Unsupervised Learning
CS 7641 Machine Learning Assignment 3
Siddharth Agarwal, GTID: sagarwal311

Datasets

Covtype dataset:

The first dataset that is used in this assignment for the studying unsupervised learning is same as the second dataset from the first assignment taken from UCI Machine learning repository. The classification problem in the dataset is predicting forest cover type from cartographic variables. The dataset includes four wilderness areas located in Roosevelt National Forest of northern Colorado. Details about the dataset are provided in the README. The dataset was already cleaned and it didn't had any missing values so no preprocessing was needed. The major constraint was the dataset had more than half a million instances so it was difficult to run the experiments fast. So 30K random instances are sampled to perform experiment 1 and 5K instances for experiment 2. The dataset had total 54 features/dimensions and 7 different classes which make it suitable for performing dimensionality reduction experiments as well.

Sensor Dataset:

The second dataset that is used in this assignment is Sensorless drive diagnosis dataset. This dataset is also taken from UCI machine learning repository. The dataset has 58K instances and 49 attributes. Features are extracted from electric current drive signals. The drive has intact and defective components. This results in 11 different classes with different conditions. Each condition has been measured several times by 12 different operating conditions, this means by different speeds, load moments and load forces. Again, since number of instances are large 30K instances are sampled from all the instances to perform the clustering experiment 1 and 5K random instances to perform dimensionality reduction experiment 2.

Experiment 1: Clustering

In this experiment performance of two clustering algorithms: Expected Maximization and KMeans is tested and compared on the above mentioned two datasets.

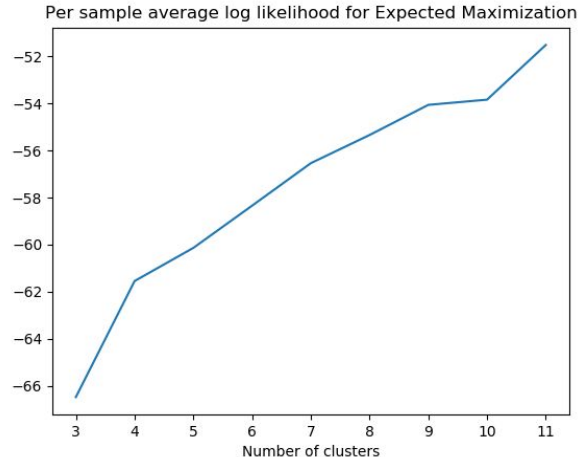
Expected Maximization:

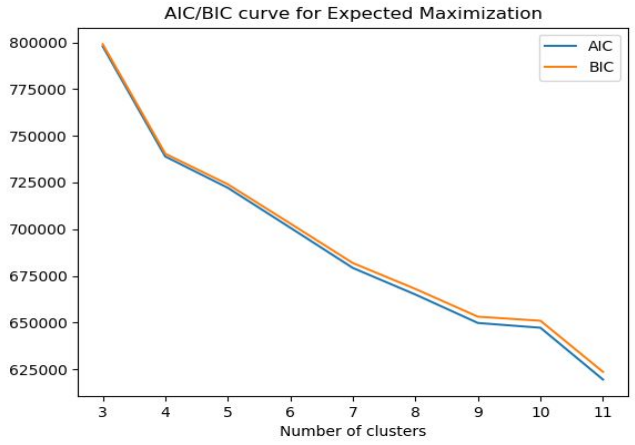
In expected Maximization algorithm mixture of gaussian distribution is fitted on the data for the probabilities. If we use mixture of Gaussian distribution to represent the unlabeled data, the main difficulty arises in separating Gaussian distribution for different sets of points. Expected Maximization algorithm is used to solve this problem by an iterative process in which one first assumes random components (randomly centered on data points, learned from k-means, or even just normally distributed around the origin) and computes for each point a probability of being generated by each component of the model, then the parameters

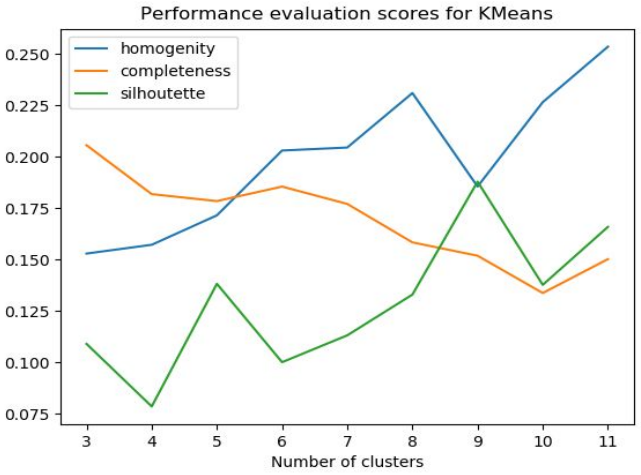
to maximize the likelihood of the data given those assignments. Repeating this process is guaranteed to always converge to a local optimum.

