**Boston University**

**Electrical & Computer Engineering**

**EC463 Senior Design Project**

Prototype Testing Report

**Neuron Spike Identification with Machine Learning**



Chen Yang, PhD

Yang Research Laboratory

cheyang@bu.edu

by

Team 2

Spike Sorters

Team Members

Victoria Carlsten [carlsten@bu.edu](mailto:email1@bu.edu)

Hao Chen [ha0chen@bu.edu](mailto:email1@bu.edu)

Claire Cropper [ccropper@bu.edu](mailto:email1@bu.edu)

Shi Gu [bengushi@bu.edu](mailto:email1@bu.edu)

**Required Materials:**

Hardware:

* Personal Laptop

Software:

* Python, 1 scripts
* 1 data files
  + *spikesortingVTJason.py*
    - Load recording data
    - Detect spikes
    - Delete unnecessary data
    - Implement Regular K-means and PCA analysis
    - Implement Optimized K-TOPS algorithm and PCA visualization
* Python, 1 script
* 2 data files
  + *spikesortingVTJason\_Data.py*
    - Load recording data files
    - Detect spikes
    - Delete unnecessary data
    - Implement Regular K-means and PCA analysis

**Setup:**

The setup only consists of one part: using our personal laptops to generate graphs via one Python script. Firstly, running the spikesortingVTJason.py will implement K-means and PCA to generate clustering results. The user will need to input 2 cluster numbers to select the interested cluster. And a new clustering result with a different algorithm will be generated to compare with the former result.

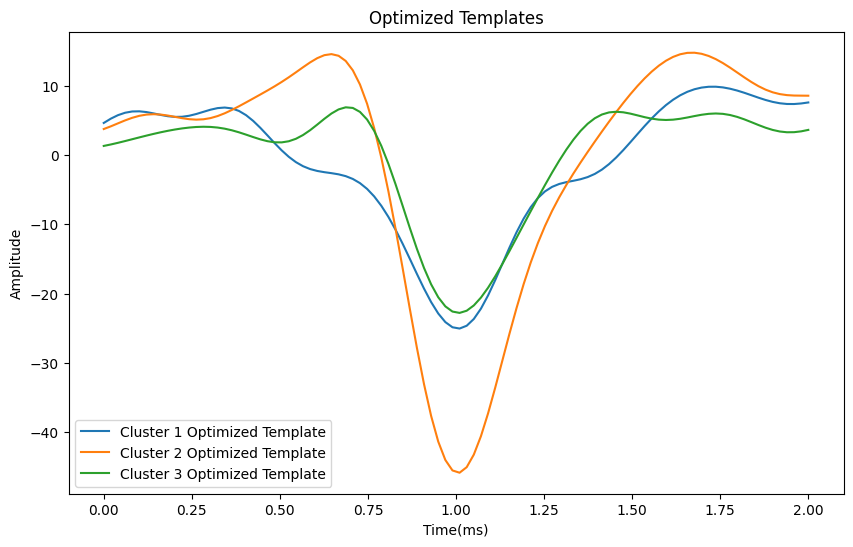
**Pre-testing Setup Procedure:**

spikesortingVTJason.py file:

