Computational Analyses Supporting Diverse Synaptic Clusters

Adam Li, Tyler Tomita

May 18, 2016

Abstract

Currently, very little is understood about the synaptic connections within our brain. Our original belief is that there are only two types of synapses: excitatory and inhibitory. However, it is now known that there is a much more diverse synapse population. We want to characterize these different subpopulations. Here we present an analysis of Array Tomography data that supports hypotheses that there is a richer set of clusters then just excitatory and inhibitory.

I Introduction

Many hundreds of distinct proteins are involved in the development of synapses and mechanisms in synaptic signaling. Although complex, this molecular architecture can enable a better characterization of synaptic populations and subgroups based on protein clustering. We are aware of excitatory and inhibitory synapses, but it is now known there is a much more diverse populations [4; 6].

Differences in protein expression patterns at individual synapses could constitute a key to understanding synaptic diversity. The dataset was produced by Weiler, et al. using array tomography (ATomo) and the columnar organization of mouse barrel cortext with stains of over a dozen synaptic molecules [5].

ATomo is well suited to proteomic mapping of synaptic circuits because ultrathin sectionining of resin-embedded tissue samples enable immunohistochemical multiplexing and high-resolution imaging of millions of synapses [1].

It is difficult to sift through millions of synaptic locations and determine subgroups by eye. With the development of increasing computational power and techniques, it is becoming easier to analyze large datasets and glean information on trends and clusters. Here, we employ a data driven approach to computationally analyze the synapse population of this dataset to glean insights into clusters and subclusters.

II Methods

II.A Data

Data was gathered from http://openconnecto.me/, the OpenConnectome project, and preprocessing was done to identify synapses based on the Synapsin expression levels in the ATomo pipeline [5]. The resulting dataset was a 1119299 x 96 matrix, with the rows representing a single synapse location and the columns representing the protein's integrated brightness (f0), local brightness (f1), distance to center of mass (f2) and moment of inertia around synapse (f3); these different measures are the metrics we have for each protein expression. There were 24 different protein measurements done per f0-f3. The protein markers belong to one of seven functional groupings as outlined in Table 1. In addition to ATomo imaging data, each synapse (or row) has an estimated location in the image space represented by a matrix 1119299 x 3. They represent 3D pixel locations at the nm scale.

We will call the general feature matrix, A. We will call the location matrix, B.

II.B Preprocessing

In our analysis, we began with an overall analysis of the four different metrics and then analyzed the metric we deemed most important separately afterwards. We utilize a kmeans algorithm, with k decided using Bayesian information criterion (BIC) to perform unsupervised clustering of the rows of matrix A.

We chose to focuse our analysis on integrated brightness afterwards for computational feasibility and highly correlated features; this new matrix is 1119299 x 24 matrix. Since, the scales of each protein channel for f0 varied widely, we applied a log-normalization transformation, which involved scaling to [0,1] and then applying a log transformation to the entire dataset. We first checked the data for f0. Certain proteins were measured twice using the same metric, such as Synapsin and VGlut1. In order to verify the validity of the integrated brightness measurement, we made a downsampled pairwise scatter plot of the Synapsin and VGlut1 f0 measurements.

II.C Cluster Analysis and Principle Component Analysis

In our cluster analysis, we used the MinibatchKMeans algorithm available in Python [3]. To determine optimal cluster number k, we implemented a Bayesian Information Criterion (BIC) analysis that scored the likelihood of the data within a cluster and the cost of number of clusters there were [2].

Although there are 24 different protein measurements, we sought to reduce the dimensionality of via principle component analysis (PCA) [7]. Here we plotted scree plots to visualize the number of components needed to account for the variance in our data.

III Results

In the analysis of all metrics f0, f1, f2 and f3, we generated a pairwise correlation plot to determine correlations within our feature set. Figure 2 shows a pairwise correlation plot and a thresholded pairwise correlation plot. It seems that f0 highly correlates with itself, which is something we should expect since it is the same metric. We therefore, choose to first focuse on f0 (integrated brightness).

In our validation of integrated brightness measurement, we saw a linear trend in the pairwise scatter plot of both repeated protein measurements. This is what we expected to see since the ATomo should only produce a linear range of values.

Looking at an initial clustering analysis of the lognormalized f0 data, we see a BIC plot of the data. An elbow can be hypothesized to be located at a k=4, which is then used to cluster the data with MinibatchK-Means. In figure 5, it shows the pairwise Euclidean distance plot between the different centroids used in KMeans of 4.

