Neural Community Detection in The Brain Using Robust Correlation Estimator from Random Matrix Theory

DUNDA, Gerry Windiarto Mohamad (20491372)

Project Report for ELEC 5450

I. Introduction

Human brain is the central of human organ that control the activities of the entire parts of the body. It is claimed that the number of neurons inside the brain can reach around billion order of magnitude. The neurons are interacting with other neurons to transmit and convey the information. The information given by the presynaptic neuron is transmitted through the axon terminals enveloped by the vesicle and received by the postsynaptic neuron. Like binary numbers as a way to represent the information in the digital communication, the information is represented by action potential which is triggered when there is a presence of the spike in the neuron electrical signal. As the neurons interact with each other, they will form different populations . The structure of the populations can be inferred based on the measure of interaction between neurons. Precisely, the neurons are collected together such that neurons with excitatory connection belongs to one group while neurons with inhibitory connection should be located on different group. Excitatory connection means high likelihood for two neurons exciting spike together and vice versa. This connection can be described mathematically through recorded time-series of neural activity for many neurons that will be used to estimate the correlation matrix of the neural population.

This report will first discuss the recent method by [1] how to detect communities among neurons in the region of interest based on sample correlation matrix constructed from the time series data. Taking preprocessed sample correlation matrix as the input, the modularity measure as the objective function will be introduced to achieve the purpose. The performance of the modularity algorithm will be dependent on null model in which Random Matrix Theory is applied for matrix filtering process. This paper will focus more on this filtering process and the numerical simulation of the community detection will be shown for different kinds of filtering process.

II. Overview of Existing Work

A. Correlation matrix from neural activity time series data

The neural activity time series data are recorded from specific region of the brain. In [1], rather than using common equipment such as EEG or fMRI with brain region as a node, they utilize single-neuron data on gene expression. This gene expression can be sampled every hour by using bioluminescence PER2::LUC. In this way, the neural activity time series data

can be obtained for each neuron inside the region of interest, in particular Suprachiasmatic Nucleus (SCN) which is responsible for circadian clock. From the data given from [1], the experiment is conducted on several different mouse. As a result, the number of detected neurons in a given brain region varies from one data to the another and the number is roughly between 100 and 200. The length of the time series also lies on this range. It is important to note that the time series data consist of real elements.

After the neural activity time series data has been obtained for each neuron, the sample correlation matrix can be directly calculated. This correlation matrix will give the information about the interaction between all neurons. For the sake of convenience, all neurons in the data being scrutinized will be labelled by sequence of natural numbers. Given neuron i and j, their interaction measure can be interpreted from the sample correlation matrix element of column i and j or, in more concise form, C_{ij}. If the value is positive, it indicates that the type of the interaction between neurons is excitatory whereas negative value indicates inhibitory interaction. Although, in general, the interaction between neurons are not symmetric, in many approximations such as mean-field theory, the interactions are assumed to be symmetric in order to meet correlation matrix property.

Knowing the sample correlation matrix, the underlying structure of the neurons can be inferred from this matrix. To begin the analysis of the matrix, one needs to perform eigen-decomposition on the matrix. Given that the total number of neurons in a system is M, the eigenvalues and eigenvectors are $\{\lambda_i\}_{i=1}^{M}$ and $\{\boldsymbol{x_i}\}_{i=1}^{M}$ respectively, the decomposition can be written as:

$$C = \sum_{i=1}^{M} \lambda_i \, \boldsymbol{x_i} \boldsymbol{x_i^T} \tag{1}$$

In other words, the matrix can be represented as the linear combination of the projection matrix given by the eigenvectors with its corresponding eigenvalues as their weights. Note that it is convenient to sort the eigenvalues in ascending order, that is $\lambda_1 < \lambda_2 < \cdots < \lambda_M$. By looking from different view, the sampled correlation matrix can be decomposed into three terms as suggested by [2]: the market-wide part or global trend, group part, and noise part consecutively.

$$C = C^{(m)} + C^{(g)} + C^{(r)} \tag{2}$$

The market-wide part contains the largest eigenvalue information, the group part contains intermediate eigenvalues, and the rest is considered as noise. The market-wide matrix can be done by straightforward construction once the maximum eigenvalue of the matrix is known. The main challenge in this decomposition is in other two matrices. Specifically, one must determine cutoff index N_g such that set of eigenvalues $\{\lambda_i\}_{i=2}^{N_g}$ are the building blocks of the group part and $\{\lambda_i\}_{i=N_g+1}^{M}$ are the building blocks of the noise part. This issue will be discussed more in part C.

