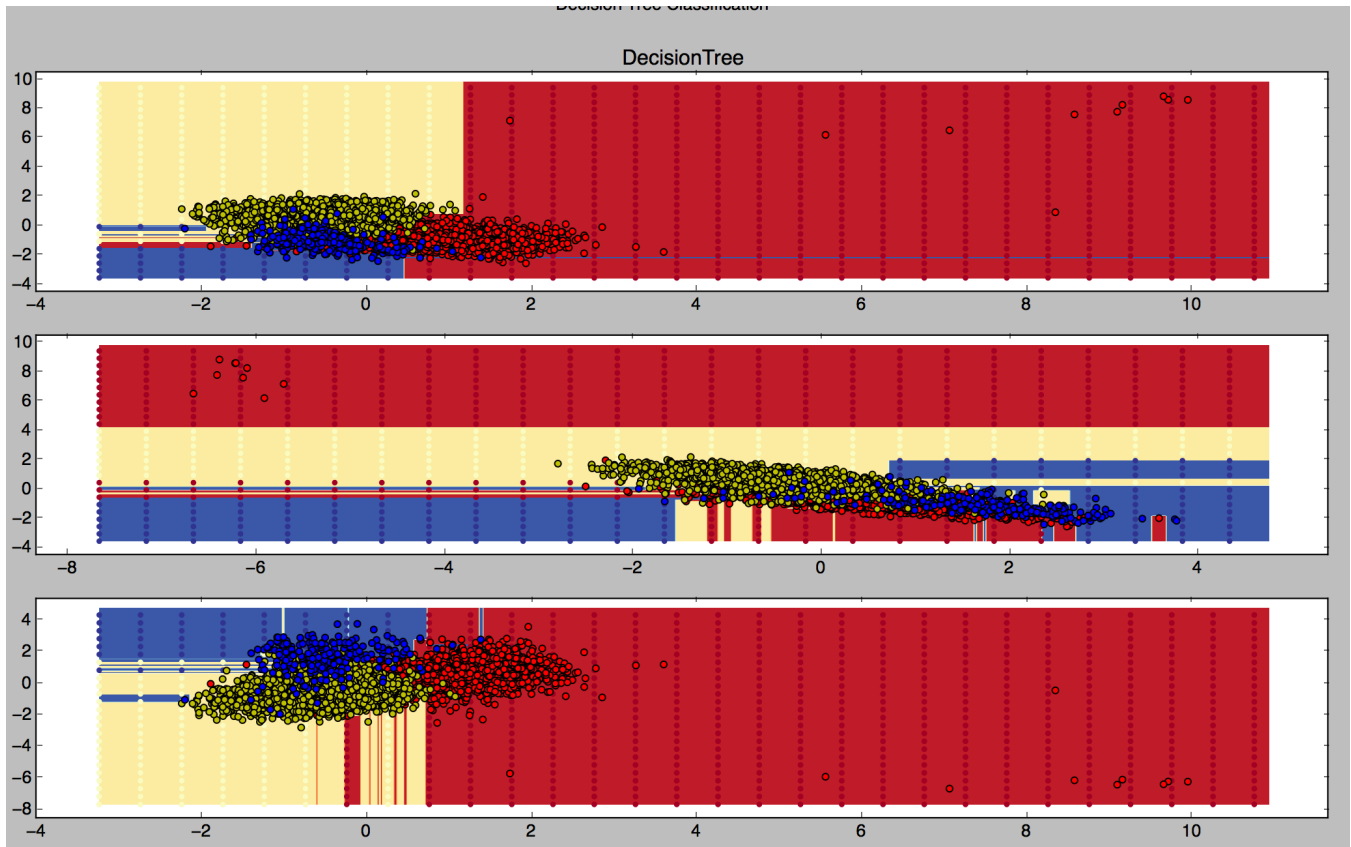6/28/2016

**Today's Goal** Start looking at methodologies for clustering.

Got the decision tree plot running. This is for subject 1038 when ran without any parameter changes (from top to bottom- EMG Power x Delta Power, Ratio x Delta Power, EMG Power x Ratio


Unfortunately, it seems like I can only plot the decision tree when it uses 2 features, so I did code extraction instead.

