7/01/2016

Today's Goal Learn about PCA and implement on whole signal to see effects of classification.

Restructured the code so that it can be used interactively.

The FASTER paper uses PCA on the whole signal (transformed with FFT) to feed into non-parametric density clustering.

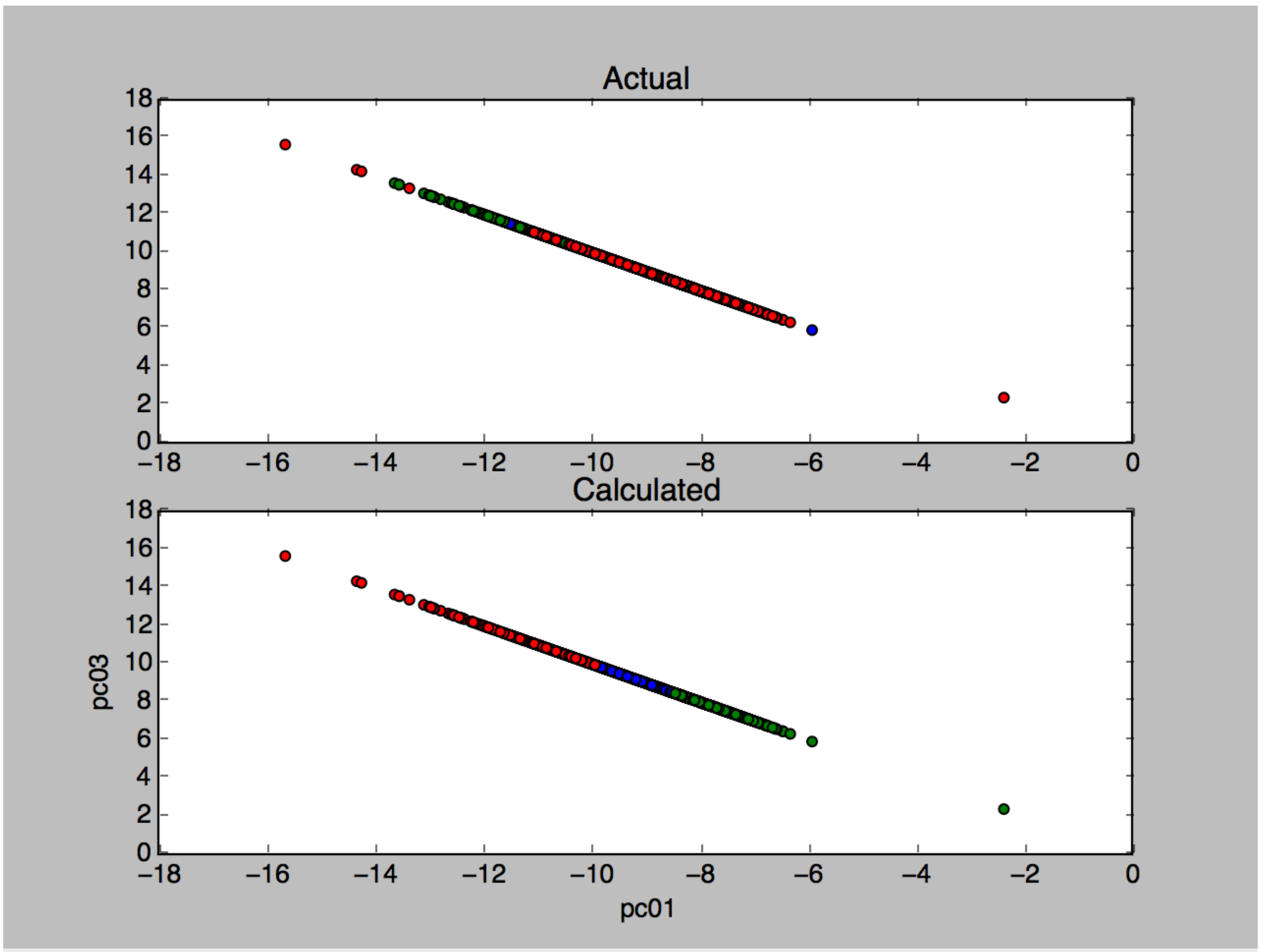
Principal Component Analysis (PCA) takes multidimensional data and compresses it into components that maximizes their variance to lose the least amount of information.

