Exam 2

Zuojun Gong

Friday, April 10, 2015

```
## Loading required package: boot
## Loading required package: MASS
## Loading required package: segmented
## mixtools package, version 1.0.2, Released May 14 2014
## This package is based upon work supported by the National Science Foundation under Grant No. SES-05187
72.
```

Introduction

Brain is the central command of a mammal's body and everything about their lives evolves around it. However, discovering how the brain operates has always been a mysterious and complicated subject because of mammals' brains' complexity. In this study the researchers aim to find out how nerve cells in a monkey's brain communicate and process information through electrical impulse transmitting. In particular, we want to discover how some of the neurons in the motor region of its brain reacts or controls hand movements. We say these neurons has directional tuning. In such neurons, we observe that when the monkey moves its hand, electrical impulses was detected from the neurons. Each of these neurons has a preferred direction, which reacts most with movement matching such direction, and it reacts the least with a movement at the opposite direction. In this report, we aim to discover the directional tuning neurons from the motor region of the monkey brain through statistical methods.

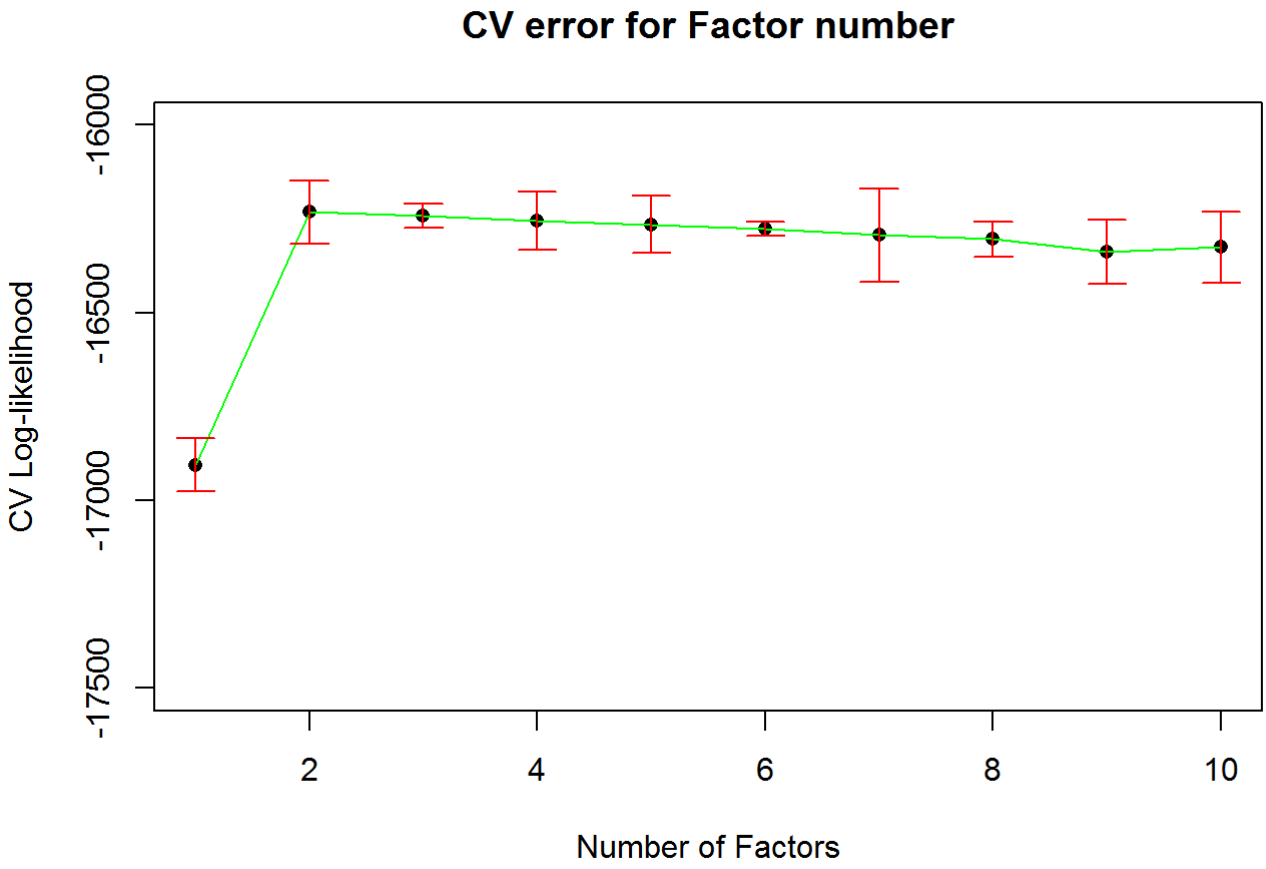
Exploratory Data Analysis

In order to preform factor analysis we must have a dataset where the mean and variance for 0 and 1. We first check our model variable distribution through summary report and we discovered that the all variables does not have a mean of 0. Therefore, we will scale the dataset and prepare for model building.