B. Community detection algorithm

The next step to establish the community among the neurons is to define a community detection algorithm. There are many algorithms available to achieve this purpose. These algorithms may yield different result of communities depending on the quality measure that describes a best partitioning that separates different communities. Following the technique from [1], this paper will adopt the modularity measure proposed by [3]. This modularity measure is relevant for community detection in neurons since it depends on the sampled correlation matrix obtained from neural activity time series data. Mathematically, this can be expressed as follows.

$$Q(\boldsymbol{\sigma}) = \frac{1}{C_{norm}} \sum_{i,j}^{M} \left[C_{i,j} - < C_{i,j} >\right] \delta(\sigma_i, \sigma_j) \tag{3}$$

Where C_{norm} is the sum of all matrix elements inside the correlation matrix. This function takes $\boldsymbol{\sigma}$, mapping between the neuron index to its corresponding population label, as the input and the function will give a scalar value to indicate how well the partitioning is. The existence of Kronecker delta means that the evaluation is based on each pair of neurons that are in the same community. Intuitively, if the pair of neurons has positive correlation that is close to 1, it means that these pairs of neurons should be in the same community and pair of neurons with negative correlation should be separated. Thus. the modularity measure can be used as a criterion to determine the quality of partitioning.

After defining the modularity measure, the algorithm to obtain desired neuron populations is necessary. The goal of the algorithm is to achieve maximum modularity measure and this problem can be described as an optimization problem. In this paper, the work from [1] will be reproduced with modified Louvain method. This method is equivalent to brute-force method since it explores every possible node's configurations. Starting from treating each node as separate community, this algorithm will migrate one node from a pre-defined community to another and this process will be evaluated by the change in modularity measure. Once there is no improvement on the measure, the algorithm will terminate and give the optimal solution of partitioning. One good point for this algorithm is that the number of communities is not needed beforehand. For interested readers, the rigorous mathematical manipulation of the weights for the intermediate step is given in [3].

From equation (3), one term that is essential in constructing the modularity measure is the null model, $< C_{i,j} >$. This term serves as the filter for the correlation matrix before establishing the modularity measure. Originally, this modularity measure is employed in pure graph problem rather than correlation matrix based on time series data. The null model in this case is used to determine whether the underlying structure of the graph is pure luck (random structure) or not. In other words, all random components should be absorbed by the null model to achieve deterministic result of partitioning. In a similar way, the obvious term containing in the null model should be the noise part of the correlation matrix as in the equation (2). From [1], the market-wide component should be filtered as well because the correlation matrix is dominated by this term and without the elimination of this term, the solution converges to the trivial community where all nodes are treated as one population. This means that the null model is the sum of noise part and market-wide part of the correlation matrix.

C. Marchenko-Pastur Distribution and Group part decision

The eigenvalue distribution of the sampled correlation matrix provides another alternative for analyzing the structure of the matrix. As seen in the previous section, the importance of noise part of the matrix is apparent during the community detection process by specifying the null model of the system. Fortunately, the noise property of the matrix can be found by studying the eigenvalue distribution on the correlation matrix. To realize this purpose, suppose that the construction of each element in the time series matrix X with size $M \times N$ is standardized gaussian with zero mean and unit variance. The eigenvalue distribution of matrix $C_r = \frac{1}{N} XX^T$ can be expressed as[4]

$$\rho(\lambda) = \frac{Q\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda}; \quad \lambda_1 < \lambda < \lambda_+ \tag{4}$$

and the new introduced variables are

$$\lambda_\pm = (1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}})$$

$$Q = N/M$$

The important characteristic arising from the equation is that the probability density function gives non-zero value when the eigenvalue is between two bounded values namely λ_+ and λ_-. These bounds only depend on the size ratio of the random time series matrix. Even though the time series data is standardized, there is some point that this paper is going to emphasize. In the equation (4), since the variance is one, the variance term does not appear. The equation will take form as follows if the variance, σ^2, is also considered instead.

$$\rho(\lambda) = \frac{Q\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda\sigma^2}; \quad \lambda_1 < \lambda < \lambda_+ \tag{5}$$

$$\lambda_\pm = \sigma^2(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}})$$

From this equation, it can be seen that the additional variable σ^2 is included in both scaling on probability density function as well as the upper bound and the lower bound of the distribution.