I implemented and ran PCA on the FFT transformed data (EEG and EMG). Below are the results:

```
Feature pc1 has F=123.71253448274315 and p=5.357171728473936e-53
Feature pc2 has F=3.5556167825974723 and p=0.028645871490168l3
Feature pc3 has F=123.71253448274315 and p=5.357171728473936e-53
Feature pc4 has F=0.6867283031222164 and p=0.5032737007409078
```

The tests are all significant and the F values are decent, but they are not nearly as large as the F values from the original features.

I attempted to use kmeans and plot the results. They looked like this:



The actual REM cluster is not very well defined, resulting in a poor fit for this particular method.

I ran Analysis of Variance on all of my features that I have ever implemented:

```
Feature delta has F=2705.7775440128944 and p=0.0
Feature theta has F=1440.078398730006 and p=0.0
Feature ratios has F=1218.7479012445829 and p=0.0
Feature emgPower has F=1812.5177705200495 and p=0.0
Feature autocorr has F=208.23518191935557 and p=3.919106362973161e-87
Feature emgMed has F=218.80922554387118 and p=2.5543945698901694e-91
Feature largeRatio has F=3573.6889823766674 and p=0.0
```

Out of these, the delta, EMG Power, and Large Ratio (0.5-22 EEG Range/ 0.5-50 EEG Range) are the best for differentiating. However, since REM is the priority, it would be useful to understand which feature separates REM the best (Analysis of Variance on REM and NREM+Wake groups).

Below are the results of that analysis:

```
>>> SEPARATION OF WAKE
Feature delta has F=3003.880303147028 and p=0.0
Feature theta has F=2839.1233573484865 and p=0.0
Feature ratios has F=685.3932118150088 and p=1.5727251753274662e-140
Feature emgPower has F=3620.7400335681873 and p=0.0
Feature autocorr has F=413.10030852914315 and p=7.471526398244455e-88
Feature emgMed has F=43.48308302430702 and p=4.781285666075207e-11
Feature largeRatio has F=5363.78369471325 and p=0.0
Feature Sign Inversions has F=7538.621549044099 and p=0.0
>>>SEPARATION OF NREM:
Feature delta has F=5319.032282306535 and p=0.0
Feature theta has F=2340.2131273745244 and p=0.0
Feature ratios has F=1771.8153950020196 and p=0.0
Feature emgPower has F=2644.747225564264 and p=0.0
Feature autocorr has F=299.4275425040895 and p=5.937569873068201e-65
Feature emgMed has F=209.013414099682 and p=2.5597819287987672e-46
Feature largeRatio has F=6929.558984580008 and p=0.0
Feature Sign Inversions has F=9215.658623388501 and p=0.0
>>> SEPARATION OF REM
Feature delta has F=296.77188438034267 and p=2.067694085839179e-64
Feature theta has F=17.759617109392078 and p=2.5567175487154313e-05
Feature ratios has F=793.6327757979561 and p=1.1136079608321774e-160
Feature emgPower has F=55.656416187089555 and p=1.034120444711392e-13
Feature autocorr has F=29.708204964074675 and p=5.2968164491443144e-08
Feature emgMed has F=317.7755752019425 and p=1.092746517216078e-68
Feature largeRatio has F=59.38301819052182 and p=1.593498144670475e-14
Feature Sign Inversions has F=34.547945470459915 and p=4.466275163729487e-09
```

Sign Inversions is extremely effective at separating Wake and NREM, but not so effective at separating REM. For REM, the Delta/Theta ratio and medium EMG are the best separators.

Even when ignoring outliers, unsupervised clustering (such as kmeans) isn't being very effective despite having all these features. My options are either to find a better clustering algorithm than the one I am using or find a feature that separates REM extremely well.

There is also the possibility of using a limited amount of data for a supervised version (like what I did with the decision tree thresholding)