Classification and Normalization for Ganglion Cells

Christopher Gong

January 2020

1 Introduction

This project, a UC Berkeley Research Apprenticeship under Amanda McLaughlin and Teresa Puthussery in Puthussery & Taylor Labs, performs preprocessing normalization and classification to group ganglion cells from the human eye. The data is cells in response to the light stimulus. All of the code to reproduce these results are in the GitHub Repository.

2 Normalization

Given the DAPI data, the data is 0-1 normalized and cutoff with a value of 0.01 (1% of the mass on the left and 1% of the mass on the right). Since low DAPI values are possibly erronous, all rows with DAPI less than 0.2 are removed. All of the other data is also normalized and z-normalized, and attached as more columns to the dataset.

3 PCA

Another prepossessing techniques we applied was principle component analysis, in attempt to emphasize the axis with the most variance. In some datasets, using as few as 3 components accounted for over 60% of the data. Other options we explored were using components that explained the variance of a specific percentage (ex: 95%, 99%) of the data.

Code: In addition to writting the Normalization procedure above, two additional functions were written. The normHists function plots three histograms of the original data, normalized data, and z-normalized data. The meanSTD function returns a mean and standard deviation for each column.

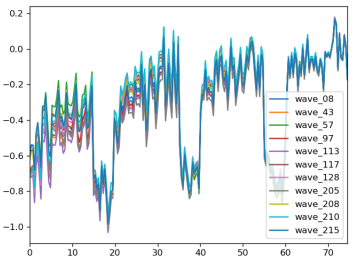
4 Classification

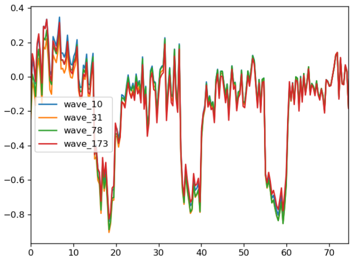
After the data has been normalized, we attempt to classify the data into separate categories. To classify, we used both k-means classifiers (k=expected number of classes) and Gaussian Mixture models. These classifications were attempted on PCA and Non-PCA data. Since both k-means and GMM need a specified number of components, we also attempted to use the DBSCAN algorithm.

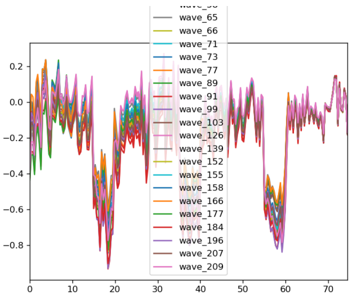
Code: This code performs the classifications above, and has the option to save the labels to with colors to a csv file.

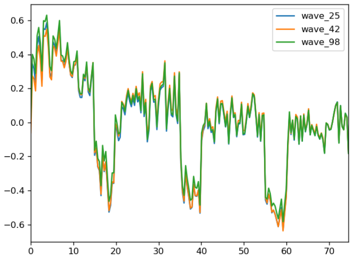
5 Results

Below are some of the plots produced by our classification techniques. DBSCAN:

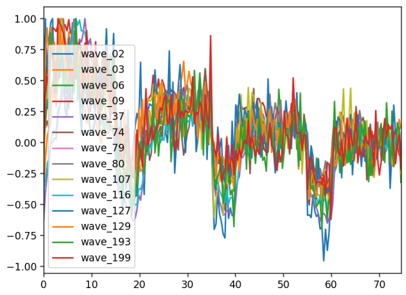


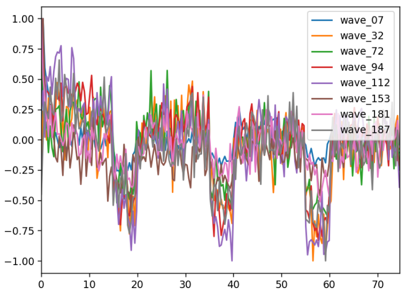


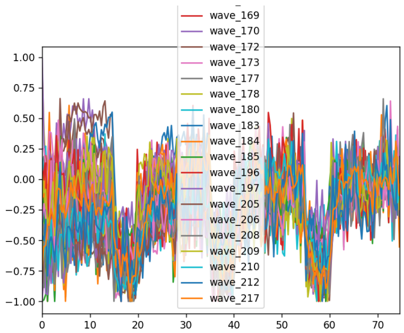


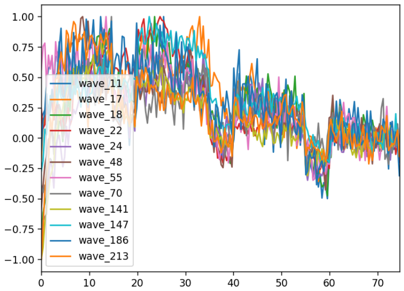


CLUSTER:





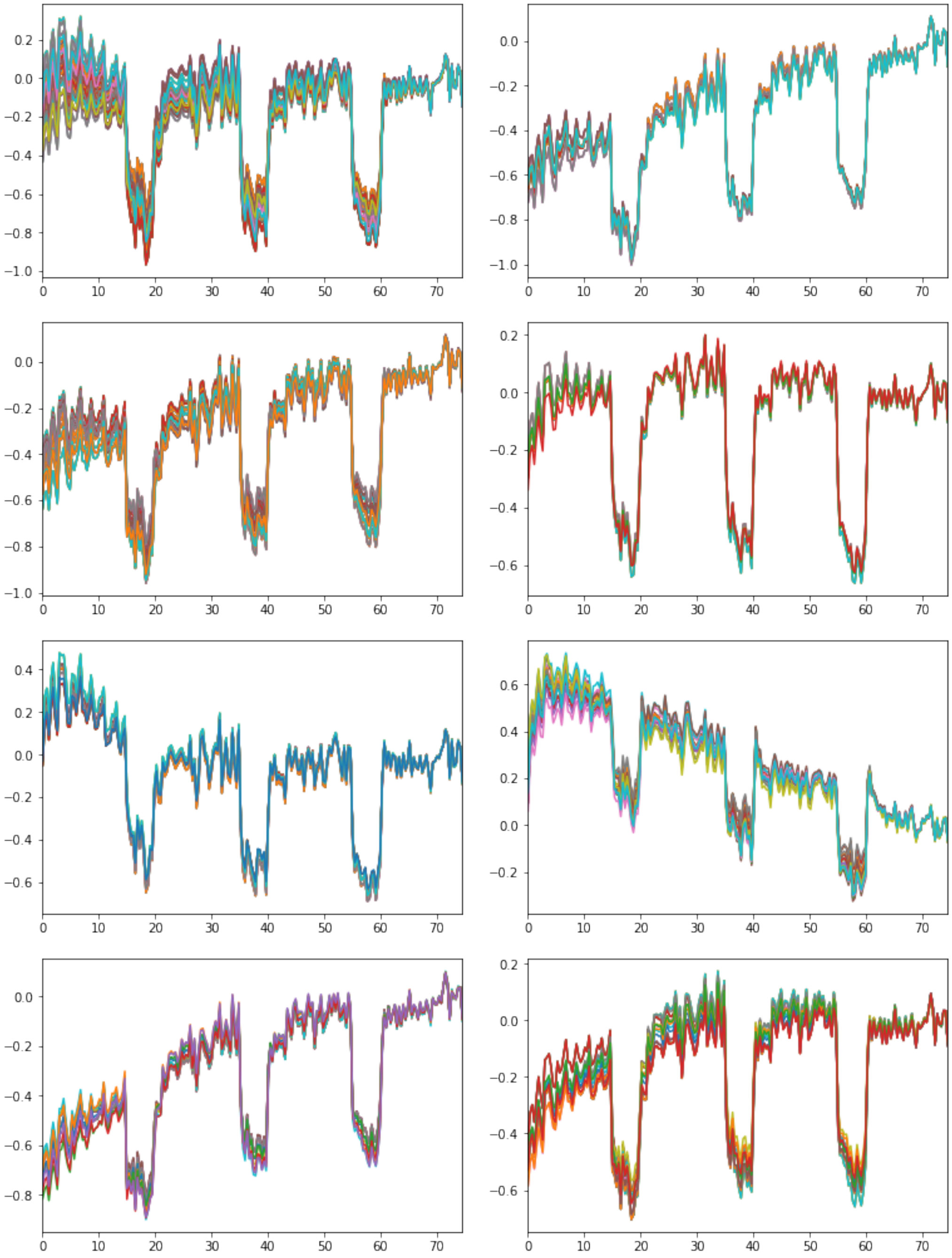




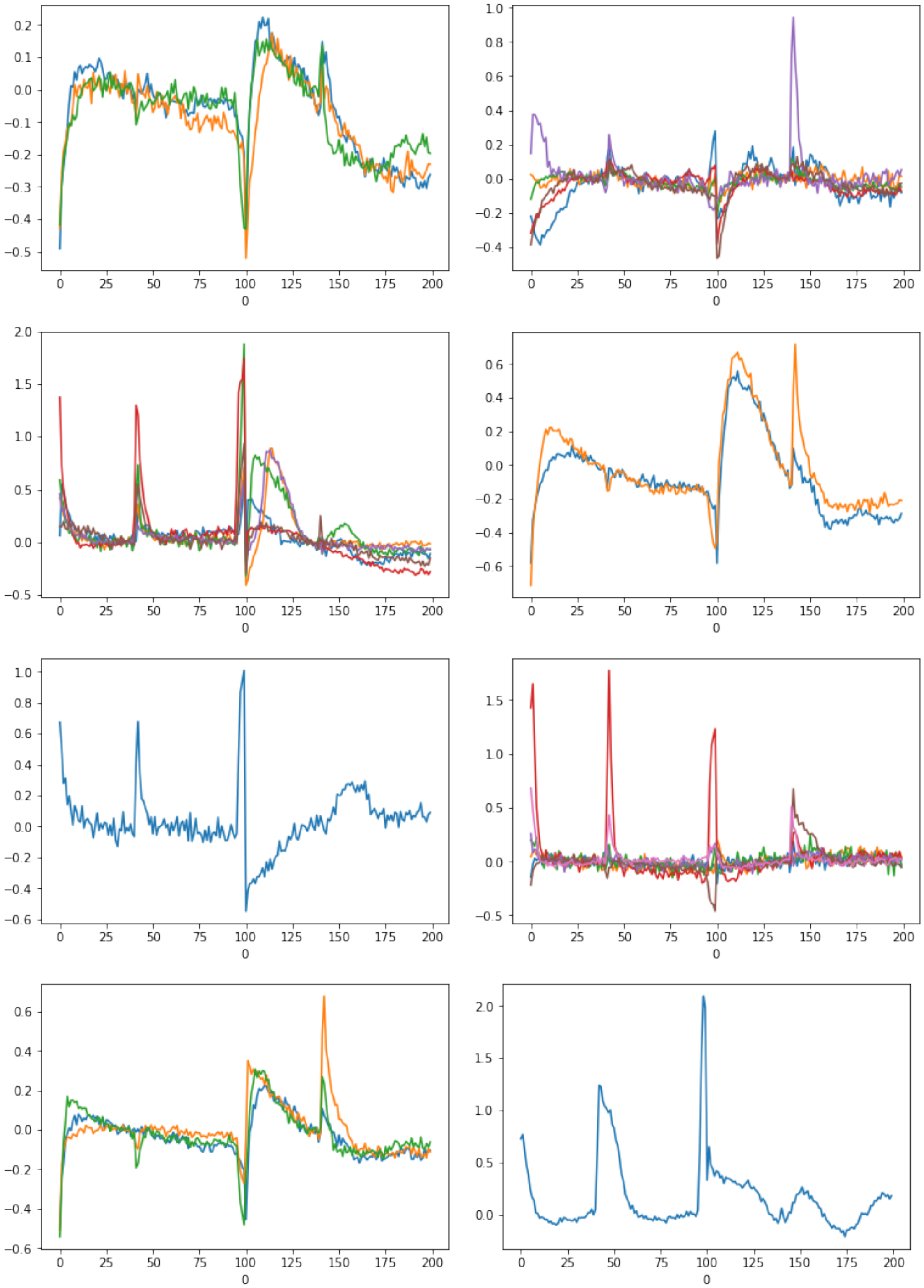
6 Conclusion

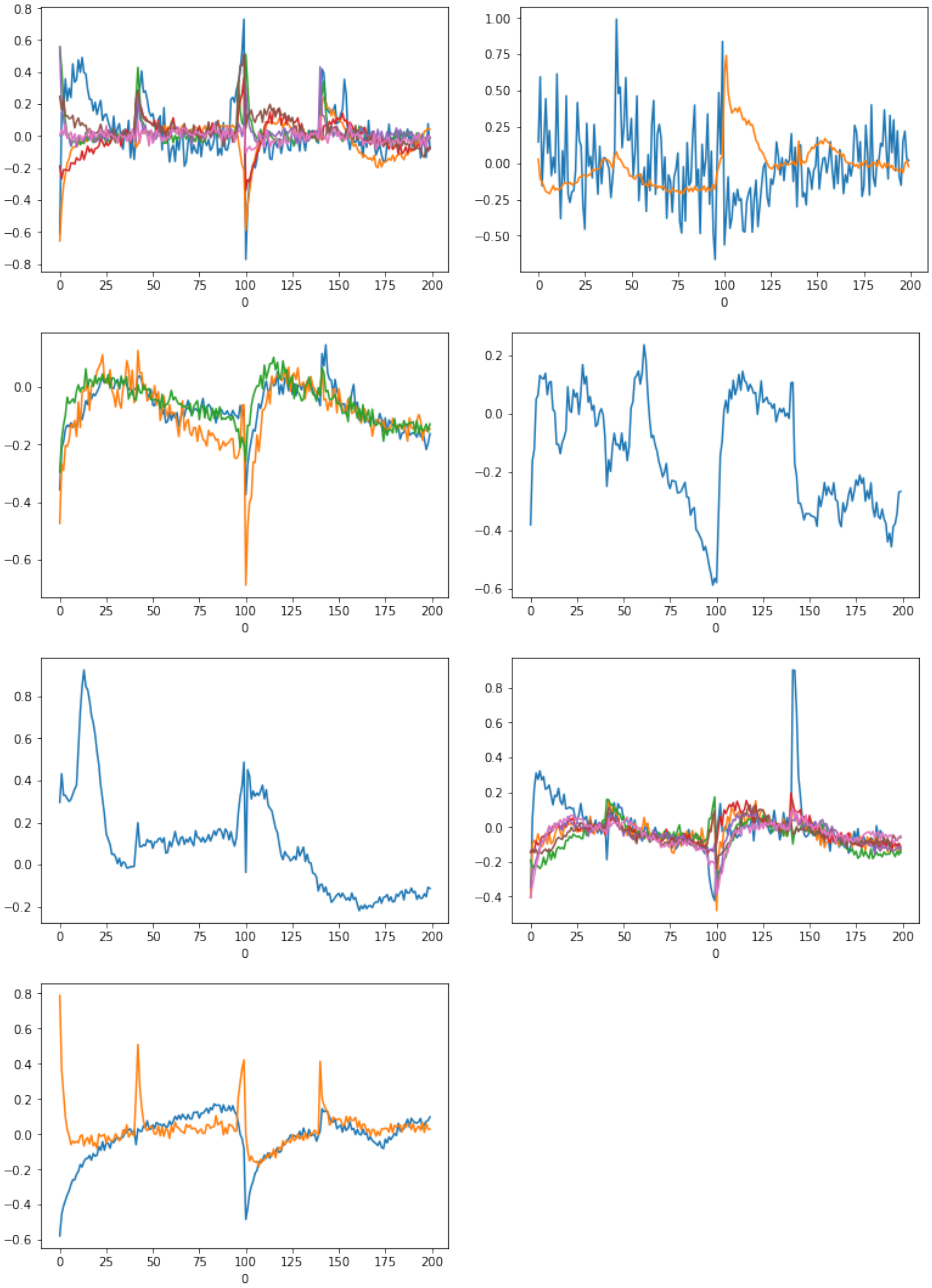
Our classification of the data set produced the best results with DBSCAN, where the clustering groups seem very closely correlated.

DBSCAN with PCA



K-means with PCA




0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
25
50
75
100
125
150
175
200
0
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
25
50
75
100
125
150
175
200
0
0.0
−0.2
−0.4
−0.6
0
25
50
75
100
125
150
175
200
0
0.2
0.0
−0.2
−0.4
−0.6
0
25
50
75
100
125
150
175
200
0
0.8
0.6
0.4
0.2
0.0
−0.2
0
25
50
75
100
125
150
175
200
0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100
125
150
175
200
0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
25
50
75
100
125
150
175
200
0