Now, the comparison between the noise part of the correlation matrix and the matrix itself will be described and the simulation result will be discussed further in the numerical simulation section. From the previous section, it is left as a problem to determine the suitable cutoff value which separates the noise part from the group part of the correlation matrix. If the eigenvalue of the sampled correlation matrix is plotted into a histogram, the eigenvalue distribution can be obtained pictorially. Superimposing the theoretical curve, which is the eigenvalue distribution given by the Marchenko-Pastur law, on the same

graph, it will be shown that there are some of the eigenvalues that lie inside this curve ($\lambda < \lambda_+$) and some are located outside $\lambda > \lambda_+$. This means that the sampled correlation matrix contains some noise distribution as well as information related to the interaction between neurons. Thus, the cutoff decision value can be determined by following this idea.

One interesting fact that is pointed out by [1] is that the existence of the global trend, the largest eigenvalue bin in the histogram, affects the shape and the boundary of the theoretical curve. In other words, the effects of noise as well as the global trend are intertwined together. This observation is taken to the consideration by the author to pull out another eigenvalue bin from inside the theoretical curve. In this way, it is claimed by the author that some of the bins may not be misclassified as the noise. This claim is supported by [5] that by adjusting the variance σ^2 of the theoretical curve, one can obtain suitable noise part of the correlation matrix. The idea is that, since the correlation matrix do not contain fully gaussian noise, the hypothesis which states that the matrix is purely noise is violated. By subtracting the unit variance with the fraction largest eigenvalue, one can obtained the modified Marchenko-Pastur distribution. This can be understood directly from the fact that the trace of the correlation matrix is equal to total number of neurons M and the sum of eigenvalue. After that, the subtraction is done uniformly on all variance for each neural activity time series distribution. Mathematically, the modification of the variance leads to the following eigenvalue distribution

$$\rho(\lambda) = \frac{Q\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda(1 - \frac{\lambda_M}{M})}; \quad \lambda_1 < \lambda < \lambda_+ \tag{6}$$

$$\lambda_\pm = (1 - \frac{\lambda_M}{M})(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}})$$

Hence, by using the modified version, the noise part and the group part can be empirically determined, and the separation is expected to give better structure.

III. Critical Evaluation of Existing Work

The construction of population correlation matrix plays major role in construction of population among the neurons. This matrix is not known and should be estimated from sampled correlation matrix. Its eigenvalues and eigenvectors are the essential components for the preprocessing before running the community detection algorithm. Specifically, the eigenvalues are essential in cutoff decision while the eigenvectors contain the correlation structure of the neurons. In [1], the estimated population eigenvectors and eigenvalues are obtained directly from the calculation of sampled correlation matrix. After that, one can empirically decide parts of the eigenvalues that belong to the group part and the other. To give more resolution in determination of noise part from the group part of the correlation matrix, the author tried to modify the theoretical curve such that the contribution from the global trend is omitted. Notice that the largest eigenvalue of population correlation matrix is

unknown and should be estimated as well. The author estimated the largest eigenvalue directly from the sample correlation matrix.

However, the estimation of eigenvalues and eigenvectors directly from the sampled correlation matrix may suffer from inconsistent estimation when dealing with different ratio of Q. The time series data provided by the author [1] is limited to 100 - 200 time samples while the number of neurons is also in this range. This means that the possible value of Q is roughly within 0.5 and 2. This direct estimation is valid only if the number of samples goes to the infinity for fixed M and this estimator sometimes referred as N-consistent. In other words, the Q value should be large enough to achieve good estimate. The consistency here means that the estimated eigenvalues as well as the eigenvectors converge almost surely to the actual values given by the population correlation matrix. For this reason, a more robust (M, N)-consistent estimator is necessary that takes the ratio Q into the consideration. This problem can be viewed as limit M and N approaches infinity with fixed constant ratio Q.

IV. Improved Estimators

In this paper, a main contribution from [6] will be discussed to improve the correlation estimator. Basically, the starting point of the approach to the estimation problem is to consider the empirical spectral distribution $dF_M(x)$ of the eigenvalues derived from the sample correlation matrix which can be expressed as

$$\frac{dF_M(x)}{dx} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\widehat{\lambda_i}}(x) \tag{7}$$

This distribution can be interpreted as the height of the histogram of the eigenvalue distribution. It is convenient to analyze this function in terms of Stieltjes transform. Performing this transform on any of two spectral distribution will yield the following

$$m(z) = \frac{1}{M} Tr((S_M - zI_M)^{-1}) \; ; \; z \in \{x \in \mathbb{C} \mid Im(x) > 0\} \tag{8}$$

where S_M is the sampled correlation matrix.