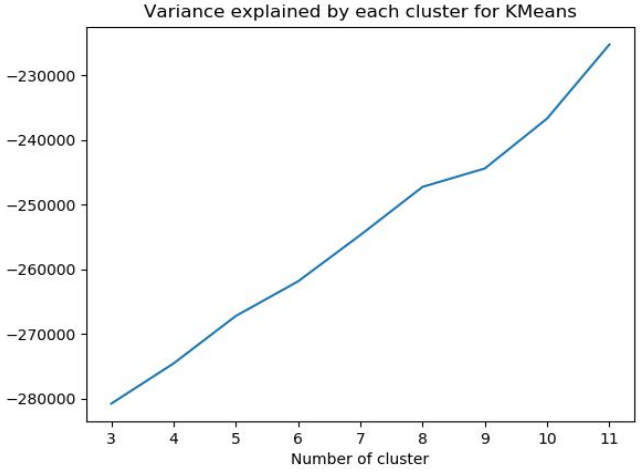
K-Means:

K-Means tries to cluster the data to seperate the samples in K groups of equal variance minimizing the criteria known as inertia or within cluster sum of squares distance. K-Means divides the samples into K cluster each described by the mean of the samples of the cluster and choses these centroids minimizing the inertia. K-Means has three steps: in the first step it initializes the cluster centroids. Then, it assigns each sample to its nearest centroid and finally it assigns new centroid to each cluster based on chosen instances in that cluster. After this step it iterates back to the second step.







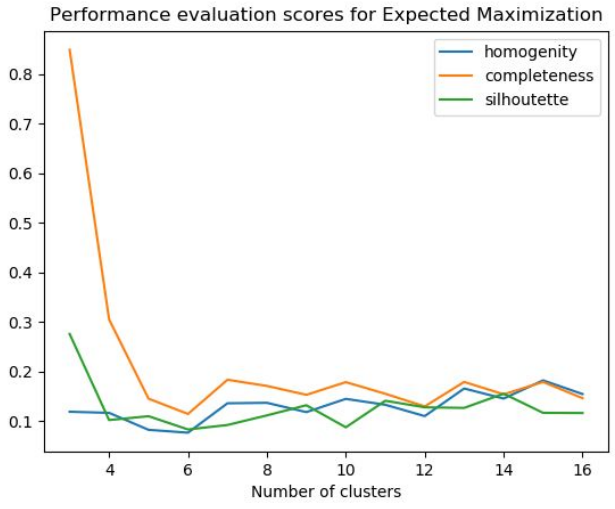


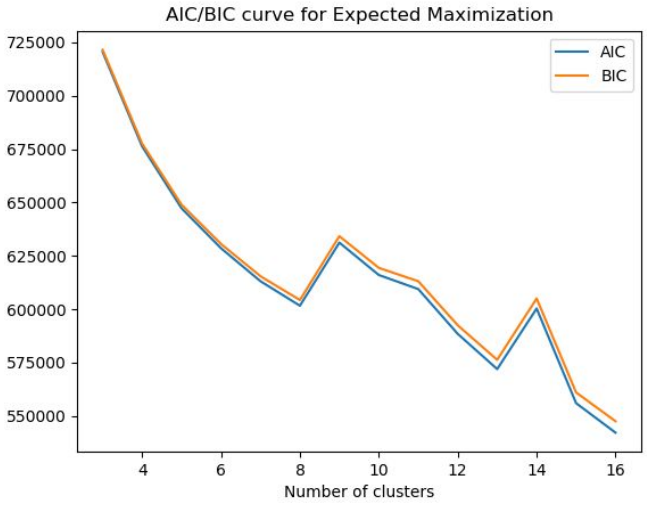
These are the plots for performance evaluation metrics obtained from Covtype dataset for both clustering algorithms with varying number of clusters. Although covtype dataset has 7 classes, the training and testing accuracies using both these clustering algorithms with 7 clusters are extremely low showing that dataset cannot

