Analysis of MEG BioMarkers To Determine Parkinson’s And Related Diseases
November 29, 2016

Summary

The EEG data provided by UCH at Denver Neuroscience Department has been analyzed using the data science techniques most likely to achieve results. The small sample was diagnosed for Parkinson’s, Essential Tremor and Parkinson’s Plus. The sample size was found to be too small to discriminate between Essential Tremor and Parkinson’s Plus but discriminating between Parkinson’s and Other Than Parkinson’s, i.e., Essential Tremor and Parkinson's Plus, showed potentially effective results.

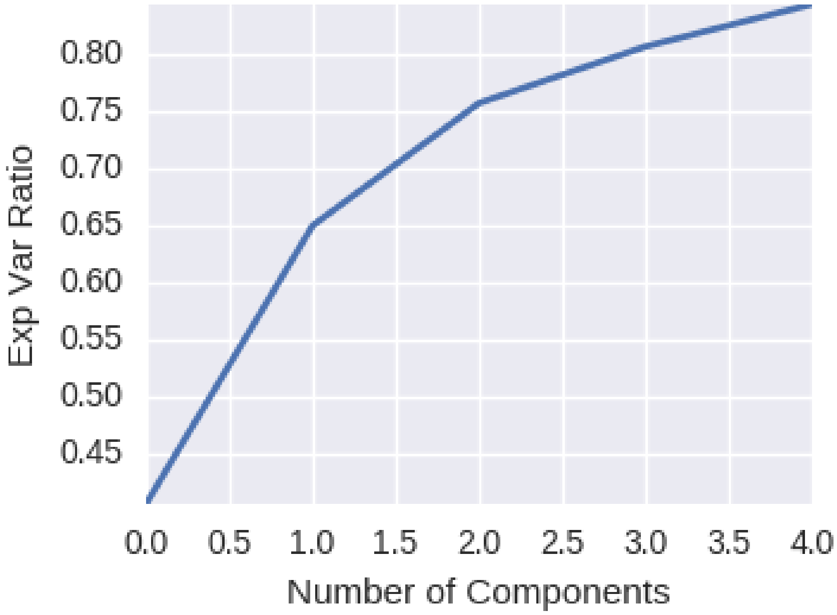
It was found that Bayesian Analysis using logistic models was effective for the larger feature sets and shows promise if the sample size can be increased. Bayesian Analysis (‘BA’) is desirable due to its statistical foundation realizing easily measured significance. The most effective method on this small sample size is a Random Forest Classifier (‘RFC’). This technique is less sensitive to small sample sizes but is less desirable since it does not readily yield statistical signicance. The RFC was effective in modeling Parkinson's and Not Parkinson's. This validates the potential use of the EEG measurements for this purpose. This report describes the results with all measurements (the 'phenotypes') as the sample base. The results for the two consolidated features is available but were not sufficently effective to merit additional description.

Samples

The sample size is extremely small, 44 samples, and the number of measurements, over 500 possible, introduce statistical significance issues. The data has two forms: (1) the samples compared on two consolidated measurements and (2) the samples compared on the entire set of measurements available from the technique ('all phenotypes'). In addition, the small sample set is divided into three classes of Parkinson’s, Essential Tremor and Parkinson’s Plus. This is an extremely small sample set for the data complexity presented. It was determined first that only discrimination between Parkinson's and Not Parkinson's (i.e, Essential Tremor and Parkinson's Plus) is feasible. A calculation of sample size required shows approximately 2,500 samples being required for an effective Bayesian Analysis including the three diagnosed diseases.

Analysis

Principal Companents Analysis('PCA'): The data set for 'All Phenotypes' was reduced to five major features using PCA techniques. This extreme reduction was necessary to reduce the number of features, i.e., components, to be significantly less than the sample size. It is found that these five components explains over 85% of all the variation in the sample. Given a larger sample the individual phenotypes comprising the components can be analyzed individually for their contribution.