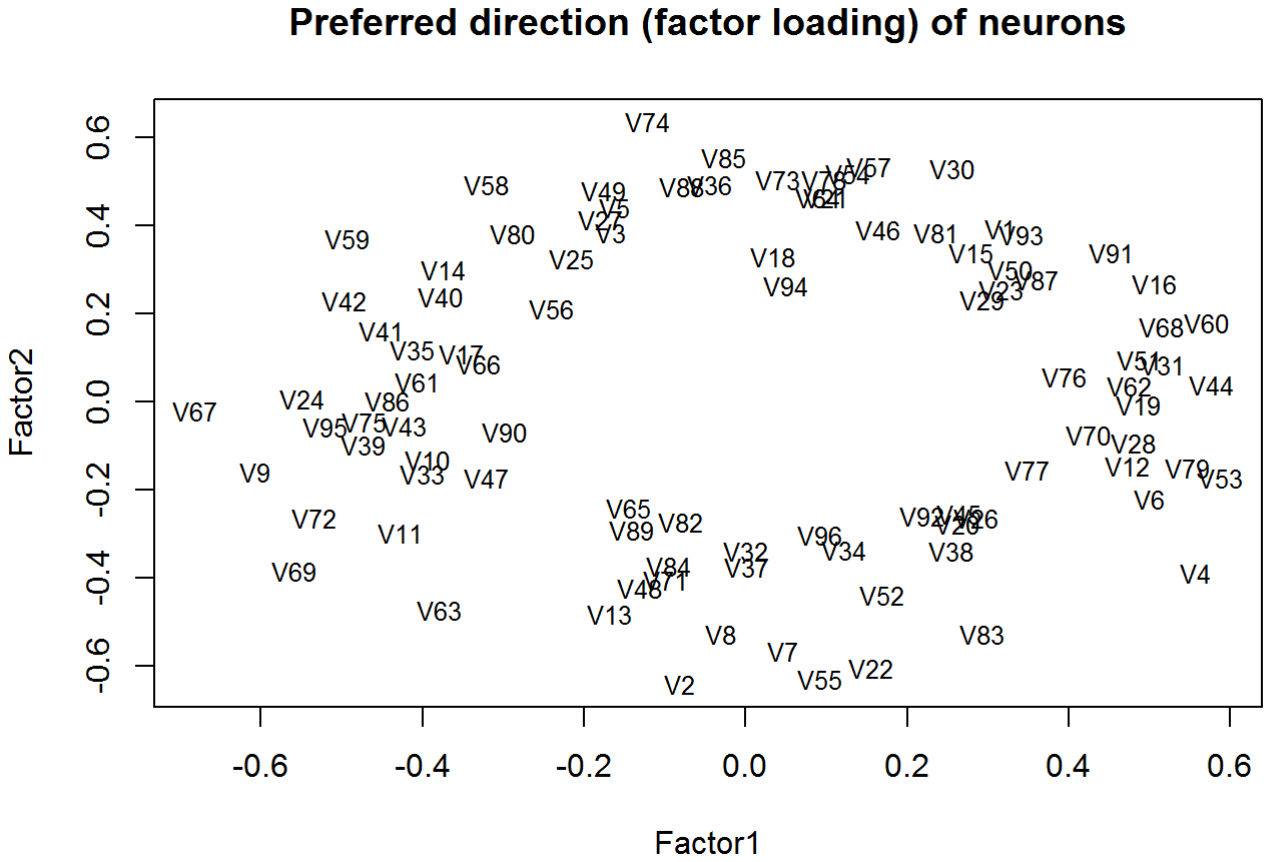
Model Building: Factor Models

We know from factor models, we explain the entire dataset X in F, w, and epsilon. Where F is the matrix of factor scores, w is the matrix of factor loadings, and epsilon being the matrix of noise. We know that the matrix of factor loadings describes how the different variables in our dataset correlates with the different underlying factors. The matrix of factor scores are the scores of each case on each factor. In other words, in what way does each data entry relates to each factor. We think that the directional tuning neurons' behaviors in accordance to hand movement, measured by electrical spikes, is similar to a factor model. In our research, we hypothesis the neuron reacts to motions with spikes, and the average number of spikes over a short interval is a+b*v. The vector b is a vector of preferred direction, which responds with the highest level of spikes to the "preferred direction" and the lowest level of spikes to the opposite direction. The vector v is a direction vector, denotes the direction of monkey's hand movement. The constant a denotes the potential normal levels of spikes in a given neuron. We believe that this is related to a factor model. We can see the preferred direction as factor loadings and the direction vector as factor scores. With recording of neuron spikes due to movement, we can compute the estimated direction vector from our factor model.

However, it is not obvious what the optimal number of factor of our factor model is, therefore we preform cross-validation on the log likelihood on factor models to determine our optimal number of factors. By comparing the log-likelihood of different factor models, we can discover the best model and the optimal number of factors.