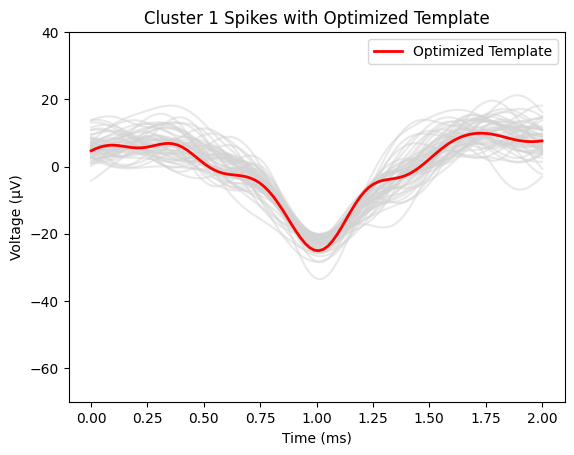
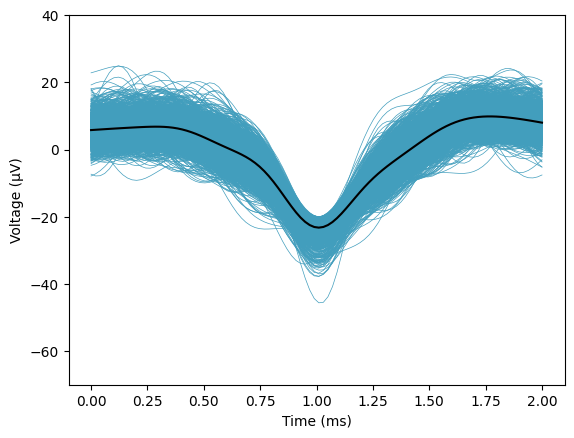
1. Make sure user is in the correct working directory
2. Make sure ‘10 min recording1.mat’ is in the working directory
3. Run all the lines of the python script, *spikesortingVTJason.py* and make sure no errors happen

spikesortingVTJason\_Data.py file:

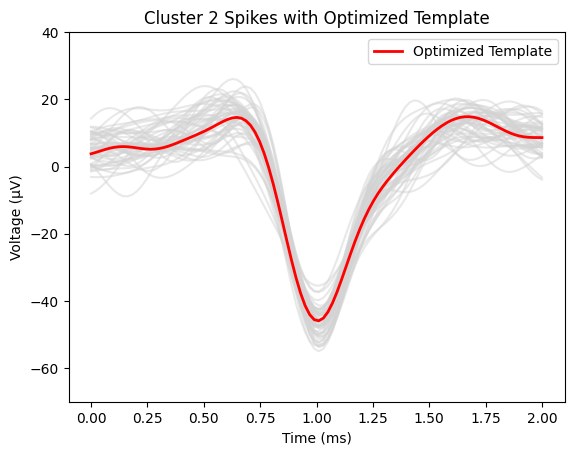
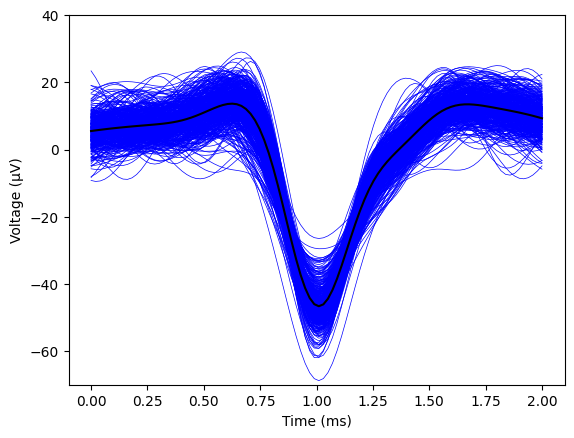
1. Make sure user is in the correct working directory
2. Make sure ‘Data\_Subject\_08\_Session\_01.h5’ is in the working directory
3. Run modified lines of the python script, *spikesortingVTJason.py*



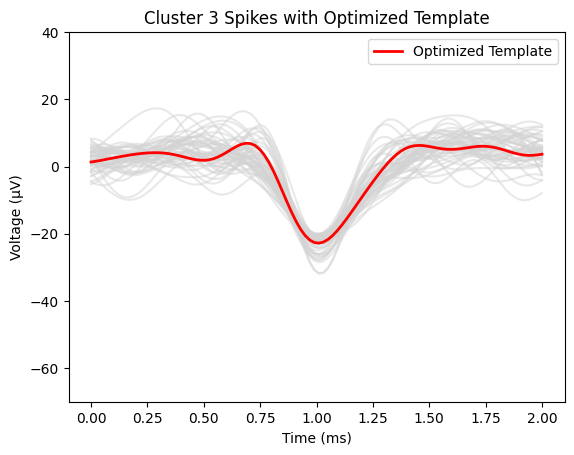
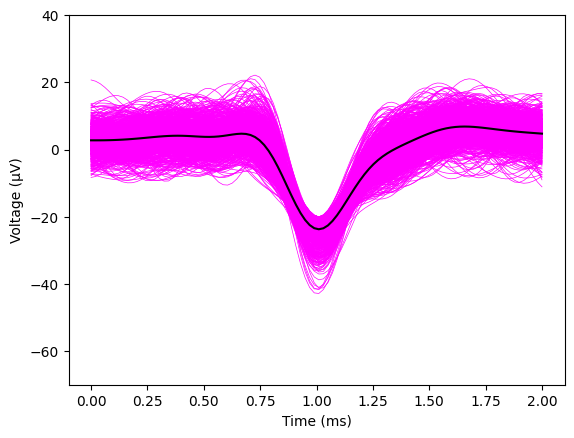
*Figure 1. Optimized Templates for 3 clusters generated by initial k-means*