When we implemented a PCA on our 1119299 x 24 data matrix, we found that only 4 principle components could account for 90 percent of our variance in our data, while 8-9 prinnciple components accounted for almost all the data variance. This means that any clustering based on variance of our data could be reduced down from 24 dimensions to 9, or even 4, if we only care about up to 90 percent of data variance.

IV Discussion

We have shown an initial analysis of the protein expression data at synapses identified using ATomo. Overall, we have shown evidence that there are indeed more then just an excitatory and inhibitory clustering of synapses. We have seen that integrated brightness is highly correlative with itself relative to the other metrics involved (local brightness, distance to center of mass and moment of inertia around synapse). By focusing our analysis on integrated brightness, we saw that an approximate optimal number of clusters was 4.

When we tested that initial assumption by plotting pairwise distance plots of the centroids, we saw that they were relatively distanced against each other. In addition, when reducing the dimensionality of the protein channels, we saw that 4, or 5 principle components accounted for 90% of the data variance. If the synapse populations are characterized by integrated brightness variance in protein expression, then this would also support our inclination that our dataset captures at least 4 unique clusters.

In upcoming work, we would like to 1. validate our clustering scheme and how the distribution of protein expression looks within each cluster and 2. test different subclustering algorithms. By validating our clustering scheme, we can determine if there is a clustering of just excitatory and inhibitory synapses first. This would mean most of the proteins expressed in excitatory are in one cluster, and vice versa. We can also try various subspace clustering algorithms that utilize l1 and l2 regularizations for sparsity and smoothness [8]. This showed a computationally efficient way of subspace clustering using elastic net and can be easily implemented.

Another interesting direction would be to take into account the connectivity among all the synapses. Since, we know that synapses are physically connected, we could build a connectivity matrix that characterizes these connections and perform spectral clustering.

All work is seen at: https://github.com/Upward-Spiral-Science/the-vat.

A Bayesian Information Criterion Definition

Variables:

- N: total number of data points
- m: total number of clusters

- n_i: size of each cluster i
- d: total number of features per data point
- D: represents the variable for clusters

The BIC score is formally defined as $\text{BIC}(\phi) = \hat{l}_\phi(D) - \frac{p_\phi}{2} * log(N)$
With:

$$\hat{l}_\phi(D) = \sum_i (n_i * log(n_i)) - Nlog(N) - \sum_i \frac{(n_i d)}{2} * log(2\pi\hat{\sigma^2})$$

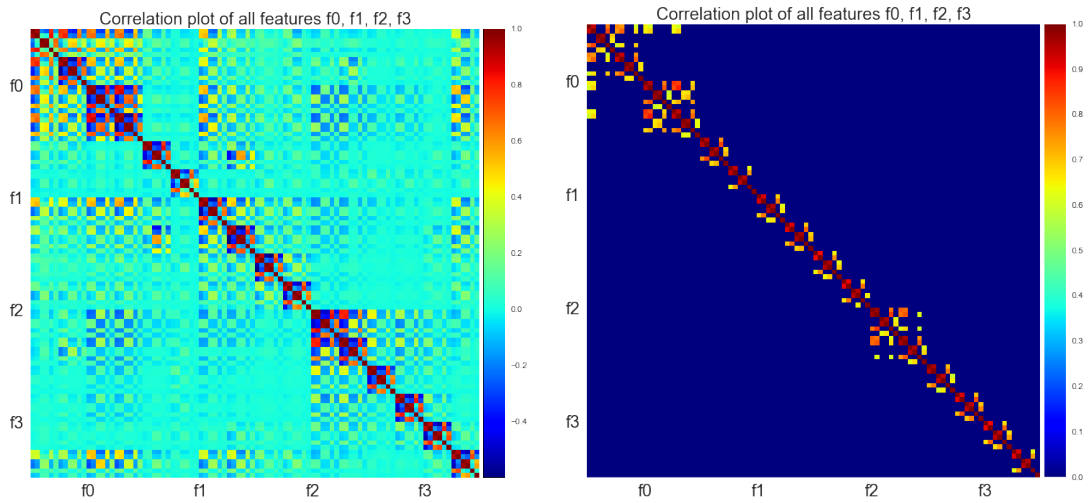
$$\frac{p_\phi}{2} * log(N) = \frac{d}{2}(N-m) - 0.5m(d+1)log(N)$$

