Neural Spike Train Analysis Using Expectation-Maximization (EM) and Generalized Factorial Coupling Analysis (GFCA)

Beomsuk Seo
beseo@ucsd.edu

Vaibhav Bommisetty
vbommisetty@ucsd.edu

Mikio Aoi
maoi@ucsd.ed

Abstract

Decision-making is at the forefront of human minds, both in the concept of how people operate in their day to day lives, but also for neuroscience, in trying to understand the human brain better. This research investigates the neural mechanisms underlying decision-making by analyzing brain activity recorded from mice performing a visually-guided choice task.

Code: https://github.com/vbommisetty/Quarter-1-DSC180A-Project

1 Introduction

This paper deciphers the neural code underlying decision-making using statistical methods such as Expectation Maximization (EM) and Gaussian-Process Factor Analysis (GPFA). Conclusions of this research can hold profound implications for understanding cognition and developing advanced brain-computer interfaces. This project delves into the intricate dynamics of neural populations during decision-making by analyzing brain activity recorded from mice performing a visually-guided choice task. Motivated by the need for models that accurately capture both temporal correlations and the non-negative nature of spiking activity, we employ a hierarchical approach. This approach utilizes variational EM to learn low-dimensional neural trajectories, providing a concise representation of the high-dimensional neural data. Furthermore, we investigate the interplay between two key brain regions, the Superior Colliculus Deep Gray Layer and the Superior Colliculus Intermediate White Layer. This investigation aims to provide a deeper understanding of how sensory information is encoded and transformed into motor commands during decision-making.

In this paper, we simulate spike train data for left and right trials, then apply EM and GFCA to analyze the data. These processes aim to demonstrate the utility of these methods on identifying neural activity.

1.1 Literature and Prior Work

Research on the neural activity that drives decision-making has rapidly evolved, driven by advancements in both neuroscience and statistical modeling. For example, a foundational study by Shadlen and Newsome done in 2001 showed how neurons in the parietal cortex drove decision-making in primates based on visual stimuli in motion (Shadlen and Newsome (2001)). Another study done by Cunningham and Yu (2014) explores different ways to perform dimensionality reduction including the PCA and GPFA methods, which we will utilize as well (Cunningham and Yu (2014)).

Expectation maximization (EM) for principal component analysis (PCA) is used to find the maximum likelihood estimates of a model's parameters, and can be used for the factor analysis model as well (Bishop (2007)). This is especially useful for when there are latent variables in the data that are unobserved. The expected value of the log likelihood is calculated in the 'E' step of EM, and the 'M' step maximizes θ, where θ is the (unknown) parameter vector that is passed as an argument into the log-likelihood function.

Gaussian Principal Factor Analysis (GPFA) allows us to uncover latent variables and trajectories from brain data with high dimensions. GPFA is advantageous because it ”simultaneously performs the smoothing and dimensionality-reduction operations” (Yu et al. (2009)). In their study, Yu et al. were able to use GPFA to improve the model's fit to the correlations in the neural data used, over more traditional models like PCA.

1.2 Data Description

1.2.1 Real Data

The International Brain Laboratory (IBL) is a group of research laboratories that aim to develop a global model in mouse brain activity. The large scale neuroscience collaboration was launched in 2016 due to a group of researchers believing that multiple labs working together would be more efficient at deepening our understanding of how the brain works (Wikipedia (2024)). The dataset we will be using is based on the IBL's Brain-wide Map project, where they work on mapping neural activity across a mouse brain at the single cell level.

This dataset contains data on around 140 mice, across three different countries There were a total of 459 sessions across all the laboratories, with around 700 probe insertions (Laboratory et al. (2021)). The IBL were able to reproduce the results between laboratories, marking a large step in the field of neuroscience as a whole. The data will need to be cleaned and processed for it to be viable for our research. For each of the mice in the dataset, we will identify the spikes in neural activity per session, and filter out the insignificant spikes. We will also be able to identify the brain region of the activity.

1.2.2 EDA of International Brain Laboratory Data:

The visualizations we made were of Raster Plots and Time-scaled Histograms of neural activity in both the SCdg and SCiw regions of the brain.

Raster Plots were used to show the general neural activity in relation to the event.

Spike Histograms were used to show the neural spikes before and after the stimulus.

We also made visualizations separating the graphs into brain activity based on whether the trial was a Left or a Right condition.

