Computational Analyses Supporting Diverse Synaptic Clusters In Mouse Cortex

Adam Li
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21210
ali39@jhu.edu

Tyler Tomita
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21210
tmtomita87@gmail.com

Joshua Vogelstein, PhD
Department of Biomedical Engineering
Johns Hopkins University
Baltimore, MD 21210
jovo@jhu.edu

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 by protein expressions. Here we present an analysis of protein expressions under Array Tomography at synaptic locations. Our initial results show evidence pointing towards a richer set of sub-synaptic clusters then just excitatory and inhibitory.

1 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 set of subpopulations [6, 4].

Differences in protein expression patterns at individual synapses could constitute a key to understanding synaptic diversity. The dataset analyzed in this paper was produced by Weiler, et al. using array tomography (ATomo) to image the synaptic molecular architecture of neighboring whisker-associated columns of the mouse somatosensory cortex [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]. These images were then reconstructed into precisely aligned image volumes, with data on millions of synapses in the image volume.

It is difficult to sift through millions of synaptic locations and determine subgroups by eye. With the development of increasing computational power and machine learning 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 synaptic clusters.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Methods

2.1 Data

Data was gathered from the Array Tomography section of http://openconnecto.me/, the OpenConnectome project. The data we received were two matrices that represented the features and the <x,y,z> image volume coordinate locations for each synapse location. Preprocessing was done to identify synapses based on the Synapsin expression levels in the ATomo pipeline, based on a set threshold [5]. The resulting feature matrix was a 1119299 x 96 matrix, with the rows representing a single synapse location and the columns representing 24 protein measurements of integrated brightness (f0), local brightness (f1), distance to center of mass (f2) and moment of inertia around synapse (f3). 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.

2.2 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 implemented a pairwise correlation computation to look at correlations among features. Then we applied an arbitrary threshold of 0.6 to see which features correlated higher then the thresdhold. By analyzing only one metric among f0-f3, the feature matrix of interest is now 1119299 x 24. Since the scales of each protein channel measurement varied, we applied a log-normalization transformation, which meant scaling all columns to [0, 1] and then applying a log transformation.

We chose to focus on f0 (integrated brightness) because of its correlative measurements. We first checked the validity of the data. Since there were repeated measurements for the same protein, the measurements should all lie within a linear regression. We created a pairwise scatter plot for repeated Synapsin and VGlu1 measurements and looked at the linear regression to check for linearity and hence valid data measurements for f0.

Since we are looking for subclusters of synapse types, it would be biased to only have synapses heavily located in one region of the tissue volume. Therefore, we inspected whether or not there seemed to be a uniformly distributed sample of synapses throughout the tissue volume by plotting histograms along the <x,y,z> axis.

2.3 Cluster Analysis and Principle Component Analysis

Due to the size of our dataset, it was computationally infeasible to apply a regular KMeans unsupervised clustering algorithm to the dataset. Therefore, we used the MinibatchKMeans algorithm available in Python, which has been shown to achieve similar results in shorter computation time [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]. Once an optimal k was estimated, we performed the MinibatchKMeans clustering on the preprocessed feature matrix, and looked at the Euclidean distance between centroids of clusters.

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.

3 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 1 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) for the remainder of our analyses.

In our validation of integrated brightness measurement, we saw a linear trend in the pairwise scatter plot of both repeated protein measurements, Synapsin and VGlut1. This is what we expected to see

since the ATomo should only produce a linear range of values. We then looked at the distribution of synapses throughout the image volume and see that it is looks distributed throughout the image volume without any visual clustering in one region over another.

Next we analyzed the BIC scores for different k using the MinibatchKMeans algorithm on the data. Since synapses were identified by the expression of Synapsin values, we took the synapses (rows in matrix) with top 25, 50 and 75% of Synapsin value and also analyzed the BIC score to see if it was sensitive to our thresholding of a synapse. We found that all BIC plots produced a relative elbow around k=4, so we proceeded with clustering the entire dataset with k=4. In figure 4, it shows the pairwise Euclidean distance plot between the different centroids used in KMeans of 4, where we see a maximum distance of 0.72 between cluster centroids and 0.30 distance between cluster centroids on the off-diagonal. The diagonal are all 0 because it represents cluster centroids compared with itself.

When we implemented a PCA on our 1119299 x 24 feature matrix, we found that only 5 principle components were needed to account for 90 percent of our variance in our data, while 12 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 12, or even 5, if we only care about up to 90 percent of data variance.

