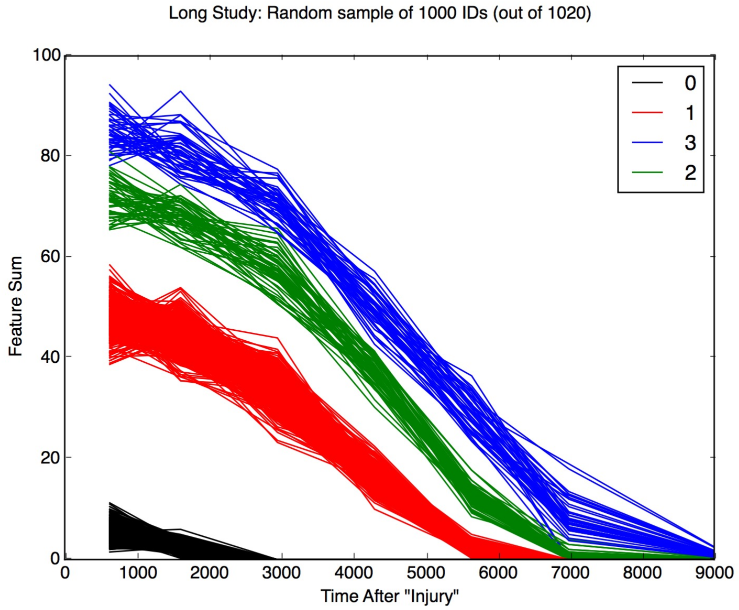
Research progress

From what we have gathered from the research so far it seems as though the group based modeling approach involves fitting polynomials to the longitudinal curves. Then statistics (t-scores) can be calculated that indicate how well a specific curve fits a particular group/class. The method of calculating these statistics is still unclear to me.

After studying Nagin's overview of Group Based Trajectory Modeling we came across his text book called "Group-Based Modeling of Development". We feel this book will provide some of the applicable techniques that his paper glossed over.

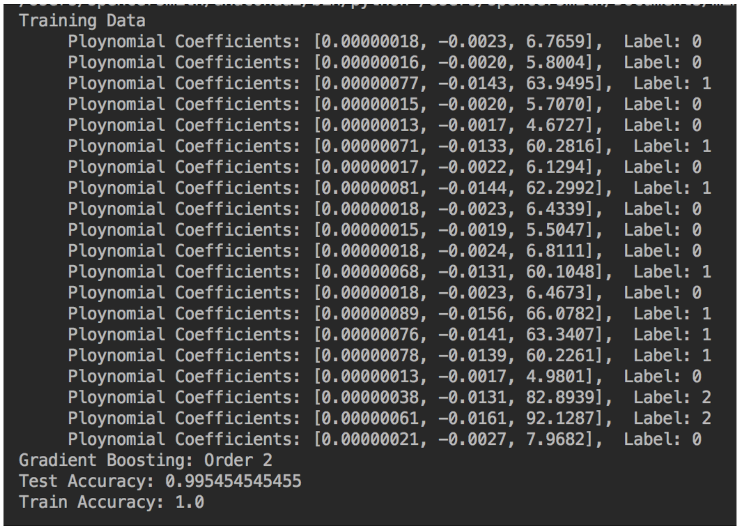
Simulated data

This week we continued working with the simulated data. If you remember from last week we were able to design a reasonably effective bucketing technique so that the data could be compared at consistent points on the time axis. Here is a plot of the results from last week:



Using this data our goal is to accurately classify each group. To do this we fit polynomials to each trajectory then the coefficients from each curve were used as features for machine learning.