******

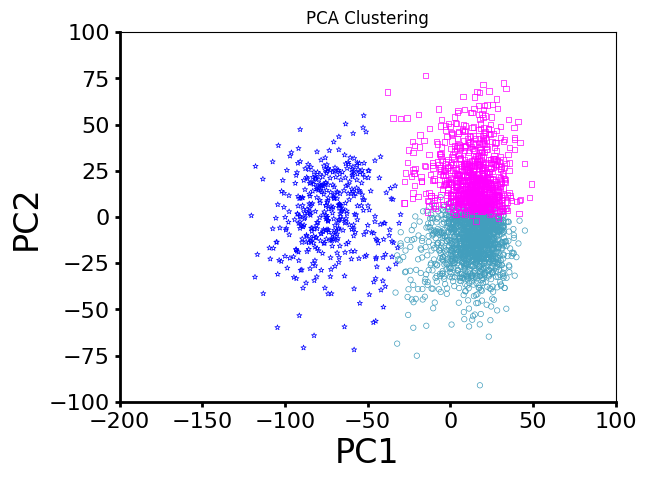
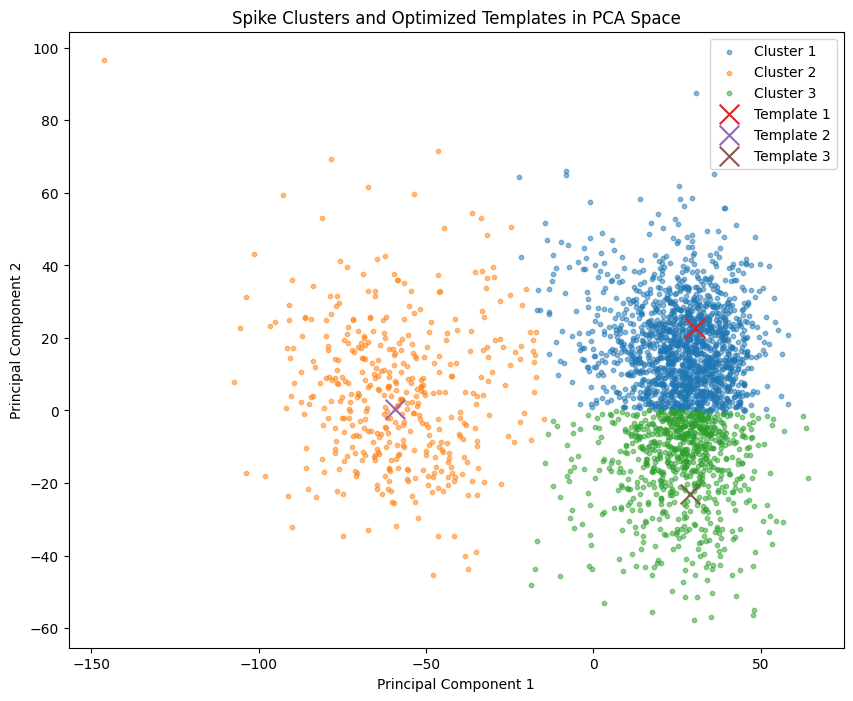
*Figure 2. Optimized Template(left) and mean of spikes(right) with all spikes in cluster 1*