The author from [6] worked on the limiting case as $M, N \to \infty$ with their ratio constant for the inverse Stieltjes transform on the empirical spectral distribution. It should be noted that the direct estimated eigenvalues and eigenvectors are denoted as $\widehat{\lambda_m}$ and $\widehat{\boldsymbol{e}_m}$ respectively, where the subscript $m \in \{1, \ldots, M\}$. With the existence of a support which is defined as the union of all set of clusters around the true eigenvalues, the author is able to derive more robust estimated eigenvalues ($\widehat{\gamma_m}$) as well as the eigenvectors ($\widehat{\boldsymbol{\eta}_m}$) in the following forms:

$$\widehat{\gamma_m} = N(\widehat{\lambda_m} - \widehat{\mu_m}) \tag{10}$$

$$\widehat{\boldsymbol{\eta}_m} = \sum_{k=1}^{M} \theta_m(k) \widehat{\boldsymbol{e}_m} \widehat{\boldsymbol{e}_m}^T \tag{11}$$

Where

$$\theta_m(k) = \begin{cases} -\phi_m(k), & k \neq m \\ 1 + \psi_m(k), & k = m \end{cases}$$

$$\phi_m(k) = \frac{\widehat{\lambda_m}}{\widehat{\lambda_k} - \widehat{\lambda_m}} - \frac{\widehat{\mu_m}}{\widehat{\lambda_k} - \widehat{\mu_m}}$$

$$\psi_m(k) = \sum_{r=1, r \neq m}^{M} \frac{\widehat{\lambda_r}}{\widehat{\lambda_k} - \widehat{\lambda_r}} - \frac{\widehat{\mu_r}}{\widehat{\lambda_k} - \widehat{\mu_r}}$$

and where $\widehat{\mu_1} \leq \widehat{\mu_2} \leq \ldots \leq \widehat{\mu_M}$ can be obtained from the solution of the following equation

$$\frac{1}{M} \sum_{k=1}^{M} \frac{\widehat{\lambda_k}}{\widehat{\lambda_k} - \hat{\mu}} = Q$$

Note that this expression is obtained from [6] with an additional assumption that there is no degeneracy of eigenvalues. The nice thing about this estimator is that the implementation is not difficult, and the solver does not take much time (about a few seconds). In the implementation, this process is done in covariance matrix first and then the resulting matrix will be converted to correlation matrix.

V. Numerical Simulations

From the time series data given, this report will simulate two of those. First, the work from [1] will be reproduced with the exception that the resulting mapping between neurons and their population will be demonstrated in the stem plot instead of topological coloring of neurons as in [1]. After knowing the sampled correlation matrix for each time series data, the eigenvalue distribution with the modified Marchenko-Pastur red curve is shown in the figures 1.