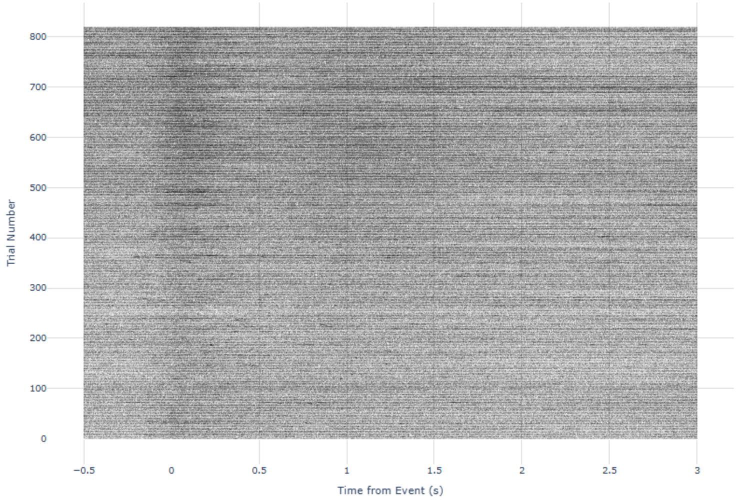

Figure 1: Raster Plot for Cluster 935, SCdg

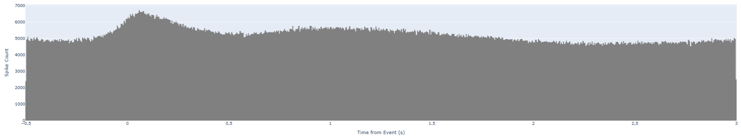

Figure 2: Spike Histogram for Cluster 935, SCdg

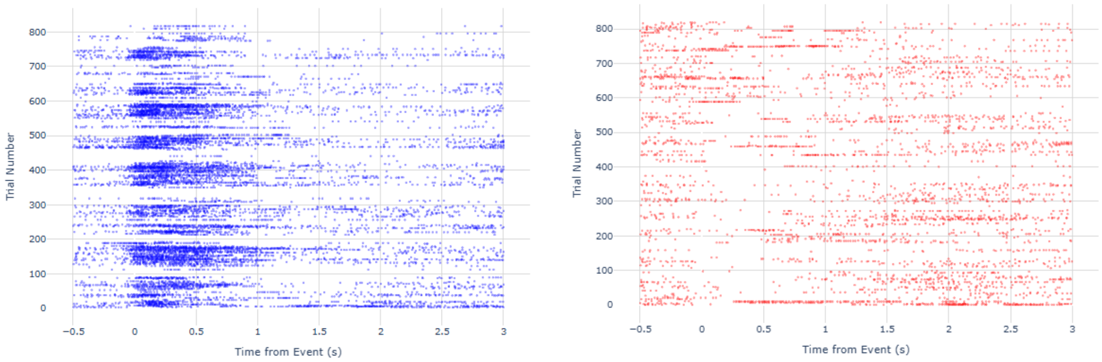

Figure 3: Raster Plots for Left (blue) and Right (Red) trials, Cluster 935, SCdg

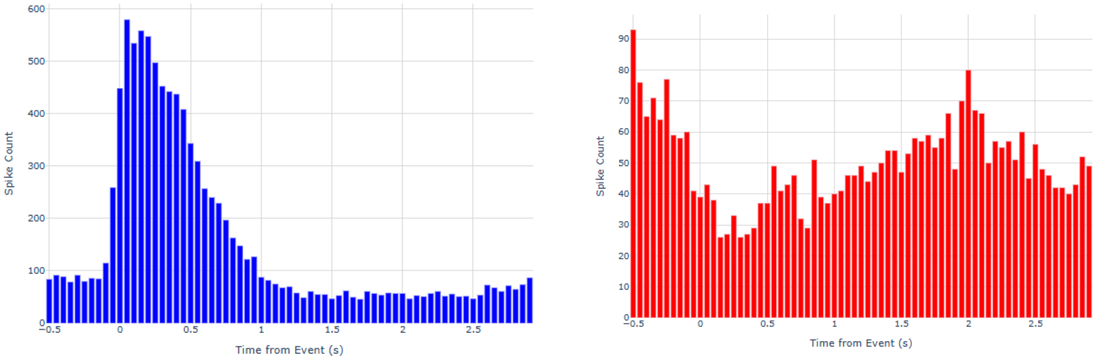

Figure 4: Spike Histograms for Left (blue) and Right (Red) trials, Cluster 935, SCdg