The Decision Tree without a max depth was highly complex with accuracy 100%. Too much overfitting for a neat threshold.
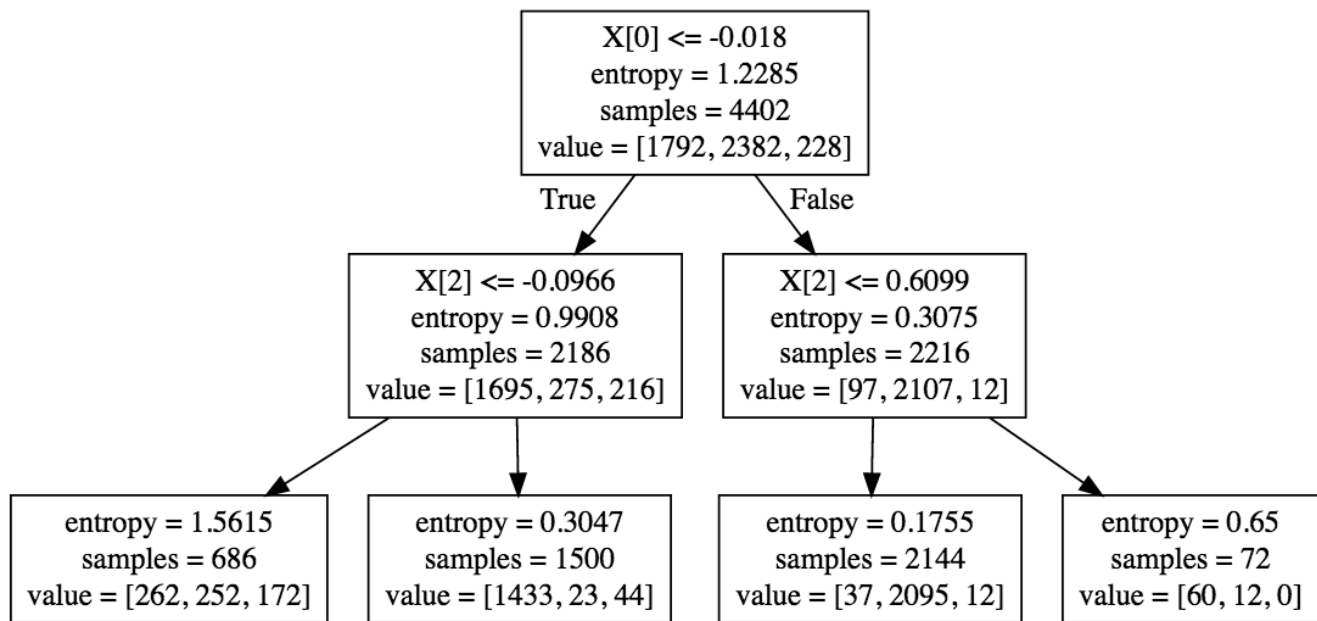
Setting max_depth=3 gave a 90.4% accuracy with a decision graph looking like the following:
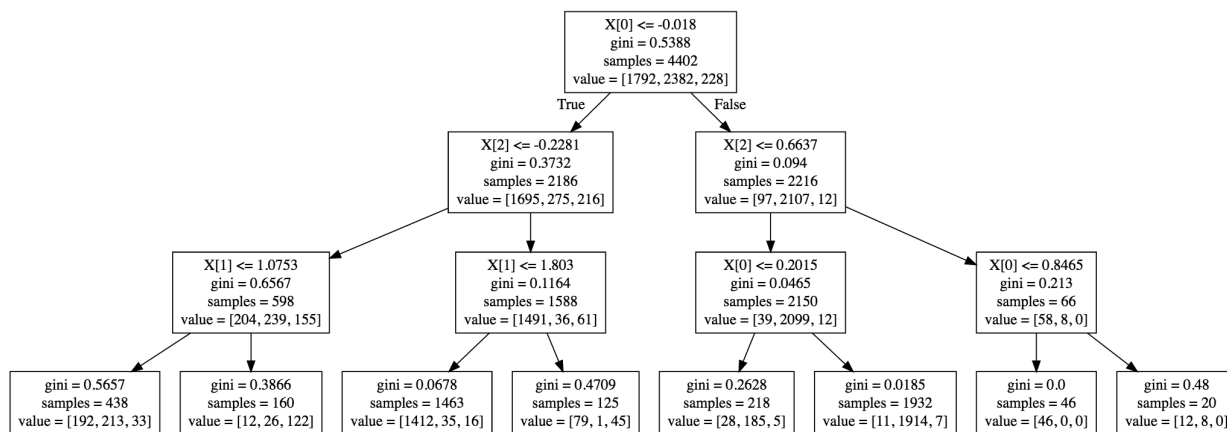
A max_depth=2 gave a 88.3% accuracy with a decision graph looking like the following:
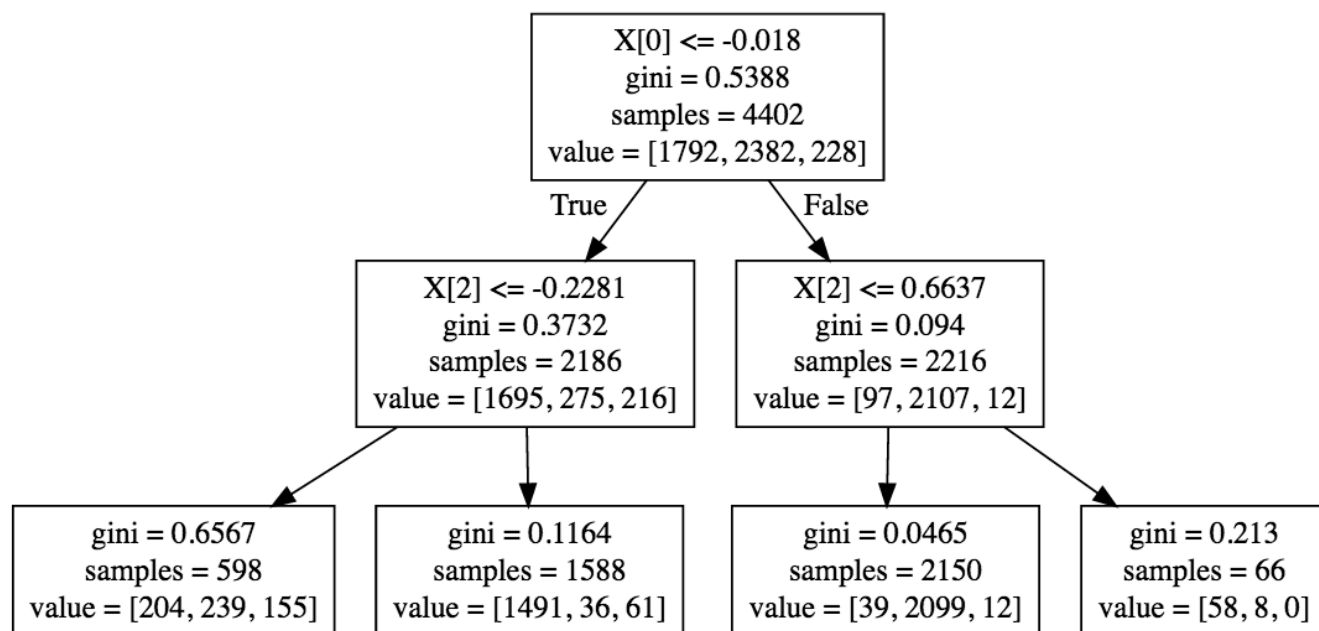
Changing criterion to entropy instead of gini impurity lowered accuracy to 87.6% and gave the following graph:

```
                          X[0] <= -0.018
                          entropy = 1.2285
                          samples = 4402
                      value = [1792, 2382, 228]
                    True  /            \  False
         X[2] <= -0.0966                    X[2] <= 0.6099
         entropy = 0.9908                   entropy = 0.3075
         samples = 2186                     samples = 2216
    value = [1695, 275, 216]           value = [97, 2107, 12]
       /          \                        /          \
entropy = 1.5615   entropy = 0.3047   entropy = 0.1755   entropy = 0.65
samples = 686      samples = 1500     samples = 2144     samples = 72
value = [262, 252, 172]  value = [1433, 23, 44]  value = [37, 2095, 12]  value = [60, 12, 0]
```

