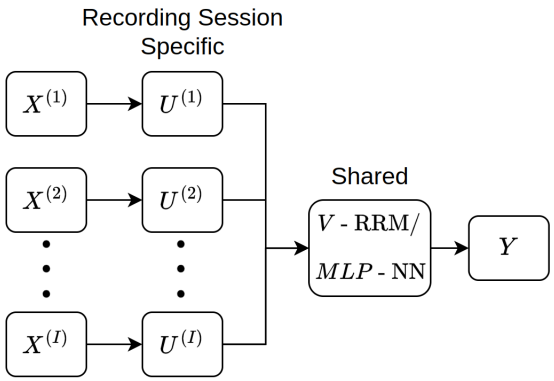

Figure 2: RRM and NN models.

4.3 Continuous Decoder Model

We implemented Ridge regression on the continuous variable of wheel speed for each recording session, with the following objective function:

$$\min_{W^{(i)}} \|X^{(i)} W^{(i)} - Y^{(i)}\|^2 - \alpha \|W^{(i)}\|^2$$

where $X^{(i)} \in \mathbb{N}^{K^{(i)} \times (C^{(i)} \times T)}$ represents neuron spike counts, $W^{(i)} \in \mathbb{R}^{T \times (C^{(i)} \times T)}$ is the coefficient matrix, and $Y^{(i)} \in \mathbb{R}_+^{K^{(i)} \times T}$ represents wheel speeds. The regularization parameter α is selected for each recording sessions by a grid search over $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$.

4.4 ARD Regression Model

Given the sparsity of neuron spikes across different recordings, trials and time bins, we also considered a Bayesian regression model with an Automatic Relevance Determination (ARD) prior (MacKay et al., 1994). From the data set of input-target pairs $\{X^{(i)}, Y^{(i)}\}_{i=1}^{I}$ where I represents the number of recording sessions, we have $X^{(i)} \in \mathbb{N}^{K^{(i)} \times C^{(i)} \times T}$ containing neuron spike counts, and $Y^{(i)} \in \mathbb{R}_+^{K^{(i)} \times T}$ containing wheel speeds. We assume that the targets are sampled from the model with additive noise:

$$Y^{(i)} = f(X^{(i)}; B^{(i)})) + \epsilon^{(i)}$$

where B is a $C^{(i)} \times T$ coefficient matrix and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian noise. Thus we have:

$$p(Y^{(i)} \mid X^{(i)}) = \mathcal{N}(Y^{(i)} \mid f(X^{(i)}; B^{(i)}), \sigma^2)$$

To avoid overfitting, Bayesian regression regularizes the parameters by defining an explicit prior distribution over them, often in the form of a zero-mean Gaussian. Given the broad differences in responsiveness across neurons and time bins, we chose to implement the Automatic Relevance Determination (ARD) prior, which furnishes each parameter with its own hyperparameter λ_j:

$$p(B \mid \lambda) = \mathcal{N}(B \mid 0, \Lambda^{-1})$$

where Λ is a positive definite diagonal matrix and $\text{diag}(\Lambda) = \lambda = \{\lambda_1, \ldots, \lambda_M\}$. ARD regression has been shown to encourage even greater sparsity than Ridge regression by assigning a hyperparameter for each parameter, and concentrating the probability mass at zero for those parameters that are "irrelevant" per the evidence (Tipping, 2001). The hyperparameters λ and $\beta = \sigma^{-2}$, which represent the hierarchical prior and noise variance respectively, are both drawn from Gamma distributions, such that:

$$p(\lambda) = \prod_{i=1}^{M} \text{Gamma}(\lambda \mid a, b),$$
$$p(\beta) = \text{Gamma}(\beta \mid c, d)$$

The graphical model of ARD regression is set forth in Figure 3:

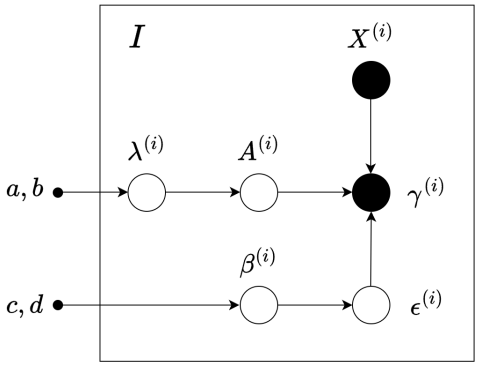

Figure 3: Graphical Model of ARD Regression.

5 Results

In comparing our models, we use test mean squared error as our performance metric, and set as our baseline a mean predictor, which outputs the average wheel speed per time-step for a given recording session. As shown in Figure 4, we see that the Continuous Decoder Model performed the best in all sessions, while the Reduced Rank and ARD Regression Models struggle to beat the baseline mean predictor.