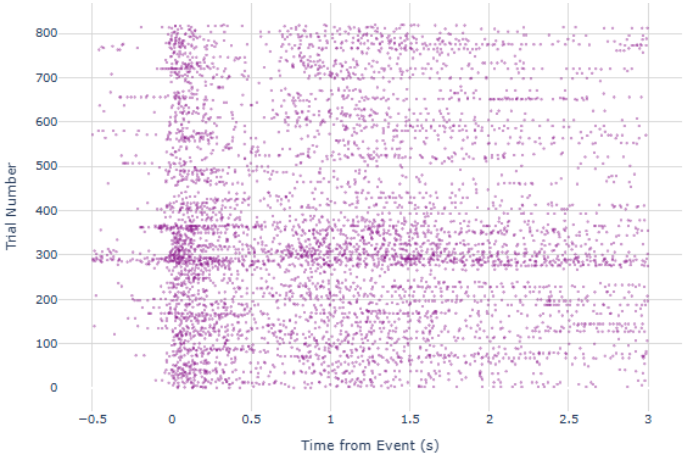

Figure 5: Raster Plot for Cluster 935, SCiw

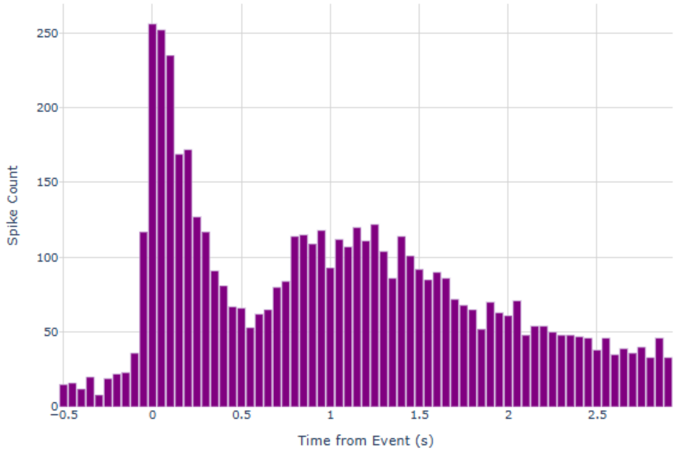

Figure 6: Spike Histogram (Cluster 1092), SCiw

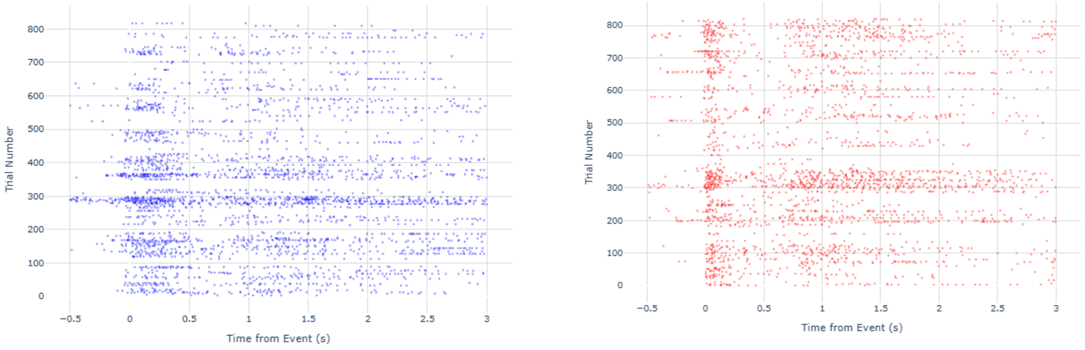

Figure 7: Raster Plots for Left (blue) and Right (Red) trials, Cluster 1092, SCiw