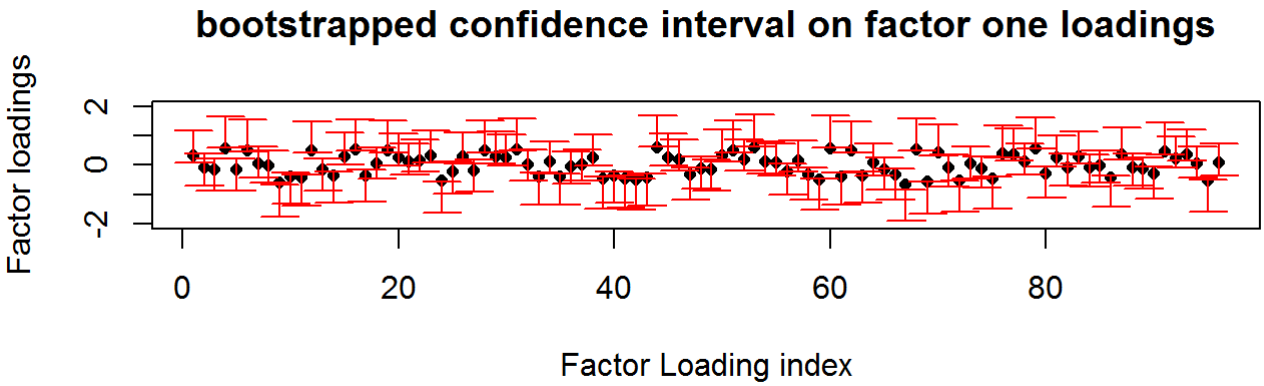


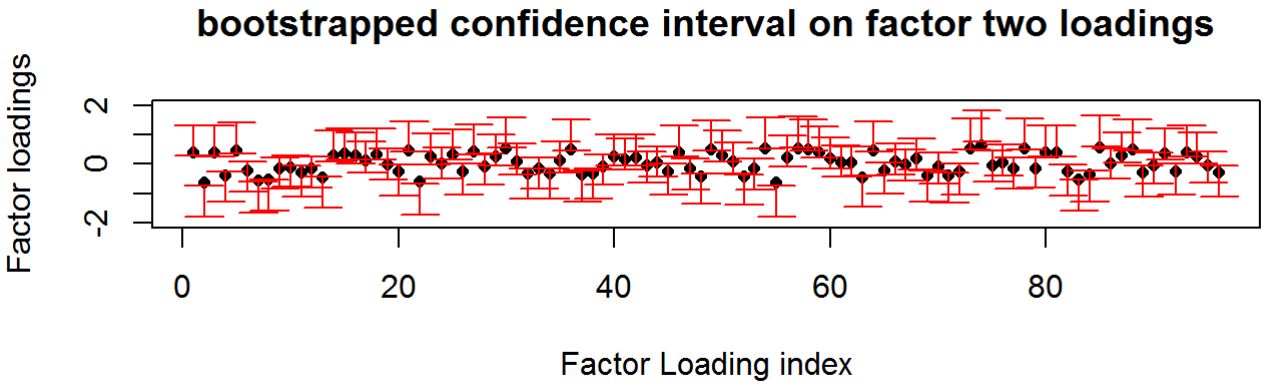
From our cross-validaiton result for log-likelihood, we determine that it is best for us to choose the two factor model since it has the highest log-likelihood value relative to all other models, indicating that this model have a higher probability of making the right predictions based on our observations. We then construct a factor model with two factors, and examine the preferred direction, which is the factor loadings, of our dataset.



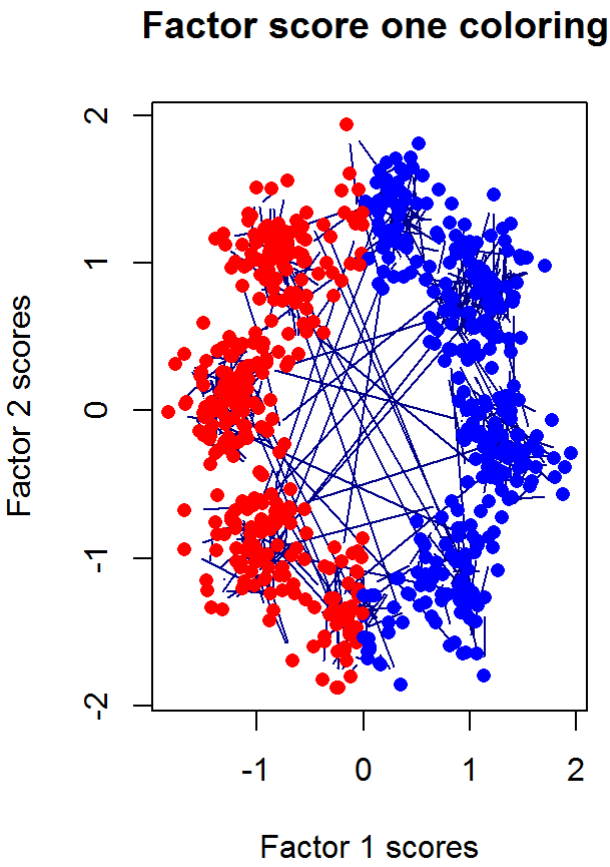
We can observe that the preferred directiorns for all 96 neurons spread out on the two-dimensional plane of factor loading 1 and factor loading 2. Each neuron's preferred direction is denoted by their neuron index number, and their coordinates on the two-dimensional plane represents their two factor loadings values. We can see that all "directions" are covered by different neuron's preferred directions, and it is coherent with our interpertation of prefered direction for neurons.