Exp Var Ratio = Explained Variance Ratio

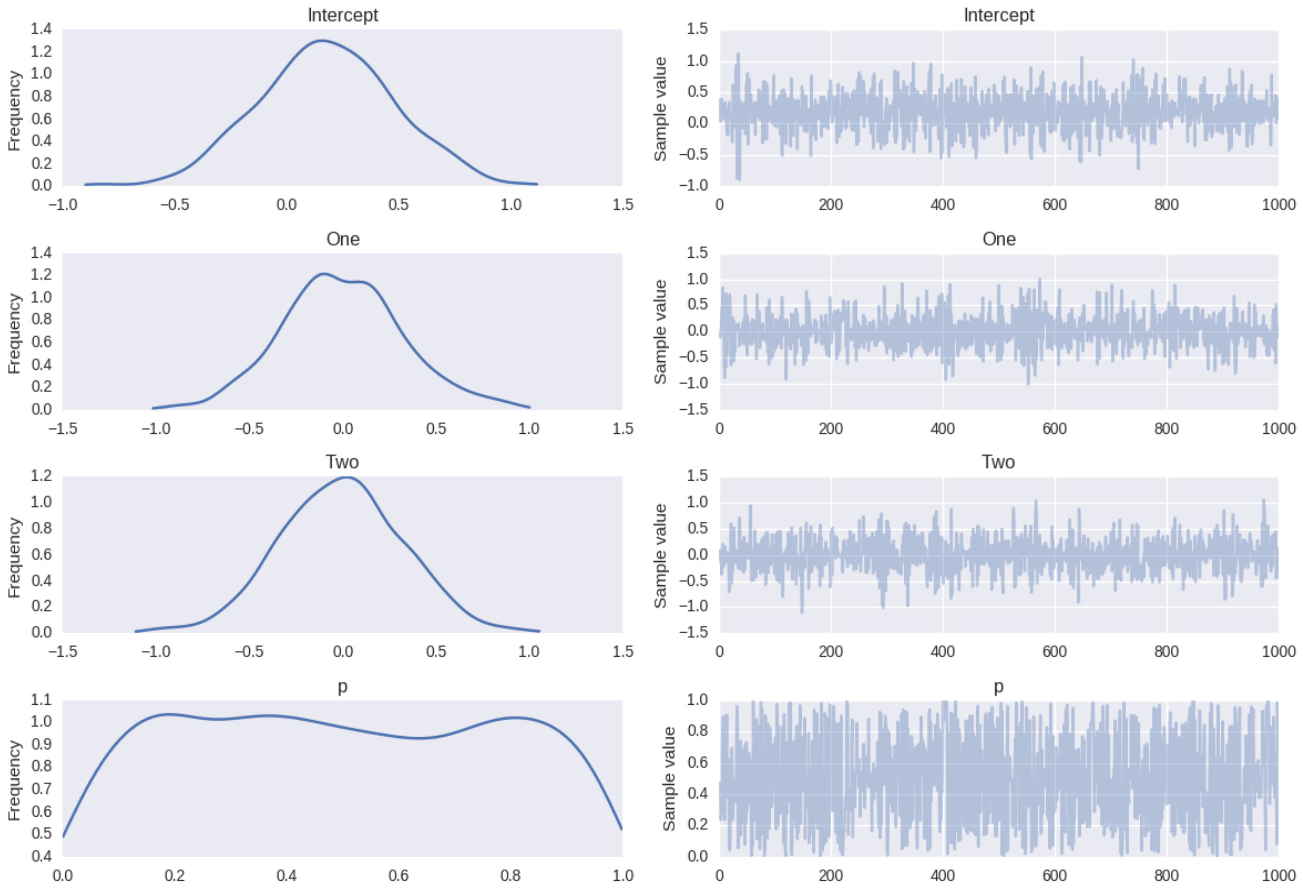
Since less than 5 components explains over 85% of the variance in the data it is likely that combinations of a few measurements will yield results in a larger sample set. The current sample set is too small to yield significant results.