Though this seems like the best choice, there were 0 REM states calculated with this tree. Here are the statistics for various parameters:

Gini purity metric with depth 2:

Score of Decision Tree: 0.8830077237619264
Wake. Calculated 1654 with actual 1792
NREM. Calculated 2748 with actual 2382
REM. Calculated 0 with actual 228

Gini purity metric with depth 3:

Score of Decision Tree: 0.904815992730577
Wake. Calculated 1654 with actual 1792
NREM. Calculated 2588 with actual 2382
REM. Calculated 160 with actual 228

Entropy metric with depth 2:

Score of Decision Tree: 0.874602453430259
Wake. Calculated 2258 with actual 1792
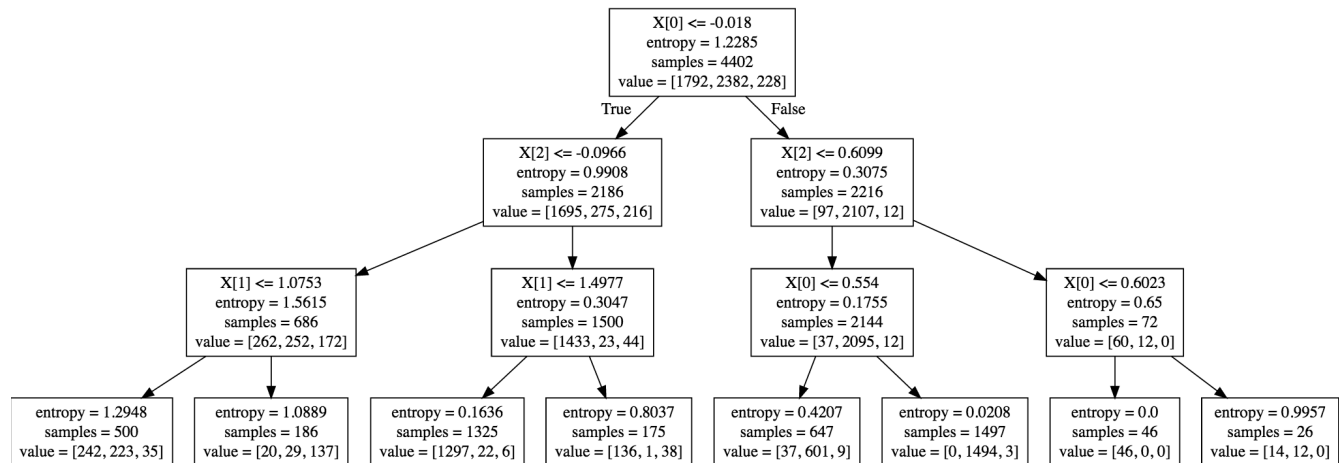NREM. Calculated 2144 with actual 2382
REM. Calculated 0 with actual 228

Entropy metric with depth 3:

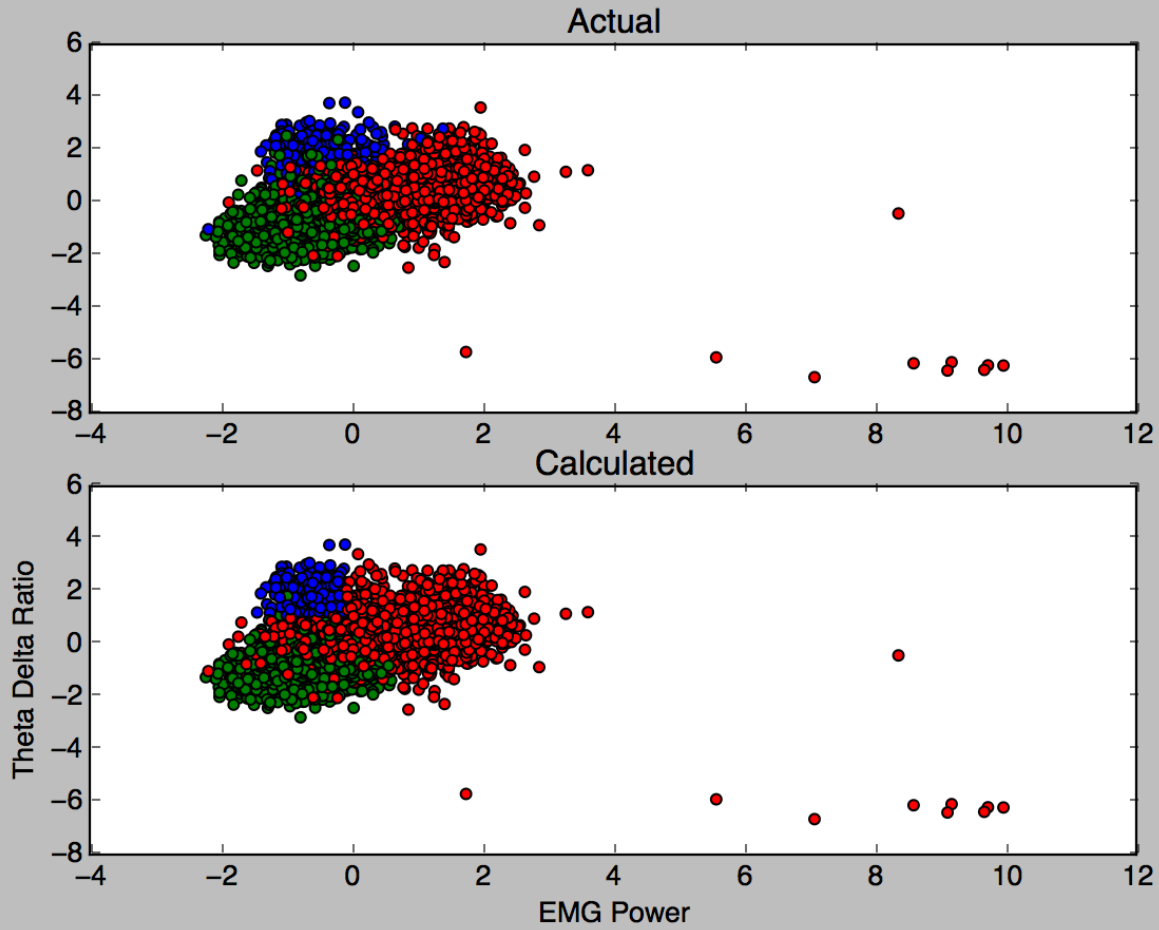Score of Decision Tree: 0.9011812812358019

So, for the current situation, the Entropy metric with depth 3 is the ideal choice.

Graph looks like this:

```
                              X[0] <= -0.018
                              entropy = 1.2285
                              samples = 4402
                          value = [1792, 2382, 228]
                         True                    False
                ┌─────────┘                        └─────────┐
        X[2] <= -0.0966                                 X[2] <= 0.6099
        entropy = 0.9908                                entropy = 0.3075
        samples = 2186                                  samples = 2216
    value = [1695, 275, 216]                        value = [97, 2107, 12]
       ┌────────┴────────┐                          ┌────────┴────────┐
  X[1] <= 1.0753     X[1] <= 1.4977           X[0] <= 0.554       X[0] <= 0.6023
  entropy = 1.5615   entropy = 0.3047         entropy = 0.1755    entropy = 0.65
  samples = 686      samples = 1500           samples = 2144      samples = 72
value=[262,252,172] value=[1433,23,44]     value=[37,2095,12]   value=[60,12,0]
   ┌────┴────┐        ┌────┴────┐              ┌────┴────┐          ┌────┴────┐
```

| entropy = 1.2948 | entropy = 1.0889 | entropy = 0.1636 | entropy = 0.8037 | entropy = 0.4207 | entropy = 0.0208 | entropy = 0.0 | entropy = 0.9957 |
| samples = 500 | samples = 186 | samples = 1325 | samples = 175 | samples = 647 | samples = 1497 | samples = 46 | samples = 26 |
| value = [242, 223, 35] | value = [20, 29, 137] | value = [1297, 22, 6] | value = [136, 1, 38] | value = [37, 601, 9] | value = [0, 1494, 3] | value = [46, 0, 0] | value = [14, 12, 0] |