4 Discussion

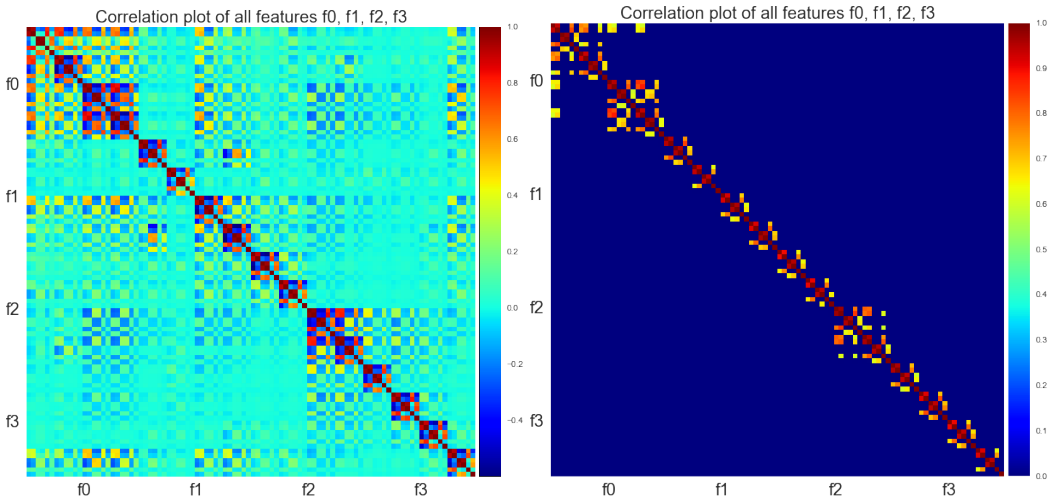
We have shown an initial analysis of the protein expression data at synapses identified using ATomo. Overall, we have shown evidence that there are indeed more then just an excitatory and inhibitory clustering of synapses. We have seen that integrated brightness is highly correlative with itself relative 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. Reducing the dimensions down to 12 would only leave <1% of variance unaccounted for, while increasing the interpretability of the data.

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.

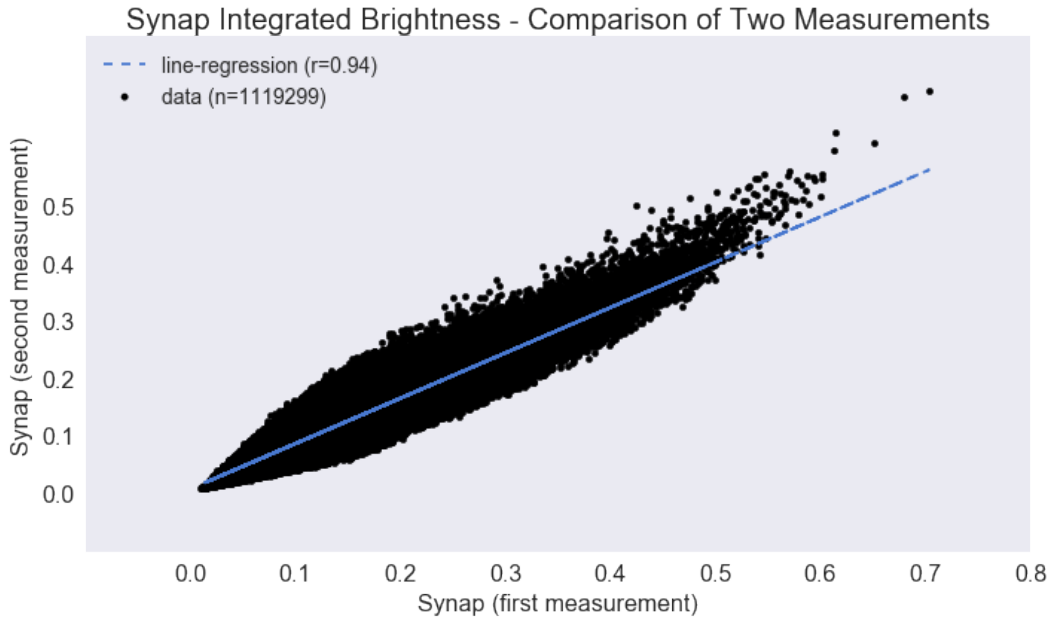
5 Citations, figures, tables, references