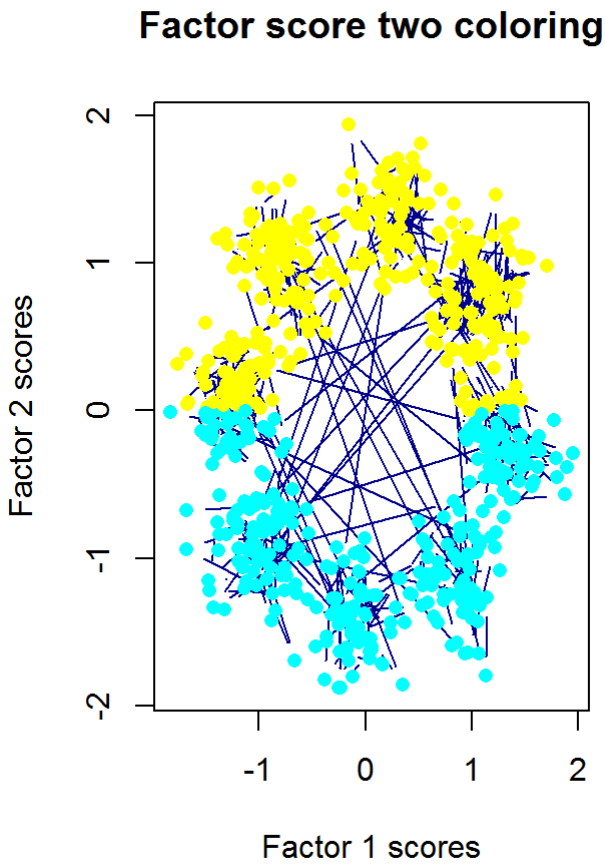
We now look at the bootstrapped 90% confidence intervals for factor loadings.





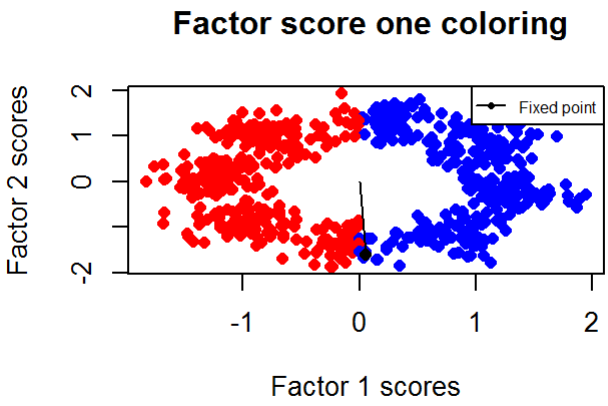
We want to visualize our factor score on a two dimensional plane with x-axis being Factor 1 and y-axis being Factor 2.

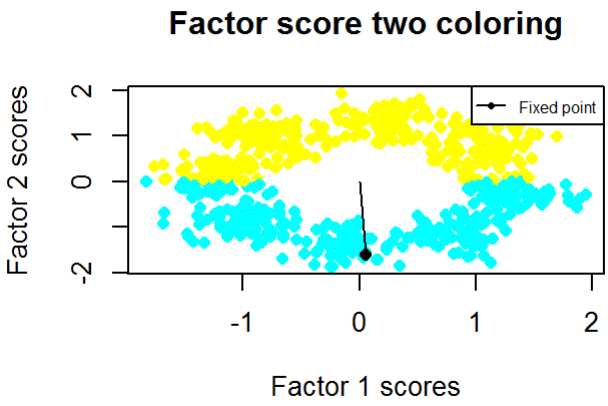


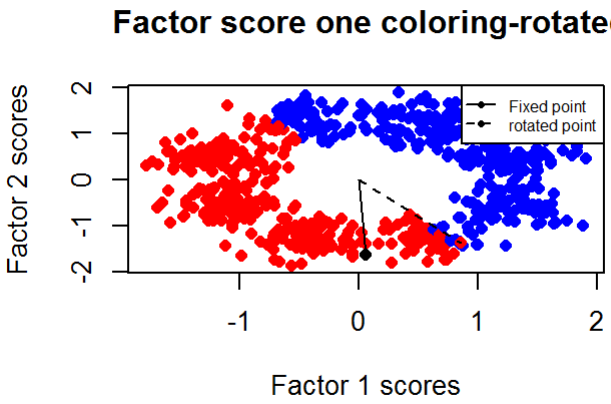


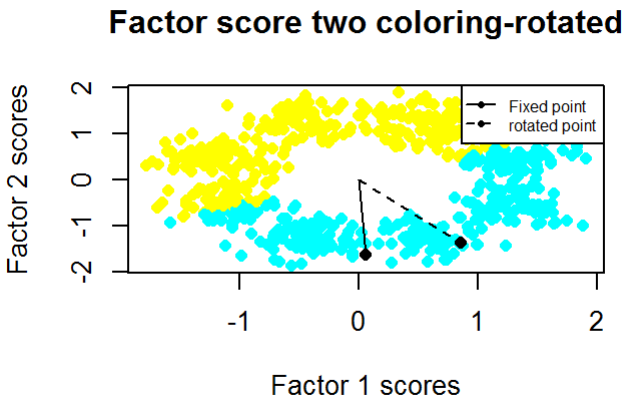
We used different colors in the graph to represent signs of values in factor one and factor two. We can observe that most of the factor scores, which is the direction vectors, are clustered in 8 clusters. This is consistent with the fact that the monkey is required to move in 8 different directions. However, we cannot determine what the actual direction is from our data or graph. We believe that our visualized factor score (movement velocity) is not corresponding to a fixed direction. Nonetheless, we believe that the relationships and differences between each cluster of factor score is relative. In other words, we do not need to find out the specific directions of factor scores to continue our analysis, and our conclusions is valid even without knowing the specific directions.

