Genomics and High Dimensional Data

Analysis 2 for MIT 6.419x Data Analysis: Modeling and Applications

Evan Rushton

March 29, 2021

1 Problem 2: Larger unlabeled subset

1.1 Part 1: Visualization

1. (3 points) Provide at least one visualization which clearly shows the existence of the three main brain cell types described by the scientist, and explain how it shows this. Your visualization should support the idea that cells from a different group (for example, excitatory vs inhibitory) can differ greatly.

To determine the number of clusters in the data in an unsupervised fashion, principal component analysis (PCA) was performed. This algorithm projects the data onto fewer dimensions in a manner to preserve as much of the variation as possible. Figure 1 shows PCA run on the raw data (red) and the log-transformed data (blue) for N components from N=1 to N=200 and shows that PCA is able to explain the majority of the variance for the log-transformed data. 85% of the variance is explained for N=12 principal components and this value of N with the log-transformed data was used whenever PCA was performed for the rest of this analysis.

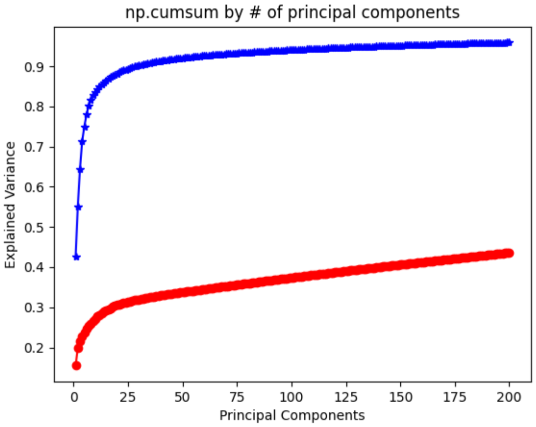


Figure 1. % of the variance explained by number of principal components. Raw data (red), log-transformed (blue).

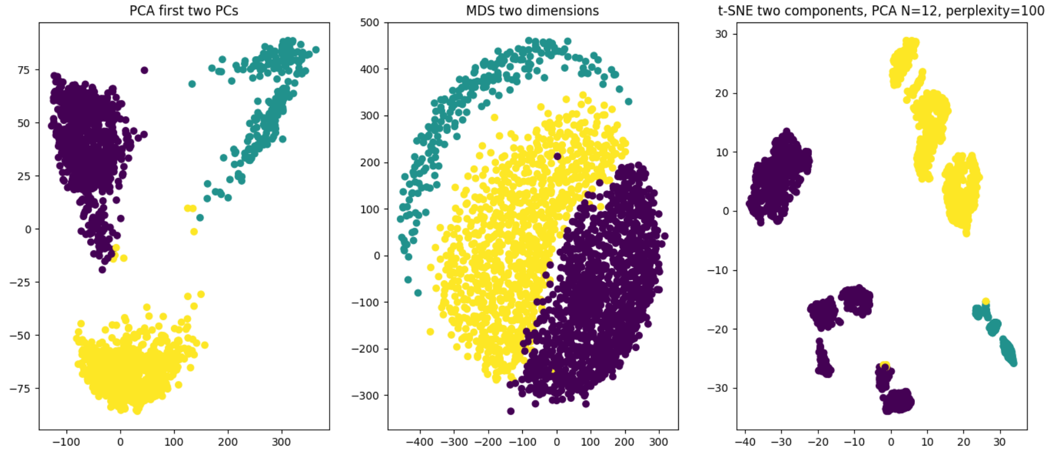
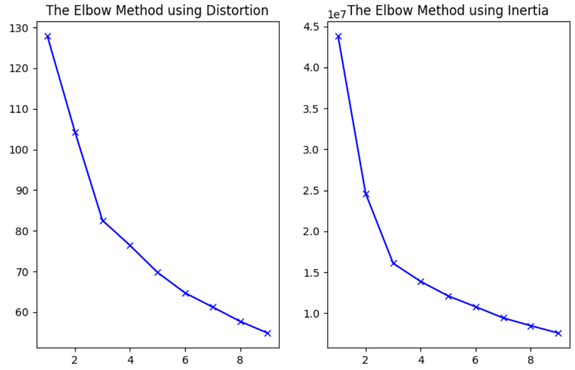


Figure 2. PCA(N=12), MDS(n=2), and t-SNE(PCA12, perp=100)

In Figure 2 the data has been visualized using a variety of dimension reduction techniques. The data was plotted along the first two principal compenents in the leftmost plot of Figture 2. Notice that there are three distinct clusters. The middle plot of Figure 2 is the multidimensional scaling (MDS) of the data into 2-dimensional space. This algorithm attempts to preserve distances between points and also suggests three clusters (albeit that the yellow and purple points could be considered a single cluster from this visualization). The rightmost plot of Figure 2 is a t-distributed stochastic neighbor embedding (t-SNE) of the PCA-transformed data. This algorithm attempts to keep similar points close together and dissimilar points far apart. With a perplexity=100 it seperates the same three clusters as the other two visualizations.