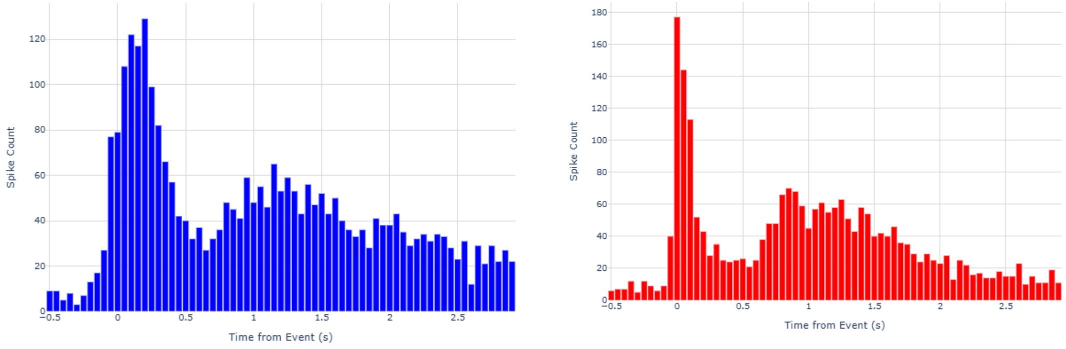

Figure 8: Spike Histograms for Left (blue) and Right (Red) trials, Cluster 1092, SCiw

Evaluation of Visualizations: In general, these graphs show a variability in neural activity across the trial, showing the stochastic nature of neural activity. They also show clear differences between the conditions (left vs right trials). Overall, these graphs show that there is neural activity in the SCdg and SCiw regions of the brain post-stimulus, showing that these regions play a role in decision making in the brain.

1.2.3 Simulated Data

We also simulated neural spike trains to represent two experimental conditions: ”left” and ”right” (binarily labeled 0 for left, 1 for right). Each trial lasted for 3 seconds and was divided into 0.05-second time bins, resulting in a total of 60 bins per trial. Using a Poisson process, spike counts were generated for 50 neurons across 100 trials per condition (left/right). The mean firing rate (lambda) for the left condition was set to 5 spikes per bin, and for the right condition, it was set to 8 spikes per bin. These parameters ensures a measurable difference between the two conditions. The formula for the Poisson distribution is shown below (Turney (2023)):

$$P(X = K) = \frac{e^{-\lambda}\lambda^{k}}{k!} \tag{1}$$

The data was then reshaped to a 2D array (flattened) where each row represents a trial, and each column contained all of the neurons' combined activity, to prepare the data to be suitable for clustering algorithms. The Expectation Maximization (EM) algorithm was employed using a Gaussian Mixture Model (GMM) to identify clusters in the high-dimensional data. This unsupervised approach aimed to distinguish trials corresponding to each condition.

Poisson Distribution:

The reason why the Poisson distribution was chose was because it is commonly used to model spike events in neuroscience, as it handles discrete events (spikes) that occur independently at varying rates, the firing rate can vary between conditions (left vs. right trials), which reflects differences in neuronal activity during different experimental states, and it captures the randomness inherent in neural activity, with the rate of events scaling linearly with time.

In this simulation, we generate the spike train data for each trial by drawing from a Poisson distribution, where the rate (mean number of spikes per second) is different for the left and right conditions. For example, in the left trials, the firing rate is set to 5 spikes/sec, and in the right trials, it is set to 8 spikes/sec. This allows us to simulate varying levels of neuronal activity between the two conditions.

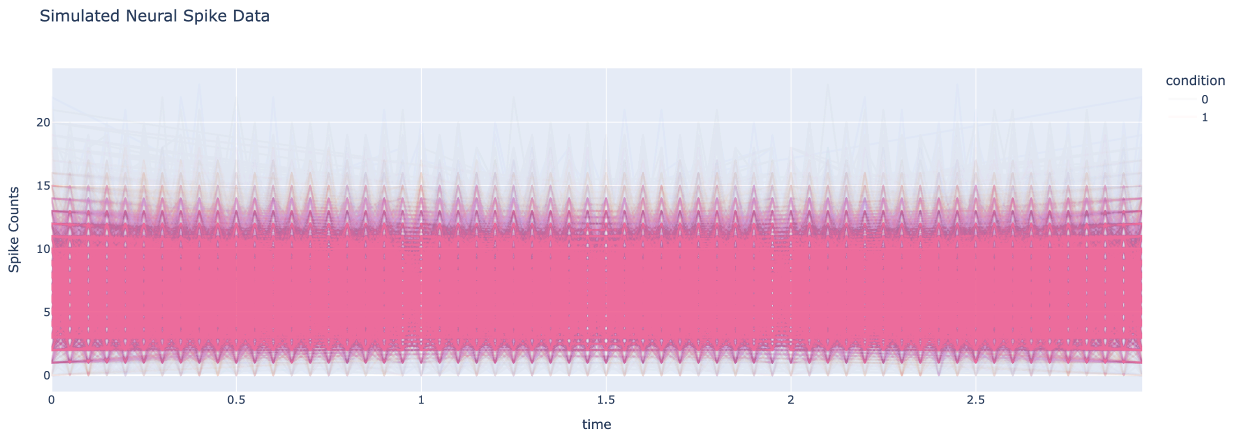

Figure 9: Simulated Neural Spike Data

2 Methods

2.1 Expectation Maximization (EM)

EM is used to perform clustering of trial data based on the spike counts across neurons.