Therefore, we want to see what influence rotation will have on our dataset. In particular, we want to find out that if we use a rotated velocity vector v, what influence it would have on our preferred direction b. In this case, we rotate the velocity vector 30 degrees counter-clockwise, we want to see its effect on the factor loadings.



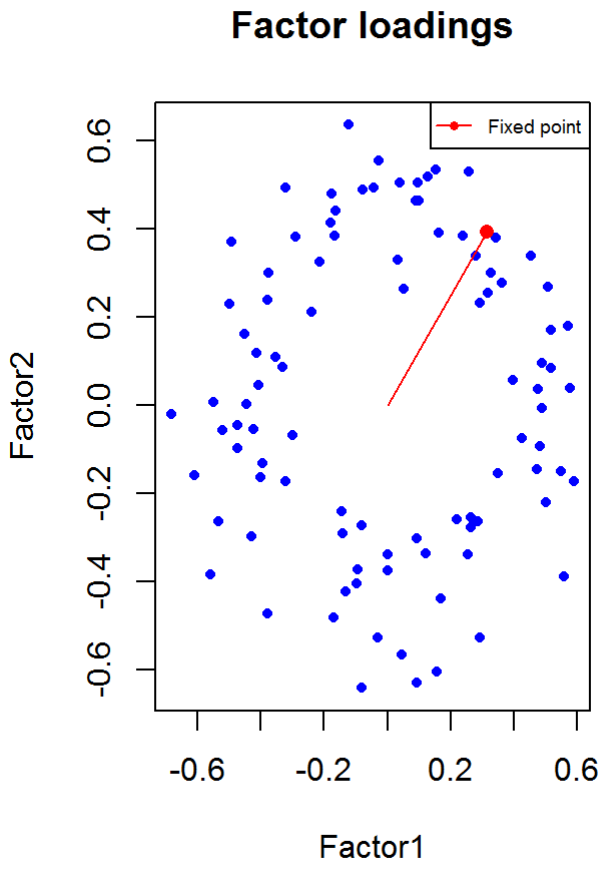


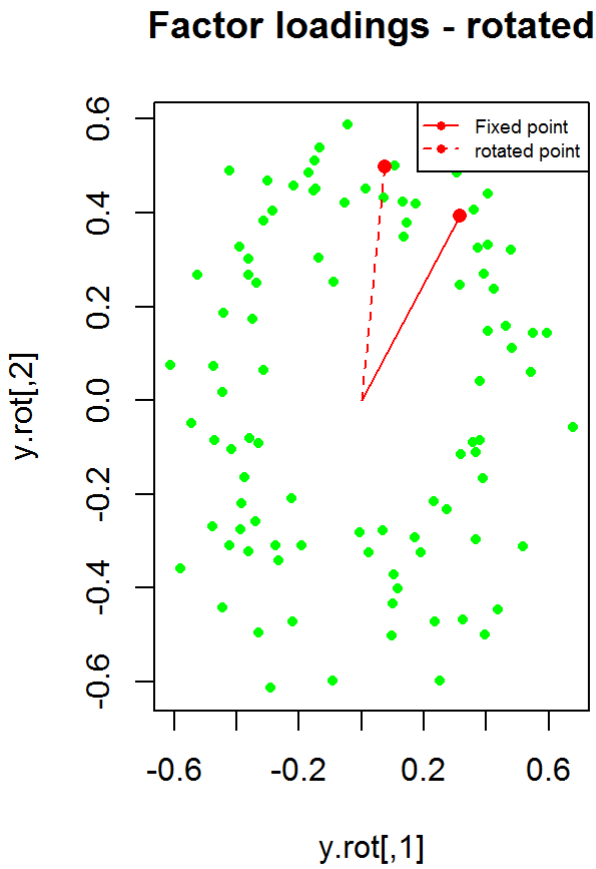




We can see from the figure above the rotation effect of our factor score, which is velocity vector, by 30 degrees counter-clockwise. In the first row of graphs we have the original graph and colored by the signs of their factor score values, and in the second row of graphs we applied the same coloring from the first graph and we can observe that all factor scores are rotated by 30 degrees. In a hypothetical situation where the velocity vectors have assigned directions, we can interpret this rotation as adjusting the velocity vector by 30 degrees to our desired direction due to offset. However, since the observed data is not going to change, we want to observe how the preferred directions will act.