The image to the right shows the structure of the training data and the results of the classification algorithm.



```
Training Data
     Ploynomial Coefficients: [0.00000018, -0.0023, 6.7659],  Label: 0
     Ploynomial Coefficients: [0.00000016, -0.0020, 5.8004],  Label: 0
     Ploynomial Coefficients: [0.00000077, -0.0143, 63.9495],  Label: 1
     Ploynomial Coefficients: [0.00000015, -0.0020, 5.7070],  Label: 0
     Ploynomial Coefficients: [0.00000013, -0.0017, 4.6727],  Label: 0
     Ploynomial Coefficients: [0.00000071, -0.0133, 60.2816],  Label: 1
     Ploynomial Coefficients: [0.00000017, -0.0022, 6.1294],  Label: 0
     Ploynomial Coefficients: [0.00000081, -0.0144, 62.2992],  Label: 1
     Ploynomial Coefficients: [0.00000018, -0.0023, 6.4339],  Label: 0
     Ploynomial Coefficients: [0.00000015, -0.0019, 5.5047],  Label: 0
     Ploynomial Coefficients: [0.00000018, -0.0024, 6.8111],  Label: 0
     Ploynomial Coefficients: [0.00000068, -0.0131, 60.1048],  Label: 1
     Ploynomial Coefficients: [0.00000018, -0.0023, 6.4673],  Label: 0
     Ploynomial Coefficients: [0.00000089, -0.0156, 66.0782],  Label: 1
     Ploynomial Coefficients: [0.00000076, -0.0141, 63.3407],  Label: 1
     Ploynomial Coefficients: [0.00000078, -0.0139, 60.2261],  Label: 1
     Ploynomial Coefficients: [0.00000013, -0.0017, 4.9801],  Label: 0
     Ploynomial Coefficients: [0.00000038, -0.0131, 82.8939],  Label: 2
     Ploynomial Coefficients: [0.00000061, -0.0161, 92.1287],  Label: 2
     Ploynomial Coefficients: [0.00000021, -0.0027, 7.9682],  Label: 0
Gradient Boosting: Order 2
Test Accuracy: 0.995454545455
Train Accuracy: 1.0
```

Gradient boosting was used for classification and we were able to get a classification accuracy of >99.5%. Initially we fit a third degree polynomial, however, we found that the higher order coefficient actually hurt the models predictive power. Second degree performed the best (first degree performed almost as good). Looking at the data it appears that the most influential predictor is the level of the intercept.

Overall, these strong results are not surprising as the came from simulated data and were designed to be very separable. However, we feel this could be a technique that can be effectively applied to EEG data.

We also explored how to calculate "confidence regions" for polynomials to get statistics (p-values) from specific curves. We are confident that there are good methods to do this in Nadin's book or elsewhere. I will continue to study this area.

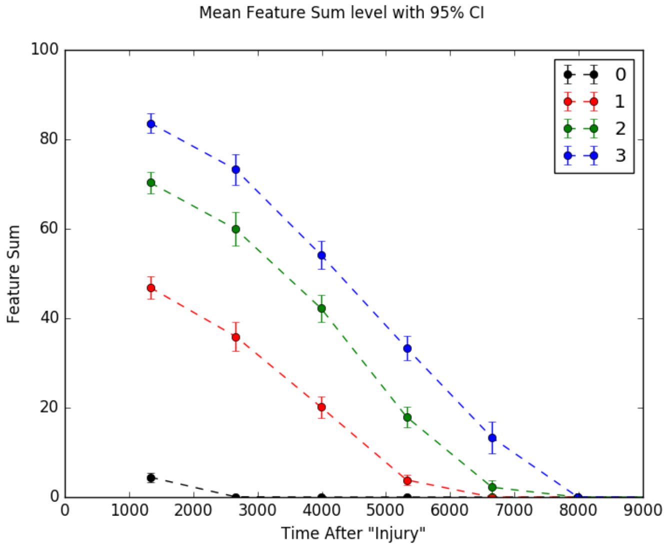
Business Case

As we are designing these classification techniques I want to make sure that they are in line with our specific business case. For instance, it is great to be able to fit a curve to a trajectory with 5+ observations and accurately classify it (as ASD or Typical), however, as you have mentioned, we need to be able to give the 'customer/patient' a classification probability after each appointment. Starting from the first appointment!

Maybe this will be a combination of single point predictions(t-test) and curve classifications. The very first appointment we will simply run a t-test to see if the "patient's" feature level is different from Typical. Then, on the second appointment we can compare linear trajectories and so on. My point is that is seems that we will need to design different classification and statistical approaches for each appointment (aka length of trajectory).

Point based comparisons

This plot shows the mean trajectories from each class with 95% confidence intervals. To be clear these are confidence intervals for each specific point in time. They are NOT confidence regions for the curve. This needs to be studied more. Something like this could be used for providing "first appointment" statistics.


Mean Feature Sum level with 95% CI
Feature Sum
Time After "Injury"