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\log p(X,Z|\theta)|X,\theta^{(t)}]$$

where:

- θ is the set of parameters to be estimated
- $\theta^{(t)}$ is the current estimate of the parameters
- X is the observed data
- Z is the latent variables
- $p(X,Z|\theta)$ is the joint probability distribution of the observed and latent data given the parameters

Steps in the EM Algorithm:

Initialization: Randomly initialize cluster assignments for each trial.

Expectation Step (E-step): Compute the probability of each trial belonging to each cluster based on the current parameters of the model (mean and covariance of each cluster).

Maximization Step (M-step): Update the parameters of the model (means, covariances) based on the cluster assignments from the E-step.

Convergence: Repeat the E-step and M-step until the algorithm converges (when the log-likelihood no longer improves).

We use GMM with two components to model the two trial conditions (left and right). The firing rates are clustered into two groups based on the spike train data. This allows us to examine whether the model can effectively separate the trials into the two experimental conditions based on the firing rates. (Bishop (2007))

2.2 Gaussian-Process Factor Analysis (GPFA)

GPFA is useful in neuroscience, as it aims to extract these latent trajectories by assuming the high-dimensional neural activity is generated from a low-dimensional process. This allows for us to get insights into the underlying dynamics of neural activity.

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\log p(X, Z|\theta)|X, \theta^{(t)}]$$

where:

- θ is the set of parameters to be estimated
- $\theta^{(t)}$ is the current estimate of the parameters
- X is the observed data
- Z is the latent variables
- $p(X, Z|\theta)$ is the joint probability distribution of the observed and latent data given the parameters

In our code, we used traditional Factor Analysis, instead of GPFA. This limitation doesn't allow for us to make temporal connections as GPFA treats latent factors as time dependent instead of like a time-series dataset.

3 Results

3.1 Simulated Data Results

3.1.1 Expectation Maximization Clustering

As shown in the figure above›, the GMM algorithm was able to successfully separate the trials into the two conditions.

Future Work: For the future, we would focus on trying to apply the GMM and EM algorithms to the IBL data.

3.1.2 Latent Factors from Factor Analysis

We can see from the figure above, that there is a clear distinction in the latent Factor 1 after Factor Analysis. However, since this data is simulated there is a limitation in how we can truly interpret our Factor Analysis algorithm.

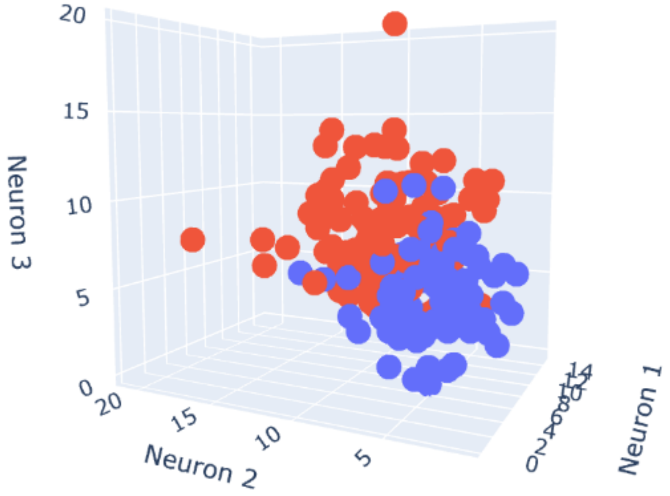

Figure 10: EM Clustering of Simulated Data, Red represents Left condition and Blue represents Right condition