**

*Figure 3. Optimized Template(left) and mean of spikes(right) with all spikes in cluster 2*

**

*Figure 4. Optimized Template(left) and mean of spikes(right) with all spikes in cluster 3*



*Figure 5. Clustering using K-means(right) and K-TOPS(left) in PCA space*

**Testing Procedure:**

1. Run spikesortingVTJason.py
2. The computer will load the recording data from the ‘10 min recording1.mat’ and identify spikes
3. Input two clusters that users are interested in to see detailed graphs
4. The computer will then generate clustering results using a different clustering approach
5. Run spikesortingVTJason.py
6. The computer will load the recording data from the ‘Data\_Subject\_08\_Session\_01.h5’ and identify spikes

**Measurable Criteria:**

1. spikesortingVTJason.py (lines 1 - 118) should correctly load the recording data from the ‘10 min recording1.mat’, which is a mat file of pre-recorded 10 minutes of Electrophysiological Data, and then select the part of interest.
2. Spikes should be detected according to the threshold value and output a plot similar to Figure 2.
3. First step of PCA analysis of the spike array should output a plot similar to Figure 3.
4. There should be three(desired by us) clusters clearly classified in the PCA graph.
5. Two clusters can be selected in a separate graph by input.
6. Three clusters using the K-TOPS algorithm should be shown in a different graph.
7. L-ratio should be around or lower than 0.02
8. spikesortingVTJason\_Data.py should correctly load the recording data from the ‘Data\_Subject\_08\_Session\_01.h5’, which is a h5 file of pre-recorded 29 seconds of Electrophysiological Data, and then select the part of interest.
9. Spikes should be detected according to the threshold value and output a plot similar to Figure 2.

**Conclusion**

Our project utilized two distinct datasets for the evaluation and validation of our algorithms for neuron spike identification. The first dataset, referred to in our testing procedure as '10 min recording1.mat', contains 10 minutes of pre-recorded electrophysiological data. This dataset served as the primary source for our initial tests, allowing us to apply our algorithms to a significant amount of data to assess their effectiveness comprehensively.

The second dataset, named 'Data\_Subject\_08\_Session\_01.h5', encompasses 29 seconds of electrophysiological data. This dataset allowed us to validate our findings from the first dataset and further refine our algorithms. Both datasets were critical in our process of loading the recording data, identifying spikes, and performing our advanced analyses, including PCA and the optimized K-TOPS algorithm.

Through meticulous testing and analysis, we detected spikes according to predefined threshold values, successfully outputting plots that visually represented the data and our clustering results. Our PCA analysis further enabled us to classify these spikes into distinct clusters. This step was crucial for visualizing the efficiency of our algorithms in segregating neuron spikes into separate groups based on their characteristics.

Our project achieved a significant milestone by demonstrating that three clusters could be clearly identified in the PCA graph, fulfilling one of our predefined measurable criteria. Moreover, our ability to select two clusters in a separate graph by input provided an interactive element to our analysis, allowing for a deeper exploration of the data.

The pinnacle of our testing was the application of the K-TOPS algorithm, which showcased its ability to produce three distinct clusters, validating the algorithm's effectiveness in handling complex electrophysiological data. The precision of our clustering was underscored by achieving an L-ratio of around or lower than 0.02, a key success criterion that indicates the high accuracy of our spike sorting methodology.

Incorporating an H5 file into our analytical code presented an interesting challenge, particularly because our primary goal was to visualize specific aspects of the dataset: the raw signal, filtered spikes, and filtered Local Field Potentials (LFP). The choice to focus on these elements was driven by the nature of the dataset and the compatibility issues that often arise when dealing with specialized data formats in neuroscience.

These datasets and the subsequent analysis underscore the potential of our machine learning algorithms in advancing neuron spike identification. The comprehensive data analysis not only validated our approach but also set a foundation for future research in this field, aiming at further enhancements and applications of machine learning in neurological studies.