The Clustering Plot seems to look identical to the actual:

Wrote the explicit decision tree for a threshold classification and generalized it to subject 1039. I changed the EMG Power qualification (left node of root) to bias the algorithm towards identifying REM. Below is the resulting clustering:
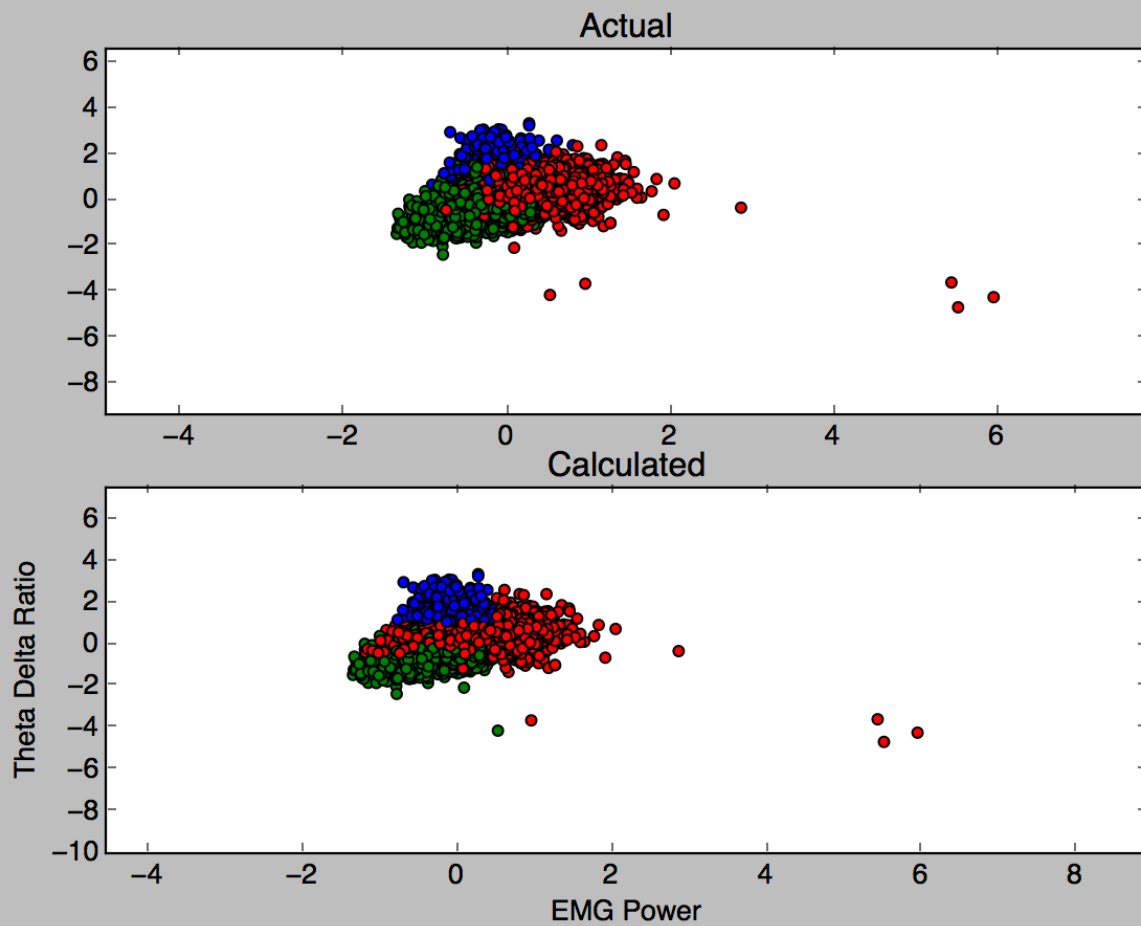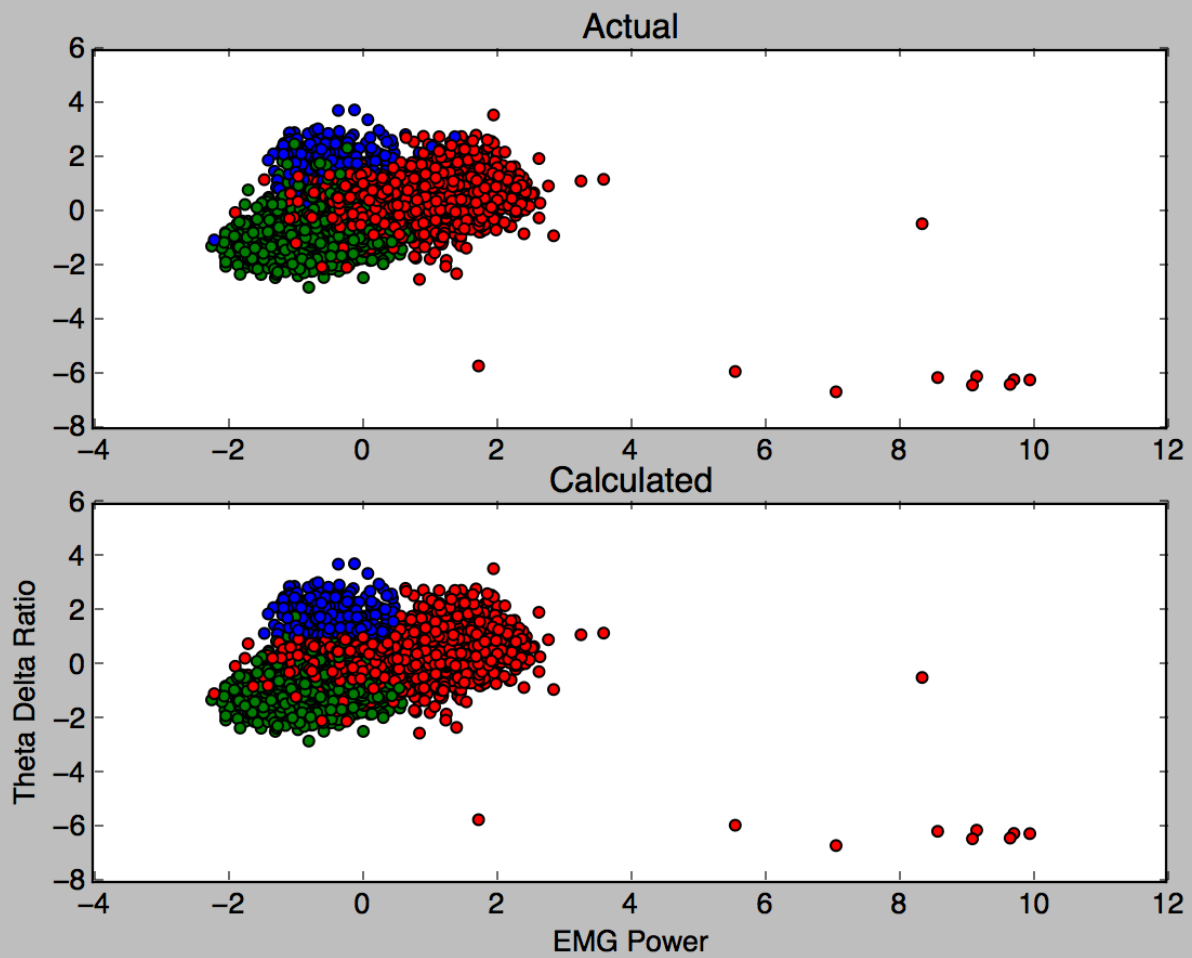
Actual