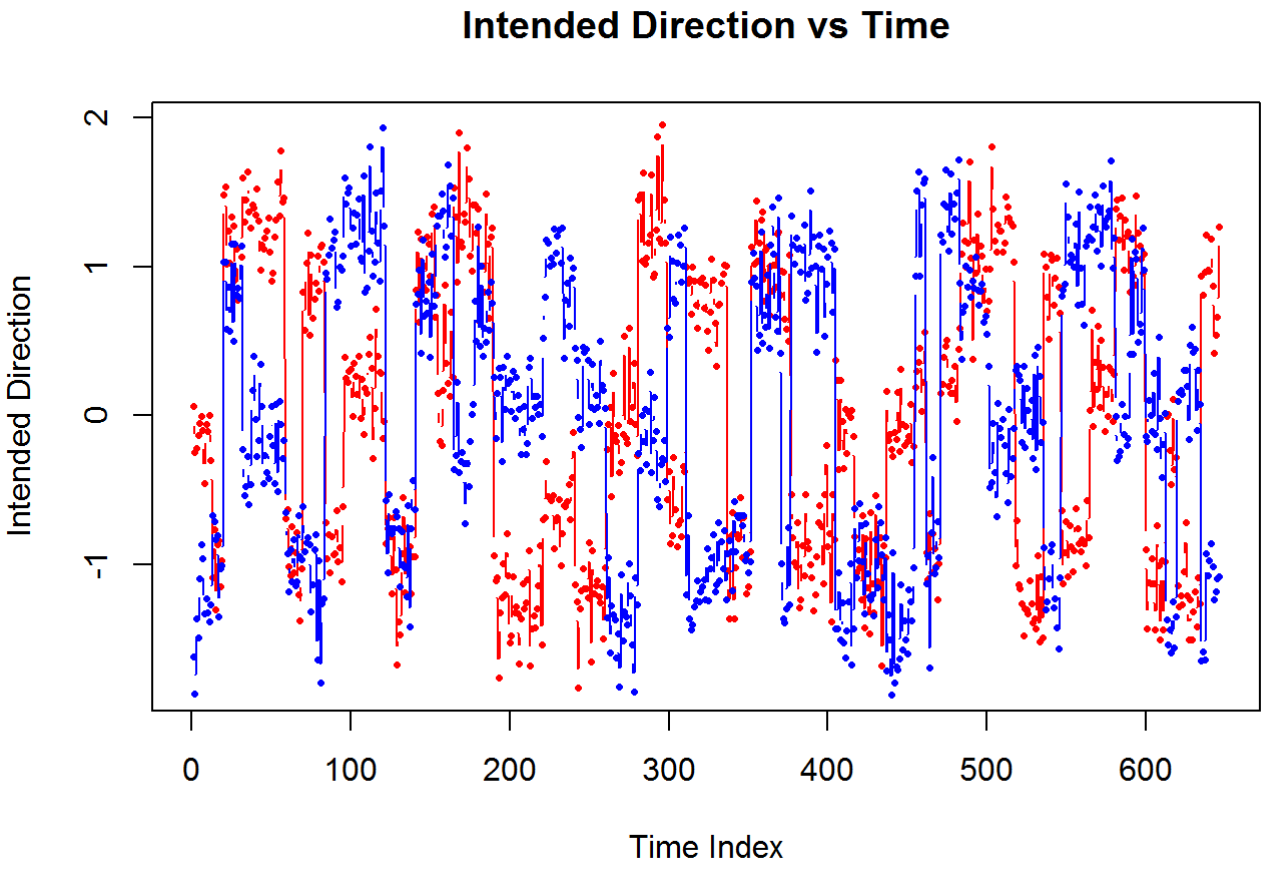
In factor models, we know that if we rotate both factor loadings and factor scores rotates the same degrees to the same direction, the matrix multiplication of the rotated scores and loading is the same as the results before rotation. We hypothesize that if the factor loadings, which is the preferred direction, rotates the 30 degrees counter-clockwise, it will yield the same matrix multiplication result with factor scores as prior to rotation.