5.1 RRM underperformance

We analyzed the role of the rank parameter for the reduced rank model (Figure 5), and saw that the optimal rank varied across recording sessions, and the differences in performance over different

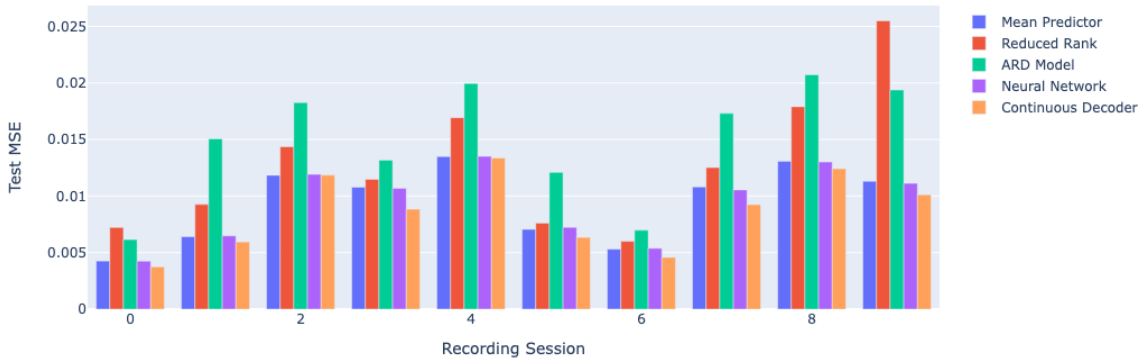

Figure 4: Comparison of model performance in 10 recording sessions.

ranks are insignificant. This suggests that despite the sparsity of the neuron spikes in the data, the rank parameter does not have a meaningful impact on the model's predictive power, and the bias of the model carries more weight than the coefficients.

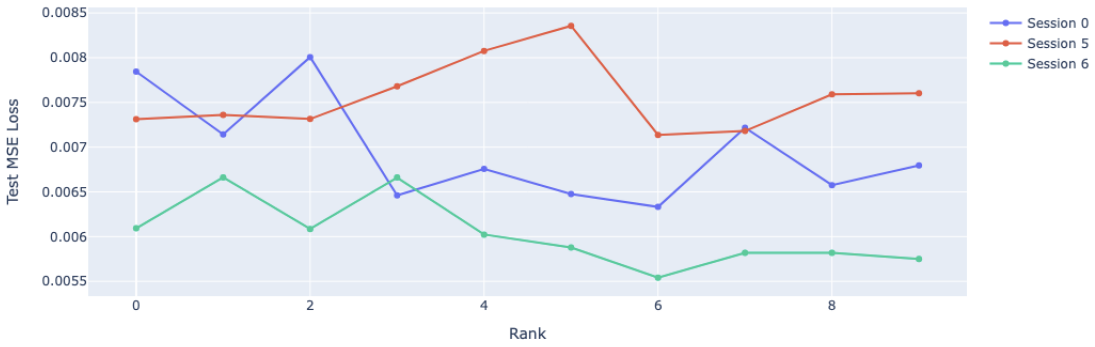

Figure 5: MSE by rank in Reduced Rank Model.

5.2 Neural Network Model analysis

The Neural Network Model failed to outperform the continuous decoder, even when it overfitted the data and when it was well-regularized. This suggests that a more complex fully-connected neural network does not provide any more information than a simple linear regression with respect to predicting wheel speed from neuron spikes. This illustrates just how difficult it is to decode behavior from neural data, and that simply increasing model complexity and number of parameters will not necessarily improve predictive power.

5.3 ARD Regression underperformance

Similar to the Reduced Rank Model, ARD regression did not perform well in predicting wheel speed, despite ARD's tendency to promote sparsity with its parameter-specific regularizers. An

explanation could be that even though the neuron spikes are sparse, there may be latent variables not captured in the data that account for the mouse behavior and observed wheel speeds. By not imposing sparsity through reduced rank or ARD prior, the Ridge regression used by the Continuous Decoder is able to capture the overall data and produce better predictive results.

6 Discussion

Using a subset of the IBL dataset, we attempt to decode wheel speed from neuron spikes in the posterior complex of the thalamus. To address data challenges related to inconsistency and sparsity, we proposed four models – Reduced Rank Model, Neural Network Model, Continuous Decoder Model and ARD Regression Model – and found that the Continuous Decoder Model, which is based the simple Ridge regression, performed the best.

As noted earlier, the data on neuron spikes is very sparse, with over half of the neurons spike fewer than once every ten time bins. In addition, we chose to use only neural data from the posterior complex of the thalamus, thereby omitting a significant portion of available neural data in our models. Using a broader range of neural data, Azabou et al. (2023) have successfully used Transformer models on decoding in primates. We suspect that our data might have been too limited, and our models could potentially achieve better results with a broader set of neural data.

Our results indicate that the predictive performance of our models vary across recording sessions and it is difficult to share information across recording sessions. A future direction could be to explore the reasons for better performance, including dataset size in terms of number of trials or neurons, and variance of neurons spikes and behavior.

References

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding, 2023.

International Brain Laboratory. Data release - Brainwide map - Q4 2022. 1 2023. doi:10.6084/m9.figshare.21400815.v6.

David JC MacKay et al. Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, 100(2):1053–1062, 1994.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Yizi Zhang et al. Neural decoding with side information (draft), 2023.