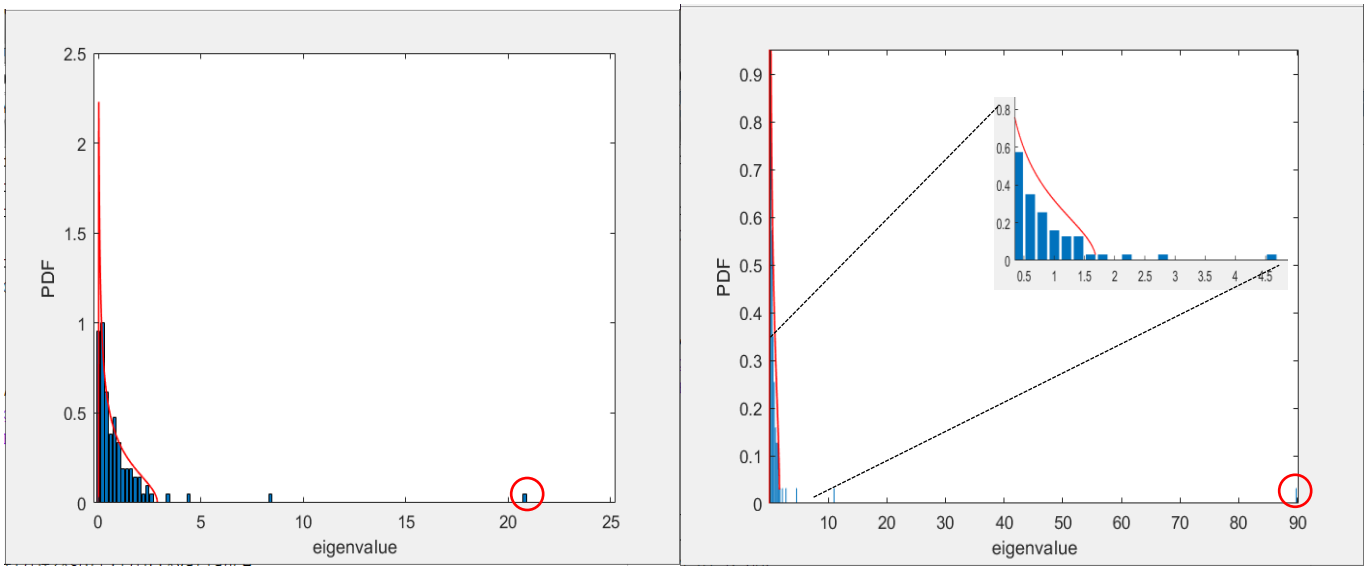

Figure 1 Histogram of eigenvalue distribution of correlation matrices from two different samples. Largest eigenvalue is labelled with red circle

Note that the sequence of showing the figures is preserved throughout this section, from left to right. From the figure, all eigenvalues inside the red curve are considered as the noise part while the deviating eigenvalues excluding the largest one are considered as group part of the correlation matrix. After the construction of the correlation matrix for each time series data, the matrix will be filtered according to defined null model and the outputs of the community detection algorithm are generated. The figure below shows the mapping between the neuron number and its corresponding population label.

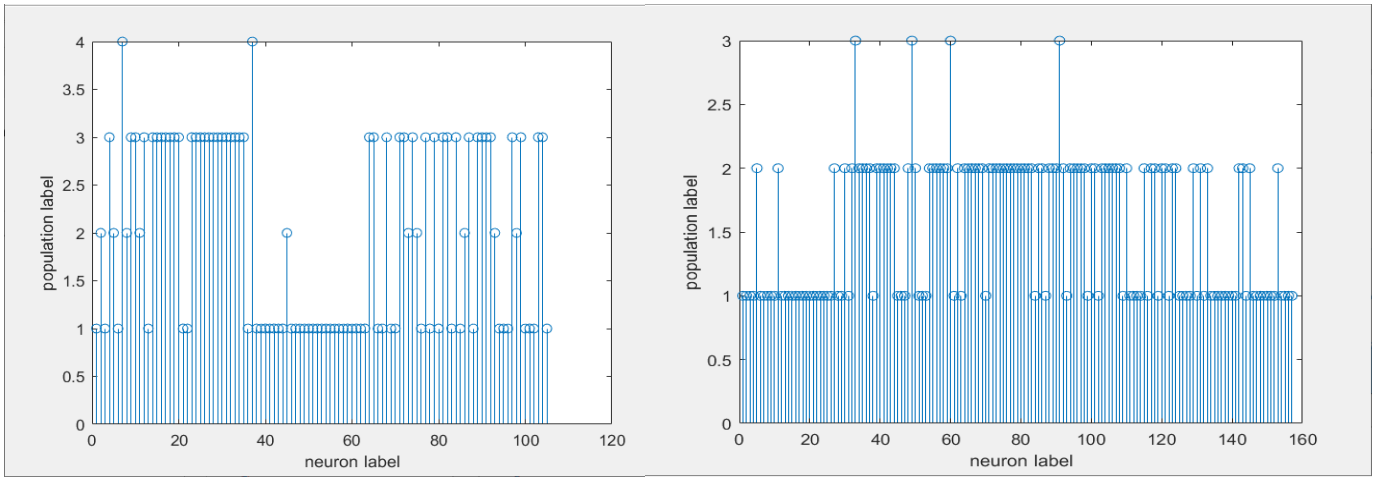

Figure 2 Stem plot for mapping the neuron label to its corresponding population label. Number of neurons: 105 (left) and 157 (right)

Now, the community detection using improved estimator will be simulated following the same flow as before and the same time series data are analyzed. The histogram of the eigenvalues with the modified theoretical curve and the outputs from the community detector algorithm are shown below.