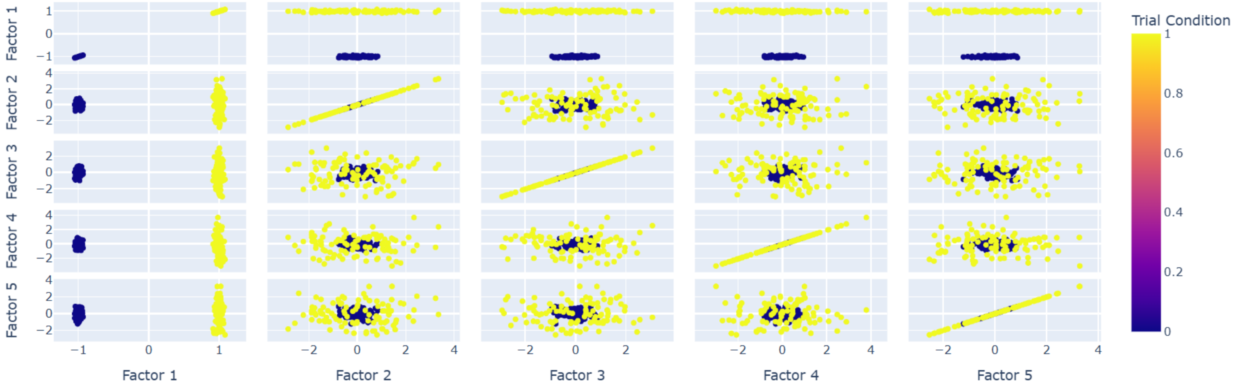

Figure 11: Latent Factors from GPFA

Future Work: For the future, we would focus on trying to create a GPFA algorithm instead of a Factor Analysis algorithm and then apply the algorithm to the IBL data. The current application of the algorithm is very clear cut, but in the future with real-world data, the data may not be as clearly separated by latent factors.

4 Discussion and Conclusion

In this paper, we explored the application of Expectation Maximization (EM) and Gaussian Factor Analysis (GFA) on simulated neural spike train data. Our results demonstrated the ability of these methods to identify neural activity patterns associated with different experimental conditions ('left' and 'right' trials). The EM algorithm successfully clustered the trials, while GFA revealed underlying latent factors that captured the variability in the data.

Our work had heavy limitations however, mostly due to the fact that we used simulated data for the EM and GFA algorithms. Although we did some data analysis on real-world mice through the International Brain Laboratory (IBL) dataset, overall our conclusions were that the brain activity in the Superior Colliculus Deep Gray Layer (SCdg) and the Superior Colliculus Intermediate White Layer (SCiw) were different due to the stimulus.

Future work will focus on addressing these limitations. We plan to apply the EM and GMM algorithms to the IBL data, allowing us to analyze real neural activity patterns. We will also implement GPFA to capture the temporal dependencies in the data, potentially revealing more nuanced insights into neural dynamics.

Furthermore, we aim to investigate the relationship between neural activity in the SCdg and SCiw regions. This could involve analyzing the correlation or shared latent factors between these regions, providing a deeper understanding of how different brain areas interact during decision-making.

References

Bishop, Christopher M. 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st edition. [Link]

Cunningham, John P, and Byron M Yu. 2014. “Dimensionality reduction for large-scale neural recordings.” August. [Link]

Laboratory, The International Brain, Valeria Aguillon-Rodriguez, Dora Angelaki, Hannah Bayer, Niccolo Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle Chapuis, Anne K Churchland, Yang Dan, Eric Dewitt, Mayo Faulkner, Hamish Forrest, Laura Haetzel, Michael Häusser, Sonja B Hofer, Fei Hu, Anup Khanal, Christopher Krasniak, Ines Laranjeira, Zachary F Mainen, Guido Meijer, Nathaniel J Miska, Thomas D Mrsic-Flogel, Masayoshi Murakami, Jean-Paul Noel, Alejandro Pan-Vazquez, Cyrille Rossant, Joshua Sanders, Karolina Socha, Rebecca Terry, Anne E Urai, Hernando Vergara, Miles Wells, Christian J Wilson, Ilana B Witten, Lauren E Wool, and Anthony M Zador. 2021. “Standardized and reproducible measurement of decision-making in mice.” *eLife* 10, p. e63711. [Link]

Shadlen, Michael N., and William T. Newsome. 2001. “Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey.” *Journal of Neurophysiology* 86(4): 1916–1936. [Link]

Turney, Shaun. 2023. “Poisson Distributions | Definition, Formula Examples.” `https://www.scribbr.com/statistics/poisson-distribution/`

Wikipedia. 2024. “International Brain Laboratory — Wikipedia, The Free Encyclopedia.” `http://en.wikipedia.org/w/index.php?title=International%20Brain%20Laboratory&oldid=1189350630`

Yu, Byron M., John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. 2009. “Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity.” July. [Link]