Bayesian Analysis Using Logistic Models

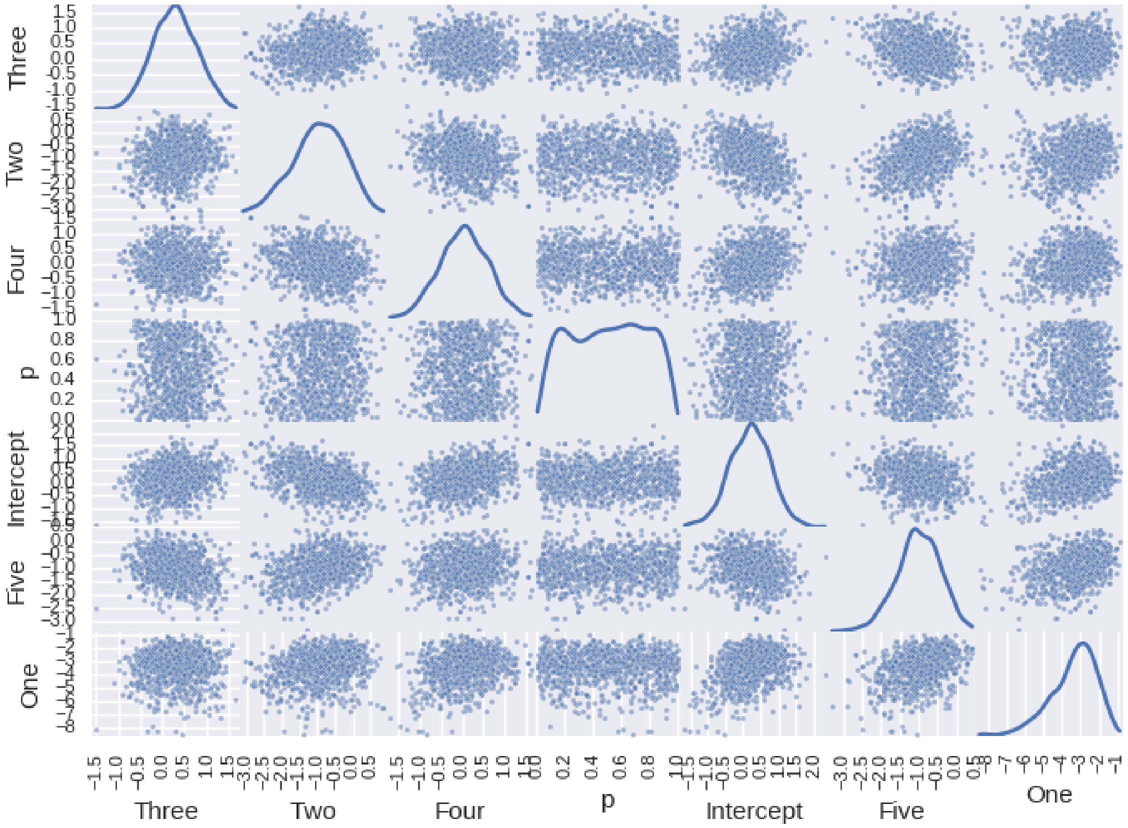
The Bayesian sampling over 2000 iterations did not completely converge. The sample size is too small. The confusion matrix did reveal a relative high degree of performance with 13 out of 24 Parkinson's diagnosis being confirmed and 10 out of 20 Not Parkinson's also being confirmed. Given the weak convergence on this small sample the results are encouraging that a large sample would be effective.[1]

[1] Splitting such a small sample into training and testing samples resulted in no convergence possible.

The convergence is measured by the sample value charts on the right. The sample values appear to be narrowing but insufficiently for convergence. The bottom chart pair is the predicted result. The predictions show random convergence and a nearly uniform distribution. The deviation in the feature variables is too high to yield significant results.



The scatter plots of the relationships between the five components does not demonstrate high collinearity but there is little convergence on the predicted value ('p') either. In order to achieve measurable results a much higher sample count is required.



The standard deviation on the means of each component and the predicted value are too large for reliable results. The sample size increase will reduce the standard deviation.

	Mean	SD	MC Error	95% HPD interval
Intercept	0.201	0.575	0.024	[-0.943, 1.262]
One	-3.370	1.247	0.068	[-5.963, -1.410]
Two	-0.954	0.726	0.033	[-2.405, 0.349]
Three	0.281	0.488	0.018	[-0.661, 1.225]
Four	0.043	0.563	0.023	[-0.940, 1.261]
Five	-1.001	0.577	0.021	[-2.136, 0.063]
p	0.504	0.284	0.008	[0.026, 0.954]

Random Forest Classifer

This non-parametric classifier is less sensitive to sample size than the Bayesian Analysis. It has been applied to determine if sufficient information exists in the data to effectively determine between the three conditions. A statistical sampling method, i.e., Bayesian, is preferable to achieve higher degrees of confidence.

The sample set is divided into two classifiers: the entire set and a split set applying a test set. The split set could only assign 5 samples to the test set and still achieve a reasonable classifier performance.

Confusion Matrix - Test Set

	Actual Not PD	Actual PD
Predicted Not PD	2	1
Predicted PD	0	2

The predictor is accurate on 4 out of 5 on the test data. This is far too small a sample to be reliable but it confirms the liklihood that a larger sample would be effective.

Confusion Matix - Entire Sample Set

	Actual Not PD	Actual PD
Predicted Not PD	20	0
Predicted PD	0	24

The prediction on the training set was exact indicating a high liklihood that an effective model could be built with a larger sample set.