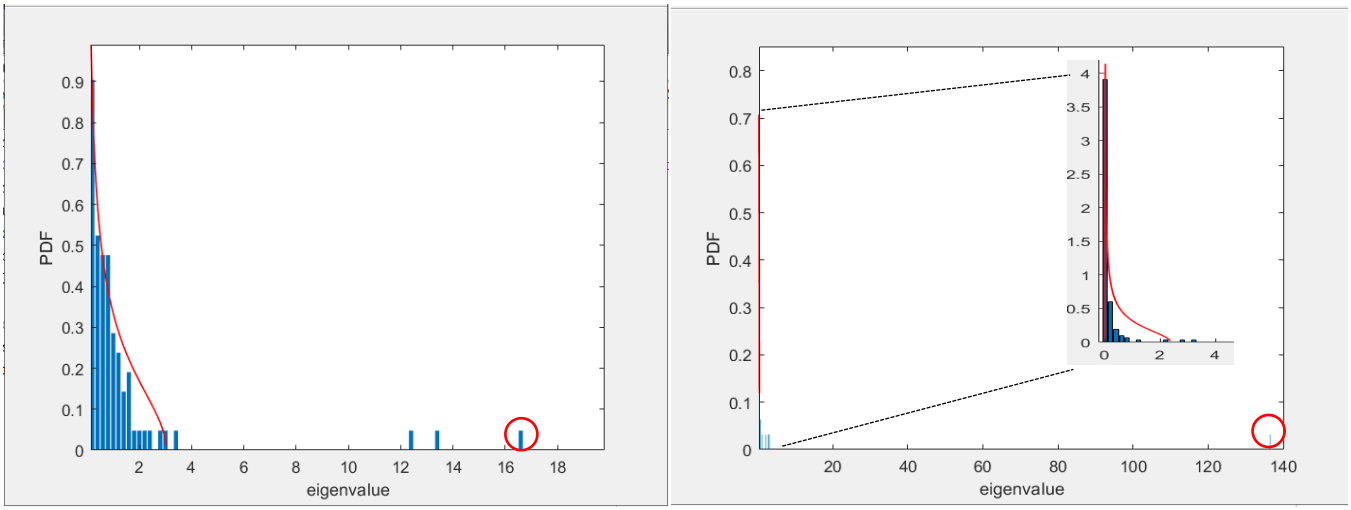

Figure 3 Histogram of eigenvalue distribution of correlation matrices from two different samples using improved estimator. Red circles are the largest eigenvalues

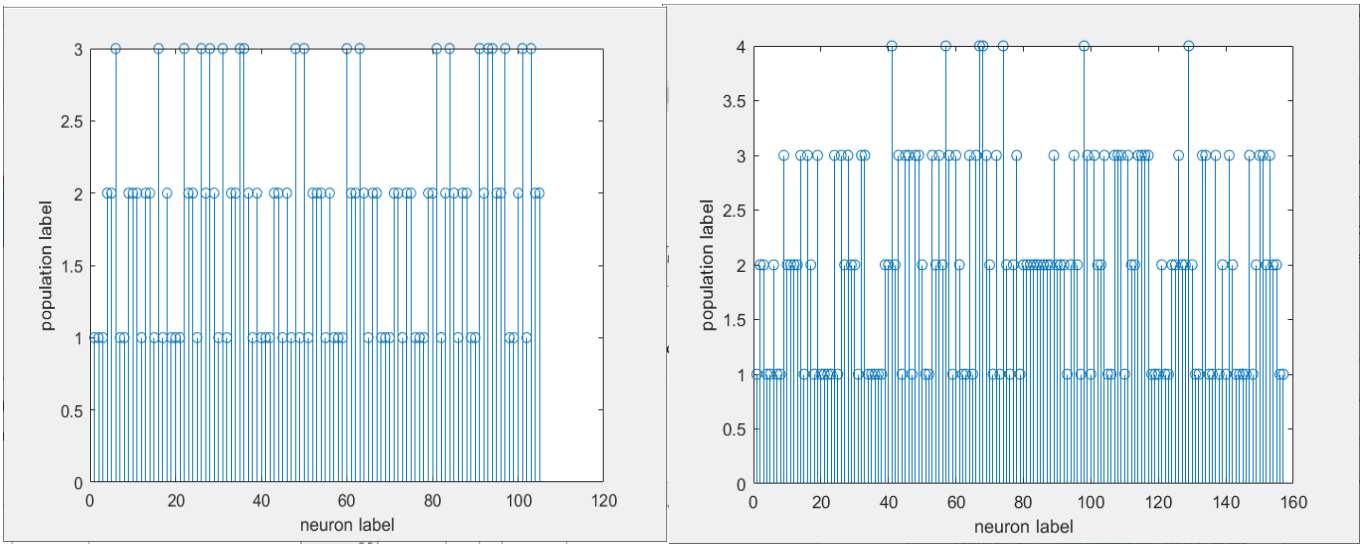

Figure 4 Stem plot for mapping the neuron label to its corresponding population label. Number of neurons: 105 (left) and 157 (right) using improved estimator