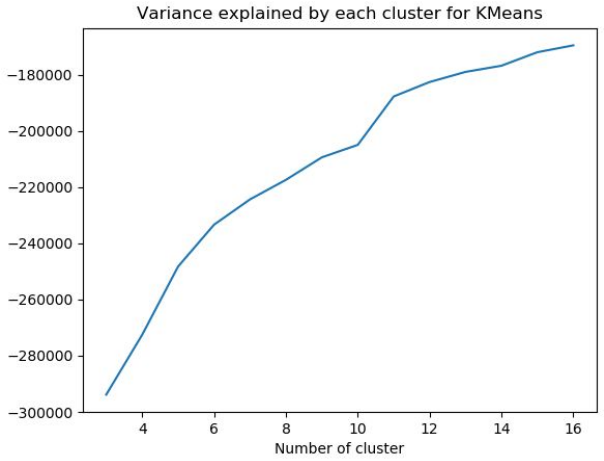
be straightaway segregated into clusters representing different classes in **higher dimensions**. Accuracies are presented in the table below.

We can use **elbow method** to find the optimal number of clusters. This method uses percentage of variance explained or any other performance evaluation metric to find the optimal clusters. Number of clusters are chosen such that adding more cluster doesn't give improvement in performance. Elbow method is not entirely dependable method for finding the number of clusters. As it can be seen that there is a short valley around K = 8 in the variance plot for K-Means. But after that variance again start increasing so, we cannot reliably conclude any optimal K using this plot. Also, for expected maximization desirable metrics homogeneity (high score when each cluster contains only members of a single class), completeness (high score when all members of a given class are assigned to the same cluster) and silhouette (how similar is instance to its own cluster as compared to other neighbouring clusters) stop decreasing at K=4 but after that they again start increasing showing improvement in performance after K=4. **So, it is safe to say that this dataset cannot be represented into clusters (representing each class) with all the attributes taken.**

	Expected Maximization		KMeans	
Dataset	**Training**	**Testing**	**Training**	**Testing**
CovType	5.54	4.33	2.97	2.833
Sensor	9.575	9.4	9.003	9.93

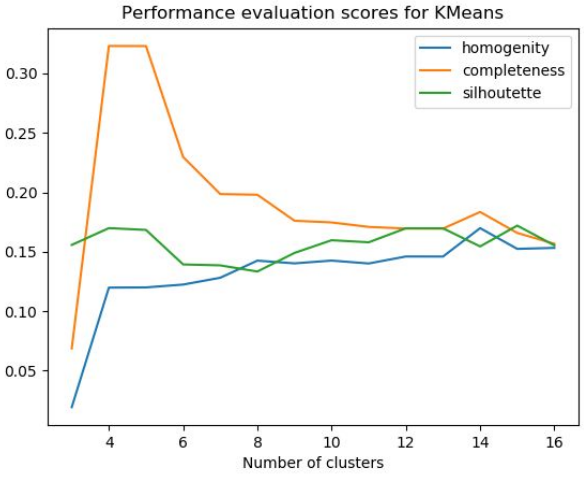


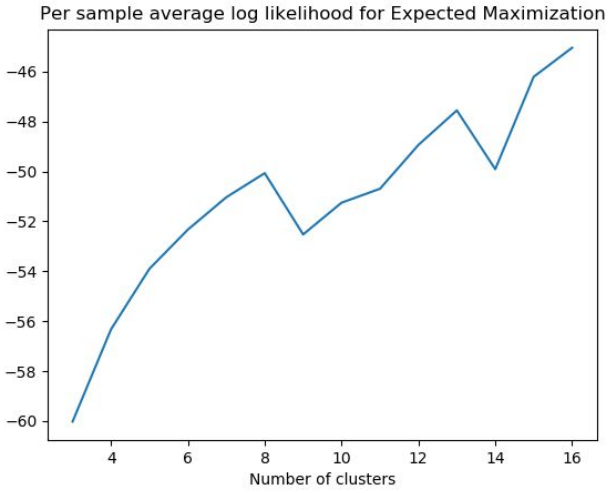




These are the same plots for Sensor data (11 classes). As we can see in the table this data set have higher accuracies if we take total number of clusters to be equal to number of classes. Although this accuracy is not appreciable at all, it is still approximately twice than that of Covtype dataset which tells us that this dataset is more clusterable than previous dataset when each cluster represents one class.

