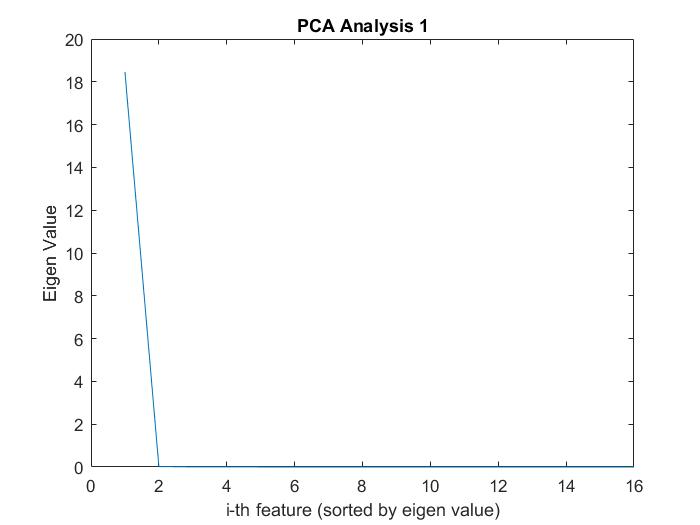
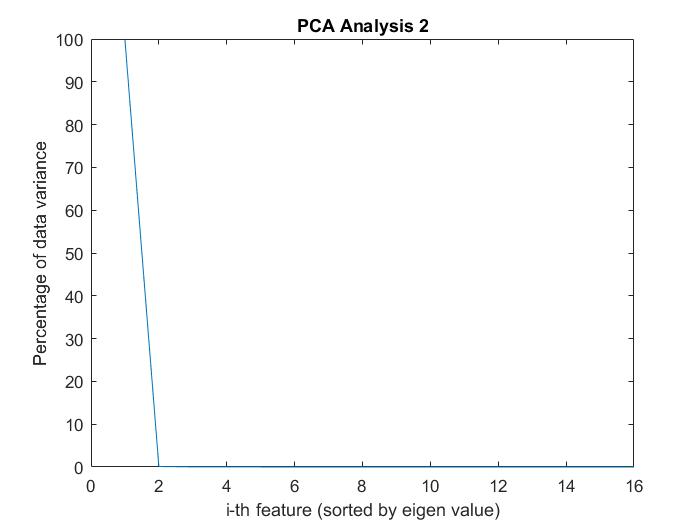
**Data Source & Description**

The dataset for the application project is Parkinsons Telemonitoring Data Set. The dataset was created by Athanasios Tsanas ([tsanasthanasis@gmail.com](mailto:tsanasthanasis@gmail.com)) and Max Little ([littlem@physics.ox.ac.uk](mailto:littlem@physics.ox.ac.uk)) of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals1.

There are 5875 instances in the dataset without any missing data. The actual number of features is 20 and there are two different label associated with each instance. The main aim of the data is to predict the motor and total UPDRS scores (Unified Parkinson Disease Rating Scale) from the 16 voice measures (the other four additional features – subject ID, age, sex and testing time – are not very relevant).

**Raw Data Analysis by PCA**

After preprocessing, the raw data is analyzed by Principal Component Analysis SVD method (refer to the two graphs below) and among the sixteen voice measures, one particular feature explained 99% of the data variance due to the fact that features may have not been normalized.



**Dataset Application**

This dataset measure the extent to which Parkinson affects the voice and hence is very useful. One of the major application is in the Parkinson diagnosis field as the voice measurements can be used to predict UPDRS scores.

**Model Specifics - Linear Regression**

The data is then modeled by linear regression with 10-fold cross validation. The error measure/loss function is the square error and the regularization is L2. To minimize the loss function, the algorithm/method used is stochastic gradient descent.

**Error Measure**

The kfoldLoss is the cross-validation loss of partitioned regression mode. For the model with motor scores as the data label, the error is 60.3566; for the model with total scores as the data label, the error is 104.3331. From my personal understanding of the dataset, the reason why linear regression works better for motor scores than total score is that motor score is better correlated to voice measurements while total score is also affected by some other factors not captured by voice measurements.