$$\hat{\sigma^2} = \frac{1}{N-m} \sum ||x_i - \mu_i||^2$$

B Figures

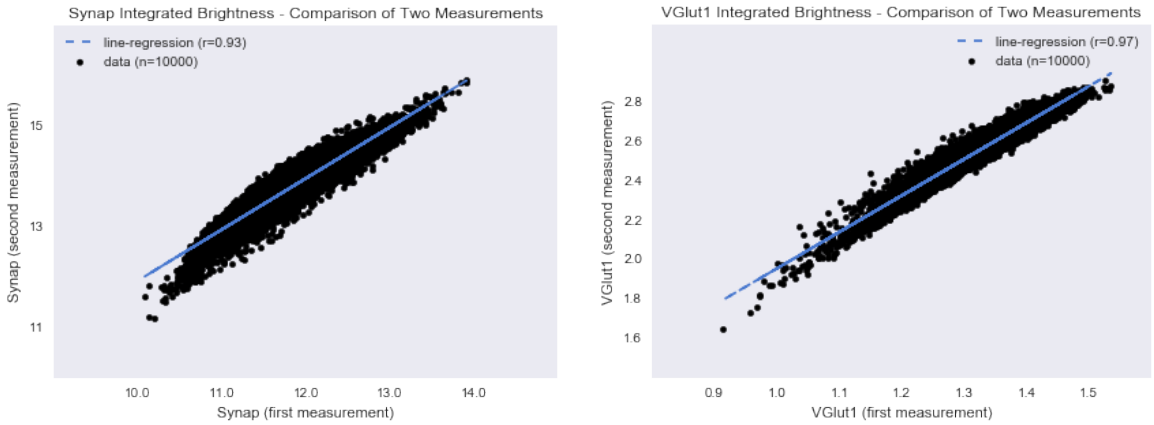
Functional Category	Markers
Excitatory Presynaptic	Synap_0, Synap_1, VGlut1_0, VGlut1_1, VGlut2
Excitatory Postsynaptic	psd, glur2, nmdar1, nr2b, NOS, Synapo
Inhibitory Presynaptic	gad, VGAT, PV
Inhibitory Postsynaptic	Gephyr, GABAR1, GABABR
Inhibitory Presynaptic (small)	Vglut3, CR1
Other	5HT1A, TH, VACht
None	tubulin, DAPI

Figure 1: Table 1: Table showing the data collectors providing domain knowledge regarding groupings of the 24 protein markers. Each marker belongs to one of seven functional groupings.



(a) (b)

Figure 2: Pairwise correlation plots of the entire feature set, with (a) being the correlation plot without a threshold, and (b) having a threshold of 0.6 applied. All correlations less than or equal to 0.6 were set to 0.



(a) (b)

Figure 3: Pairwise scatter plots of the repeated measurements of Synapsin and VGlut1, showing the integrated brightness.