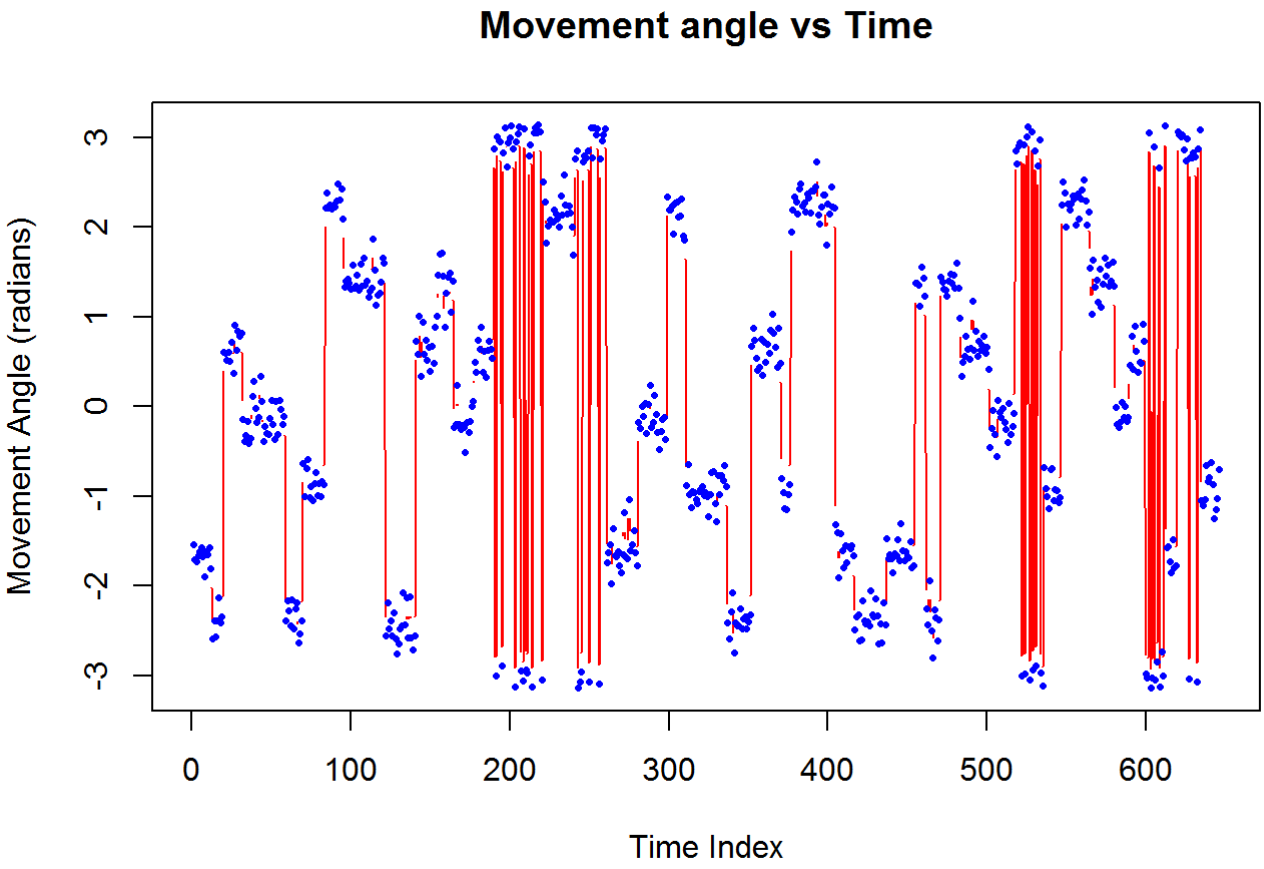


We picked a fixed point in the pre-rotation factor loadings and colored it red and we highlight the same point in our rotated loadings scatter plot. We can observe from the second graph that our factor loadings rotated 30 degrees counterclockwise. We examined through calculation that this rotation does not yield a different result than what we have piror to rotation. This implies that rotation will not change how we interpert factor score estimates and factor loadings regardless of rotation. Our interpertation of the factor score estimates and factor loadings are relative – we look for the relationship between each other but not their relationship with pre-determined values or directions. Therefore, we do not need to know what is the actual direction of our velocity vector estimates, and all our inference and estimates are relative to each other without any set starting direction.

With that conclusion, we want to disocver the procedure of the experiment. In particular, we want to find out when the distinct break between the trails where the subject monkey change its hand movement direction. We look into our factor scores across both factors.



In the graph above, factor one scores is denoted by color red and factor two scores is denoted by color blue. We can see from the figure that factor scores of factor one and factor two behave in clusters, and across time the clusters have different patterns and changes, which indicates that the monkeys changed the direction of their hand movement. We want to find out when these distinct breaks happen during our experiment. Since we interpert our direction of movement as factor scores, which is a two-dimensional plane with factor one being the x-axis and factor two being the y-axis, we can interpert the relative distance between factor scores as an angle.



In the graph above, the x-axis in the graph is the time index, and the y-axis is the radiant computed from arctangent of our factor scores. Negative values indicates that the movement is occurring at the 'lower quadrant' of the two-dimensional plane. Each point represent an angle of movement that the monkey is required to do over a short period of time, and each cluster indicates the same movements is being repeated over the time period. Therefore, each different cluster represents a distinct break where the monkey change its direction of movement. The exception is around time index 200-250, 530-630 where over some of these time span the points are on the opposite end of the y-axis. This is because in our graph -3.14 and 3.14 radiant is the same direction, the movement is 180 degrees from our relative starting point of 0 radiant. Even if the points are on both extremes of the y-axis, they are still movements around the same angle. Therefore we can consider them also as one cluster.

Model Building: Mixture Models

We want to explore our models beyond the factor model. In factor analysis we assume that our latent variables are being continuously adjusted with the distributions of observables changes continuously. However, we want to discover through mixture model if some of the latent variables are categorical or ordinal. First, we begin to fit a three mixture model. This is approiate because any linear factor model with q factors is equivalent to some mixture model with one more clusters, because the two models have the same means and covariances ((Bartholomew, 1987, pp. 36-38). Following this rule, the three cluster model is can match the covariance matrix of 2 factor model very well and is likely to have a good log-likelihood.

```
# First we fit a three mixture model
mix.mdl.3 = npEM(neur,mu0=3)
```

And then we prefrom cross-validaiton to find out the log-likelihood of our mixture model with it's standard error.

The log-likelihood for our three mixture model is -9349.2908 with standard error 500.3004. # comment more

We then decide to see how a eight-cluster mixture model will preform on our dataset. Eight-cluster model is valid due to the same reasons as three-cluster model, and it has a higher number of clusters than factors. Also, the eight clusters is resonable because each cluster may indicate each directional vectors.

