Data Set

The goal of this work is to use experimental seismic signals to predict the times of laboratory earthquakes. Ultimately the goal of this work would be transferring these models to real time monitoring in order to predict the time to actual seismic activity.

PCA

Principal component analysis (PCA) is a technique used for understanding the dimensional structure of a data set. PCA transforms data in such a way that it converts a set of orthogonally correlated observations into a set of linearly uncorrelated variables called principal components. This transformation maximized the largest possible variances for each principal component. There can be as many principle components are there are feature dimensions in the data set and each principal component accounts for the largest possible variance between entries.

In this work we use three different PCA visualization methods to help understand the dimensional structure of the data.

The x-axis of the graph below labels each principal component for the featurized data set. The y-axis accounts for the proportionality of the total variance contained within the data set. As expected, the first principal component accounts for the largest amount of variance and each consecutive principal component account for more variance than the once after it. That total variance across all principal components sums to 1. The red line shows the cumulative variance after each principal component is formed.

The dashed line is an indication of 99% variance of the data. Once can see that the dashed line crosses the red line at the 9th principal component. This indicated that 99% of the variance within the data is still accounted for when the dimensionality of the data is reduced from 16 to 9.

The next plot shown divides the first and second principal component into two separate heat maps. Each heat map indicates which features contribute are correlated most significantly within each component. Yellow indicates a strong positive correlation while purple indicates a strong negative correlation. In the first principal component the most significant features are the ‘Roll\_std\_pXX’ components as well as the “MFCC\_mean02” components. In the second principal component the most correlated features appear to be “mean”, “FFT\_std\_max”, and “Roll\_mean\_p05”, and “MFCC\_mean02”. Knowing this correlation relationship could provide a framework for identifying the most important features within the model.

The final graph within this section is a heat map which shows the correlation between different features. Dark red indicates that features have a very strong positive correlation while dark blue indicates that there is a strong negative correlation between the features. This heat map provides further insight into which variable are linearly independent and which variables linearly dependent. For example, the ‘Roll\_std\_p60’ and ‘skew’ features are virtually linearly independent. In both of their columns and rows there is no correlation between the variables.