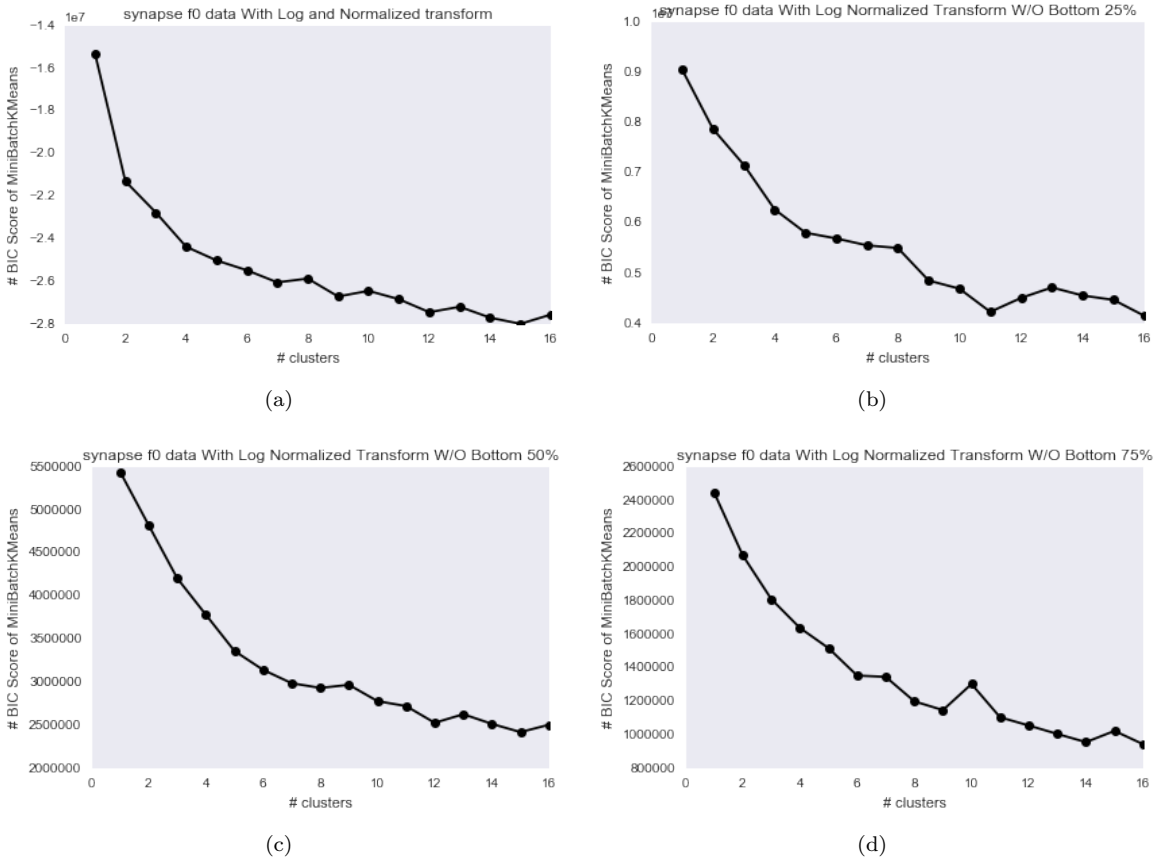


Figure 4: Figure (a) showing a BIC plot generated using MinibatchKMeans with a k=1,..,16 on the integrated brightness values. (b,c,d) show the BIC plots with the bottom 25, 50, and 75 percent Synapsin values filtered out respectively.

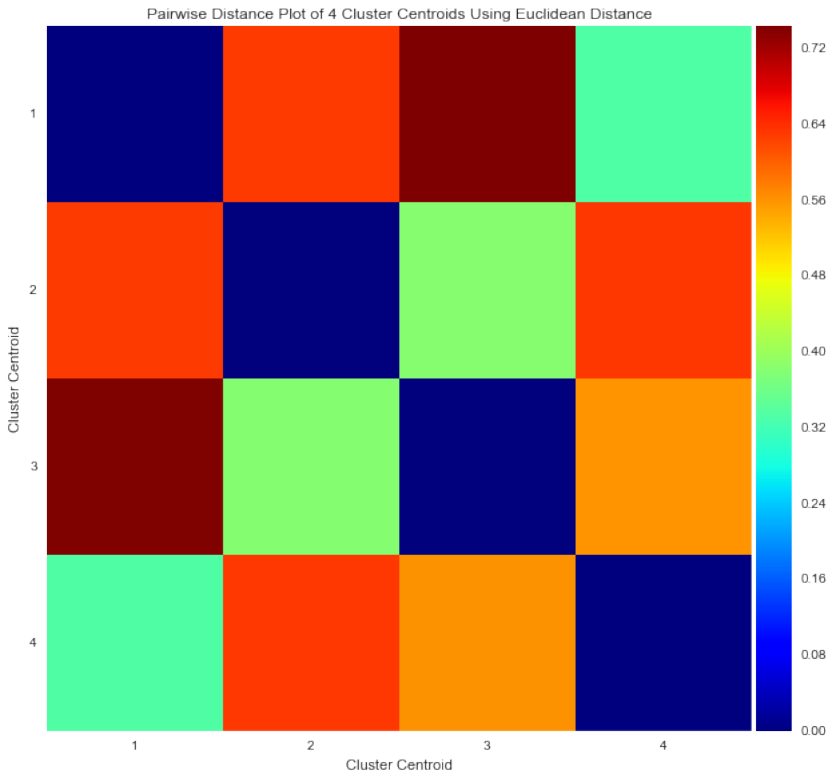


Figure 5: KMeans clustering with k=4. Along the diagonal all pairwise distances are equal to 0.0, while the off diagonals show relative euclidean distance. All off diagonals are relatively spaced away from each other with a minimum euclidean distance of 0.3.