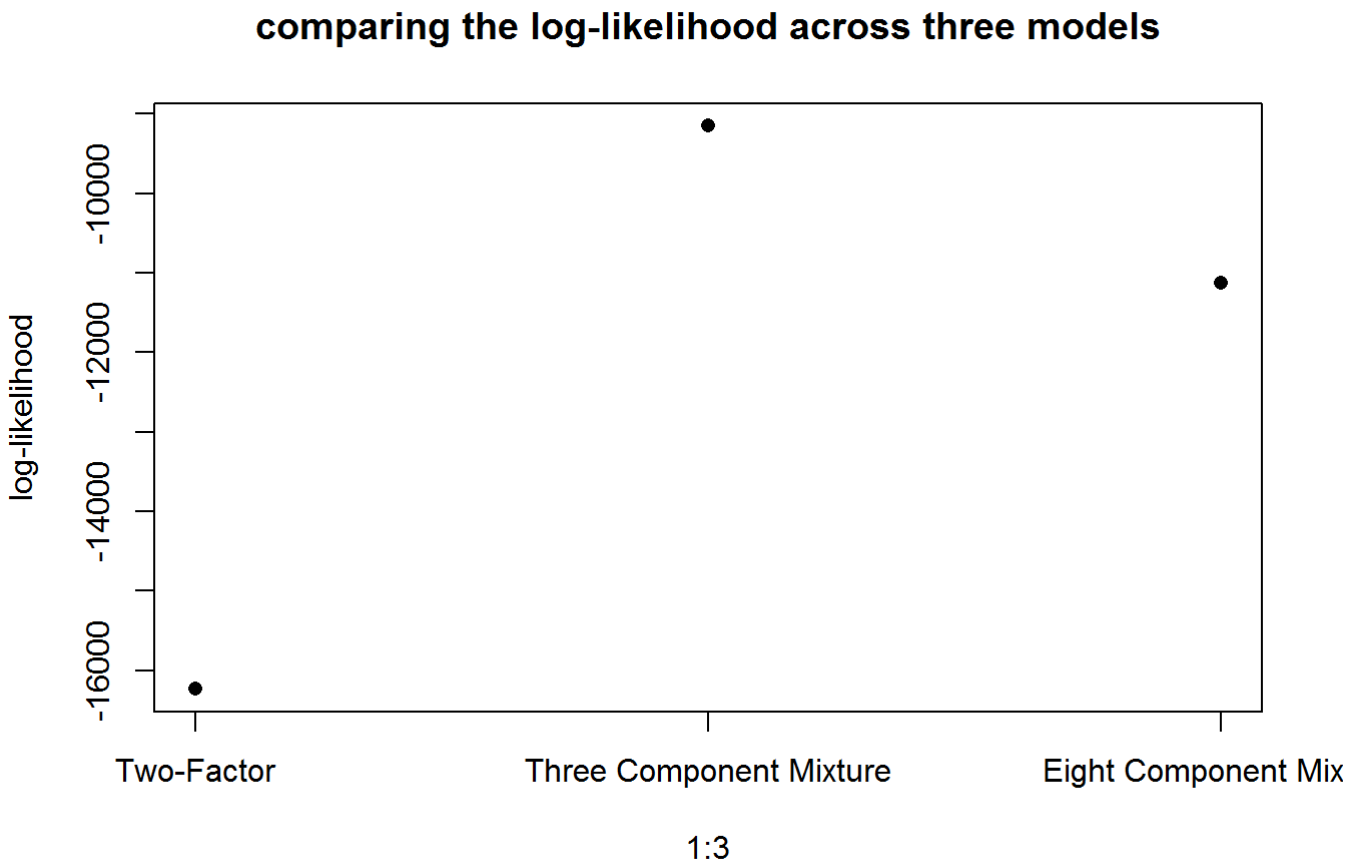
```
# First we fit a three mixture model
mix.mdl.8 = npEM(neur,mu0=8)
```

Then we find out the cross-validated log-likelihood in order to compare it with the previous models on its goodness-of-fit.

The log-likelihood for our eight mixture model is -11210.4754 with standard error 334.3583.

Model Comparsion and Model Selection

Now we have three models, and we want to find out our optimal model through model comparsion. We compare the cross-validated log-likelihood across all three models, as these values can provide us a good estimate of how the goodness-of-fit of each model is.



We can see from our plot above that three component mixture model has the highest log-likelihood amongst all three models, indicating that it has the best goodness-of-fit across all three models. Since our log-likelihood is computed through cross-validation, this indicates that the three mixture clustre model is better at predicting new data.

Conclusion

We examined three different models constructed on our dataset: two factor mode, three cluster mixture model and eight cluster mixture model. The two factor model provides us more direct graphical visualization with regard to clustered velocity vectors, allowsus to see the movements better on a two dimensional plane, and allows us to make more direct inference on the preferred direction of neuron or directional tuning. However, the cross-validation log-likelihood for this model does not prefrom as well as the three-component mixture model and the eight component mixture model. Therefore, it does not make a good fit on a new set of data. In this case, we choose the three component mixture model as our final model.

Since the three component mixture model have three clusters, it suggests that there is three discrete or categorical latent variables in our data, which is different than the two-dimensional plane interpertation we previously have with the two factor model. It may possibily be that the preferred direction can have something to do other than the monkey's movement, or the movement speed may influence how much the neurons produce spikes in the monkey's brain. It is possible for us to discover the preferred directions for the neurons, but the preferred directions may have to combine with some extra variables.