Comparing the community detection result from [1] and the improved estimator, it is clear that the results are noticeably different. Although there is no available benchmark currently because of no additional information available in the dataset provided by the author [1], the comparison will be done qualitatively based on the apparent features in the simulation results. First, the numbers of detected communities are different in both cases. The number of detected communities in the improved estimator may be higher or lower compared to the simulation results based on the original work [1]. Also, based on the histogram of eigenvalue distribution, the claim of the author that some of the eigenvalue bins are misclassified as the noise without the modification of the Marchenko-Pastur distribution is not always the case. It can be seen from the histogram from the time series data on the right that some eigenvalue bins are attracted to the bottom of the theoretical curve which leads to the reduction of the number eigenvalues deviating from Marchenko-Pastur. The underlying reason is that the estimated largest eigenvalue by the improved estimator can differ significantly from direct estimator affecting the boundary of the theoretical curve. Lastly, the distribution of the neurons in the population is also important to notice. The result based on the improved estimator seems to predict more "distributed" result than the original work. This means the number of neurons in each population tends to not differ much. From the left figure of both cases, it can be observed that in the simulation result based on the original work many neurons are labelled as population 1 while only two neurons are labelled as population 4. In contrast, the result from the improved estimator does not give such extreme difference. The notion of "distributed" population may be quantified with Shannon entropy and it can be shown that the entropy of the result from improved estimator is larger than the original. All these contrasting results emerges from the fact that the correlation matrix estimator is sensitive to the ratio value Q and should be handled carefully to ensure (M, N) consistency.

VI. Conclusion

In this report, the attempt to approach the problem of community detection in the system of many neurons has been discussed. The interaction between neurons can be described using the correlation matrix from the neural activity time series data. This correlation matrix can be decomposed into three different terms and one of the terms contains the information about the grouping of the neurons. This report also demonstrated the challenge in separating the noise part from the group part. Because of that, the Random Matrix Theory approach is used to solve this problem, in particular Marchenko-Pastur distribution and (M, N) consistent estimated population correlation matrix. In the numerical simulations, it has been shown that the community detection results with two different preprocessing of correlation matrix exhibit significant difference. Even though both processes involve filtering of noise part and global trend, the approach of using robust correlation matrix estimator seems to be more promising. This is because the robust estimator takes (M, N) consistency into the account.

VII. Statement

The idea of the estimator and its derivation is done by others. Also, the dataset of the neurons, the backbone code, as well as the original idea mainly are from others. However, the interpretations and the figures of the simulation, the supplementary code for the simulation and the linkage between ideas are done by the author of this paper.

VIII. References

[1] Almog A, Buijink MR, Roethler O, Michel S, Meijer JH, Rohling JHT, et al. (2019) Uncovering functional signature in neural systems via random matrix theory. PLoS Comput Biol 15(5): e1006934. https://doi.org/10.1371/journal.pcbi.1006934

[2] Kim, Dong-Hee & Jeong, Hawoong. (2005). Systematic analysis of group identification in stock markets. Physical review. E, Statistical, nonlinear, and soft matter physics. 72. 046133. 10.1103/PhysRevE.72.046133.

[3] MacMahon, Mel, and Diego Garlaschelli. “Community Detection for Correlation Matrices.” Physical Review X 5.2 (2015): n. pag. Crossref. Web.

[4] A.M. Sengupta and P.P. Mitra Distribution of Singular Values for Some Random Matrices, cond-mat/9709283

[5] Laloux, Laurent et al. “Noise Dressing of Financial Correlation Matrices.” Physical Review Letters 83.7 (1999): 1467–1470. Crossref. Web.

[6] X. Mestre, "Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates," in IEEE Transactions on Information Theory, vol. 54, no. 11, pp. 5113-5129, Nov. 2008.