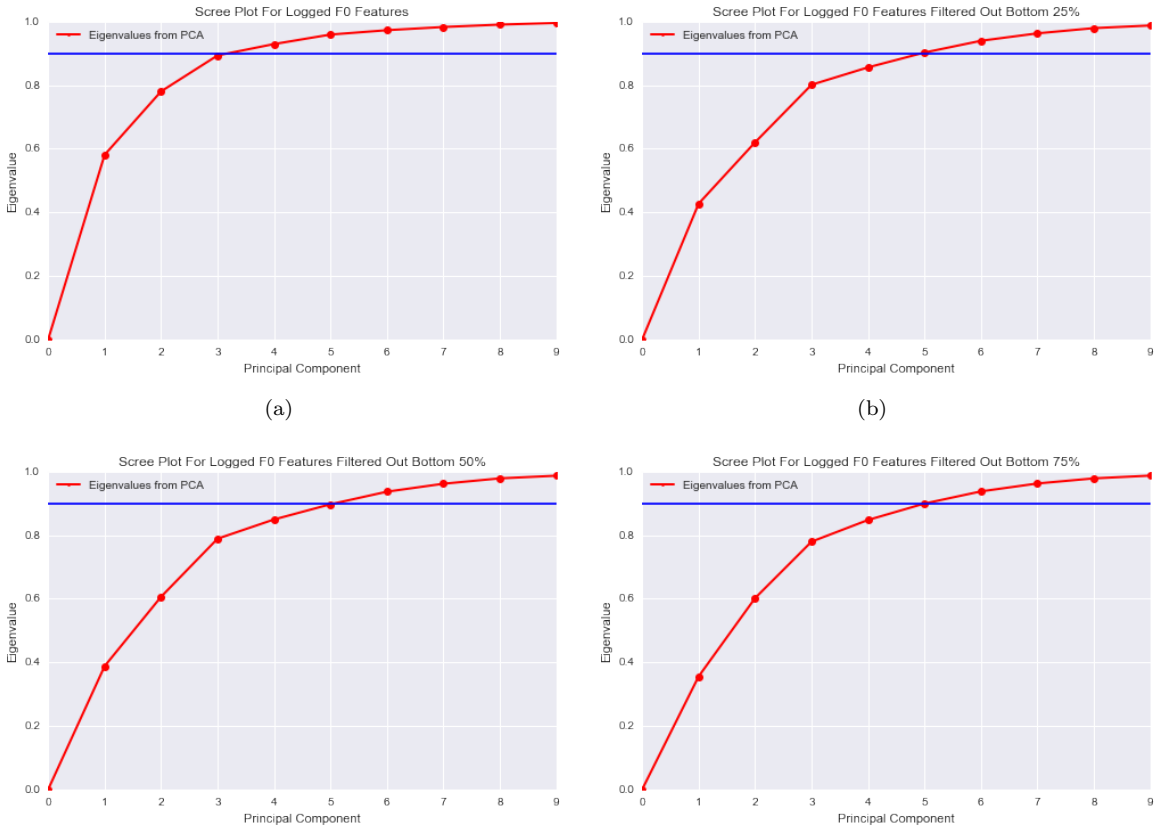


Figure 6: Figure (a) showing a scree plot on the entire lognormalized dataset on the integrated brightness values. (b,c,d) show the scree plots with the bottom 25, 50, and 75 percent Synapsin values filtered out respectively. The solid line represents 90% variance.

References

[1] Kristina D. Micheva and Stephen J. Smith. Array Tomography: A New Tool for Imaging the Molecular Architecture and Ultrastructure of Neural Circuits. *Neuron*, 55(1):25–36, 2007.

[2] Moore A. Pelleg D. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Journal of Chemical Information and Modeling*, 53(9):1689–1699, 2013.

[3] D Sculley. Web-scale k-means clustering. *Proceedings of the 19th international conference on World wide web WWW 10*, page 1177, 2010.

[4] Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, Sten Linnarsson, and Patrik Ernfors. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Publishing Group*, 18(1):145–153, 2014.

[5] Nicholas C Weiler, Forrest Collman, Joshua T Vogelstein, Randal Burns, and Stephen J Smith. Synaptic molecular imaging in spared and deprived columns of mouse barrel cortex with array tomography. pages 1–20, 2014.

[6] Diek W Wheeler, Charise M White, Christopher L Rees, Alexander O Komendantov, David J Hamilton, and Giorgio A Ascoli. Hippocampome . org : a knowledge base of neuron types in the rodent hippocampus. pages 1–28, 2015.

[7] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[8] Chong You, Chun-guang Li Daniel, and P Robinson Ren. Oracle Based Active Set Algorithm for Scalable Elastic Net Subspace Clustering. 2016.