Calculated

Theta Delta Ratio

EMG Power

Results for subject 1040:
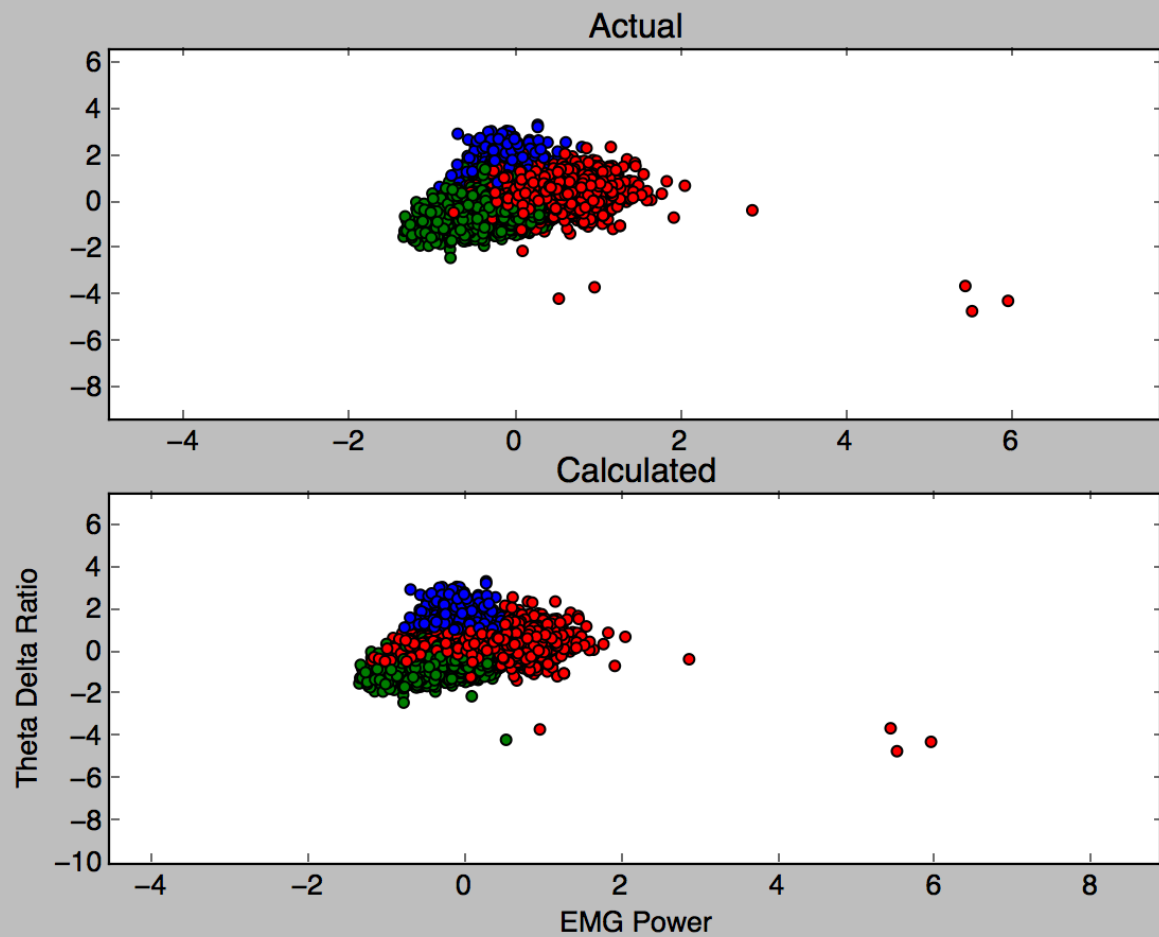
Wake. Calculated 2045 with actual 2143
NREM. Calculated 2423 with actual 2316
REM. Calculated 228 with actual 237

Results reran for subject 1032:

Wake. Calculated 1984 with actual 1792
NREM. Calculated 2144 with actual 2382
REM. Calculated 274 with actual 228

Hypnograms show false positives at regular intervals, but do hit REM when its supposed to.