5.1 Figures

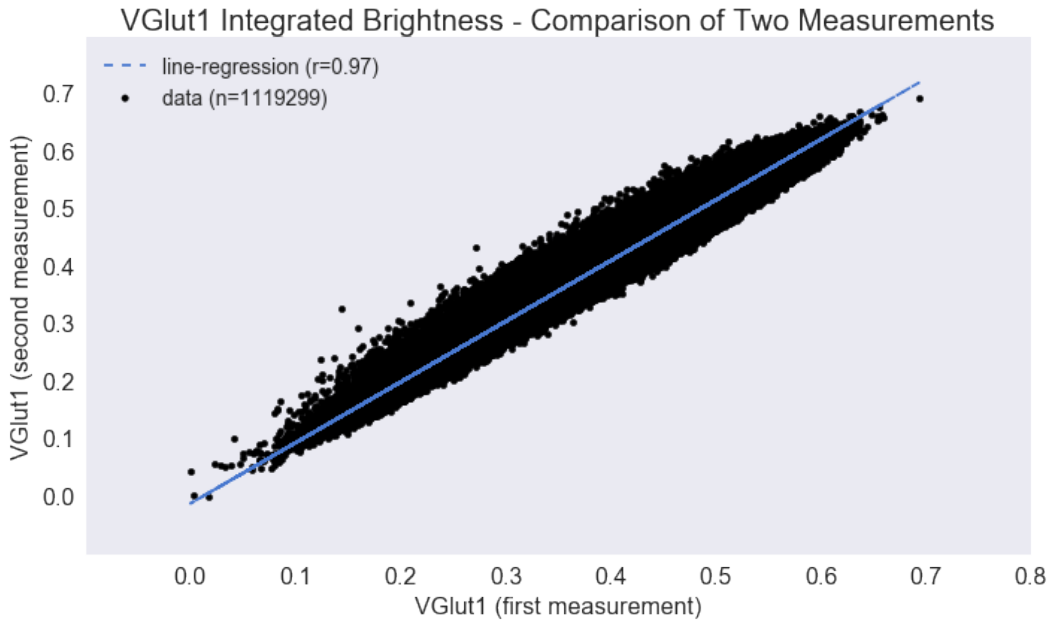


(a) (b)

Figure 1: 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 2: Pairwise scatter plots of the repeated measurements of Synapsin (a) and VGlut1 (b), showing the integrated brightness.