The colors are assigned by K-means with the number of clusters set to k=3. This value was chosen by looking at the elbow plots of distortion and inertia for k=1 to k=10 clusters as shown in Figure 3.

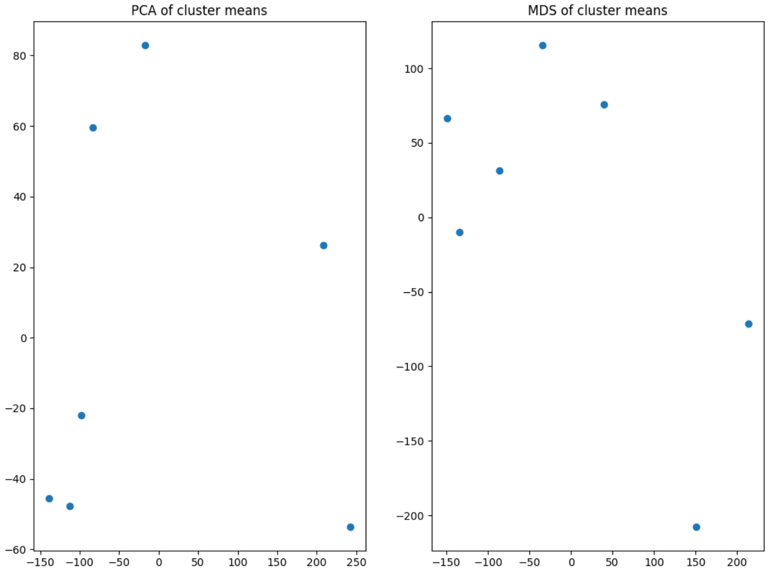

The Elbow Method using Distortion

The Elbow Method using Inertia

Figure 3. K-means distortion and inertia plots versus number of clusters.

2. (4 points) Provide at least one visualization which supports the claim that within each of the three types, there are numerous possible sub-types for a cell. In your visualization, highlight which of the three main types these sub-types belong to. Again, explain how your visualization supports the claim.

From the previous analysis, we have identified three clusters that may indicate three cell types: purple, yellow and green. From the t-SNE plot and the elbow plots it seems reasonable that there may be 7 or 8 clusters. To see if there are sub-types within each of these three main cell types, a PCA and MDS of the cluster means was performed on the resulting K-means data with k=7 clusters. These plots in Figure 4 suggest that the yellow cell-type has three subtypes (third quadrant of the PCA plot), the purple cell-type has two subtypes (second quadrant of the PCA plot), and the green cell-type has two subtypes($X > 200$ on the PCA plot).


PCA of cluster means

MDS of cluster means

Figure 3. K-means distortion and inertia plots versus number of clusters.

1.2 Part 2: Unsupervised Feature Selection

(4 points) Using your clustering method(s) of choice, find a suitable clustering for the cells. Support your choice of clustering with appropriate visualizations and/or numerical findings. Be sure to briefly explain how you chose the number of clusters.

As demonstrated in Figure 2 above in the visualization section, the most suitable clustering of the cells looks to be 3 main cell-types. This is supported by PCA, MDS, t-SNE and K-means. The elbow plots suggested 3 clusters and since the coloring displayed by K-means looks to seperate out these cell types quite well, I will use K-means with k=3 to label the data for logistic regression.

(6 points) We will now treat your cluster assignments as labels for supervised learning. Fit a logistic regression model to the original data (not principal components), with your clustering as the target labels. Since the data is high-dimensional, make sure to regularize your model using your choice of L1, L2, or elastic net, and separate the data into training and validation or use cross-validation to select your model. Report your choice of regularization parameter and validation performance.

I am flying to Atlanta and will attempt to complete more of this report before the deadline. I am very sad to not complete in time, but alas, life calls.