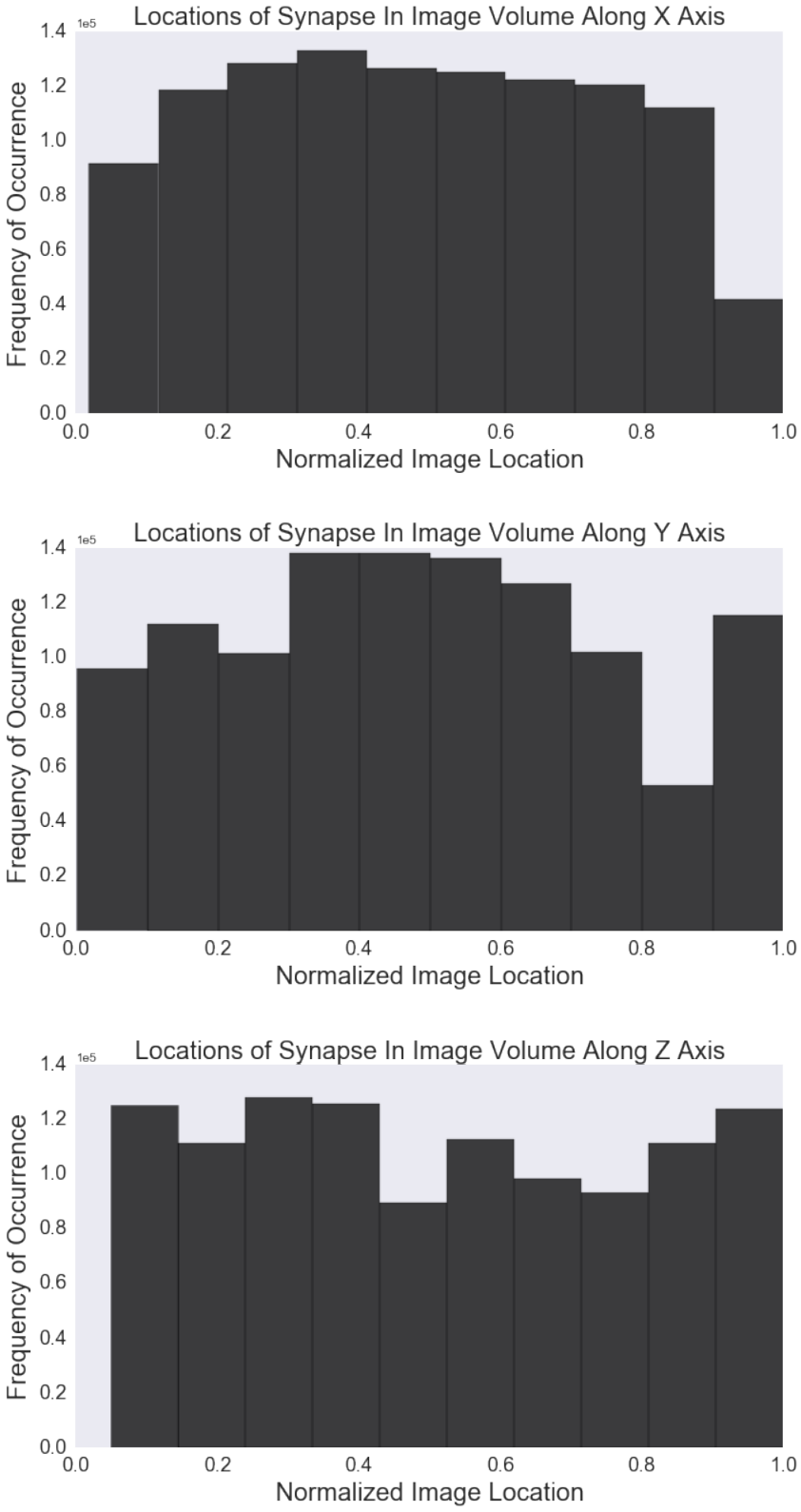


Figure 3: Histogram plots showing the distribution of synapses at locations throughout the image volume.

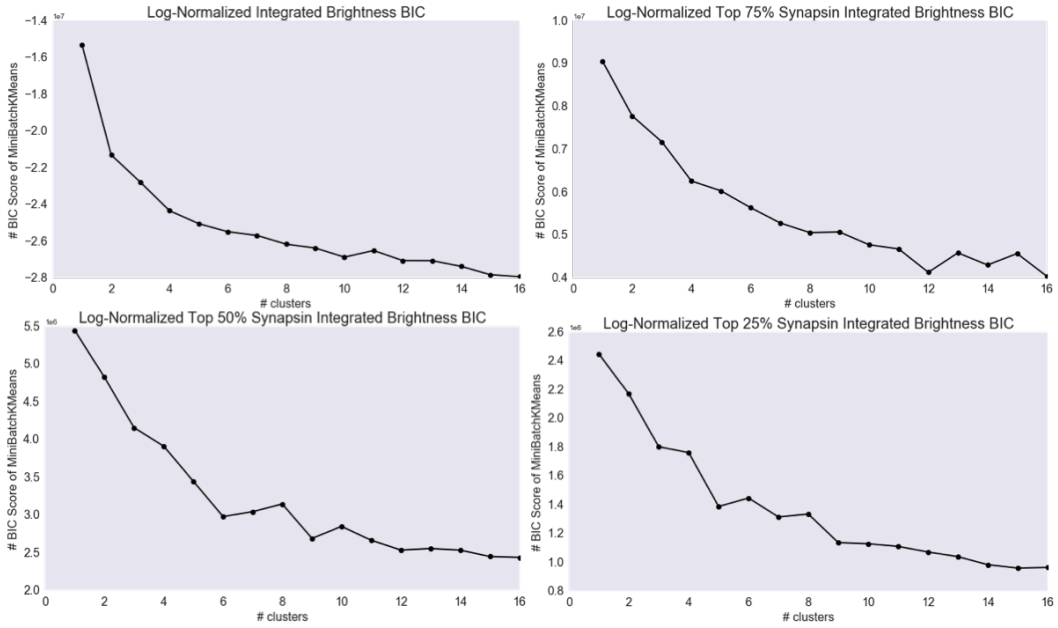


Figure 4: Figure (a) showing a BIC plot generated using MinibatchKMeans with a k=1,..,16 on the integrated brightness values. (b,c,d) show the BIC plots with the bottom 25, 50, and 75 percent Synapsin values filtered out respectively.

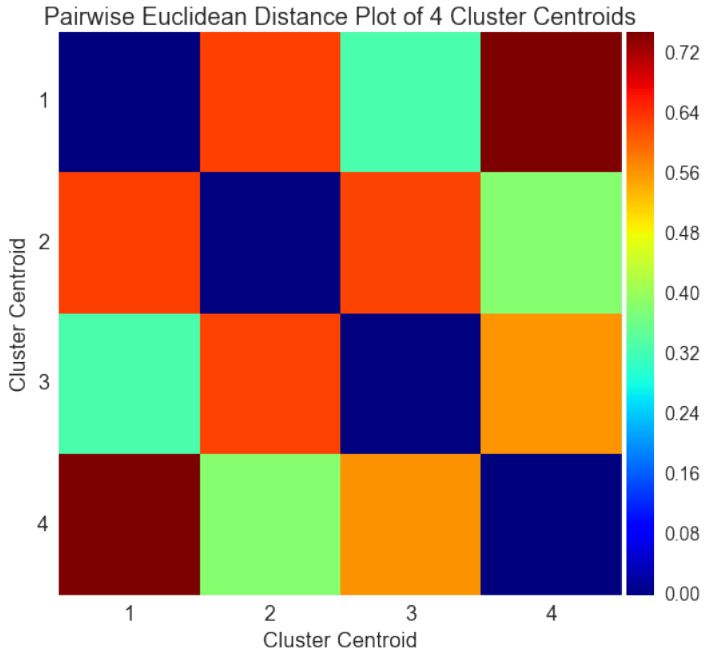


Figure 5: KMeans clustering with k=4. Along the diagonal all pairwise distances are equal to 0.0, while the off diagonals show relative euclidean distance. All off diagonals are relatively spaced away from each other with a minimum euclidean distance of 0.3.

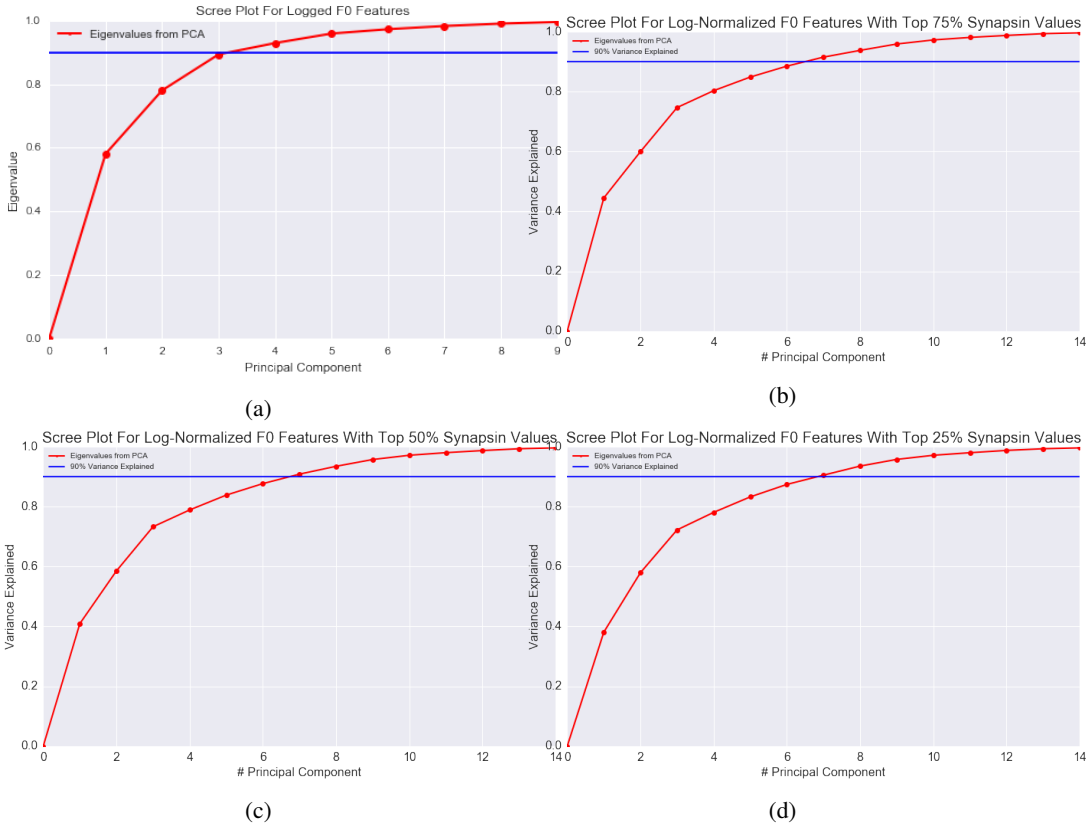


Figure 6: Figure (a) showing a scree plot on the entire lognormalized dataset on the integrated brightness values. (b,c,d) show the scree plots with the bottom 25, 50, and 75 percent Synapsin values filtered out respectively. The solid line represents 90% variance.

5.2 Tables

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

Table 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.

5.3 Equations

Bayesian Information Criterion 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 BIC(ϕ) = $\hat{l_\phi}(D) - \frac{p_\phi}{2} * log(N)$

With:

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

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

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

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.