On observing the performance evaluation score for expected maximization, score stops





increasing approximately after 6-8 clusters.

And for K-means, variance is continuously increasing in the whole graph so nothing can be concluded from this plot. But, it can be observed in the performance evaluation plot for K-Means that these metrics are not increasing after K = 6-8. So, **using both the methods, optimal K for this dataset would be approximately between 6-8 clusters in higher dimension.**

Experiment 2: Clustering after dimensionality reduction

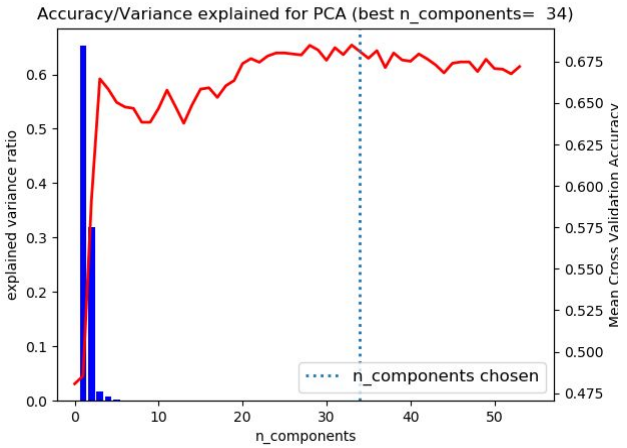
Dimensionality Reduction Algorithms:

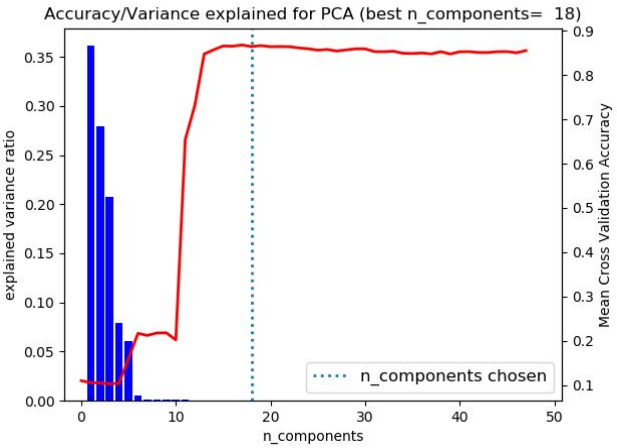
Principal Component Analysis: PCA is used to decrease the dimension of multivariate dataset into set of orthogonal components such that they explain maximum variance. Principal component or the first component has the highest variance and each successive components in turn has highest variance possible under the constraint that they are orthogonal to all preceding component. PCA is performed using eigenvalue decomposition and it is sensitive to initial relative scaling of the attributes.

Independent Component Analysis: ICA separates a multivariate time signal into additive sub components that are maximally independent. Typically, ICA is used to separate signal into independent component instead of reducing the dimension of the dataset.

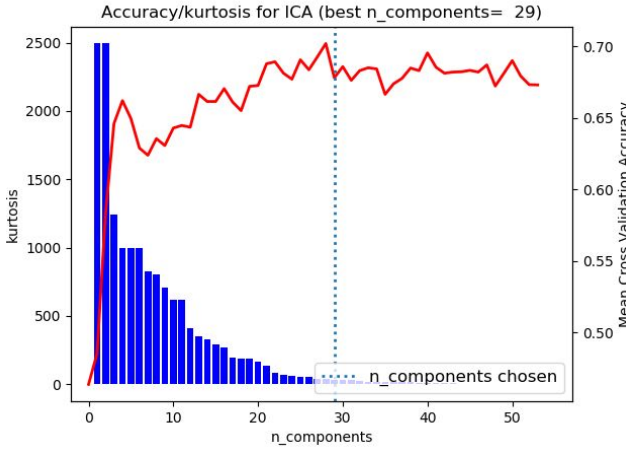
Random Projections:
Gaussian random projection reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from a gaussian distribution with zero mean and variance equal to inverse of number of components. Random projection is used for distance based method and pairwise distance between any two instance of dataset is preserved. This is done by varying dimensions and distributions of random projections matrices.

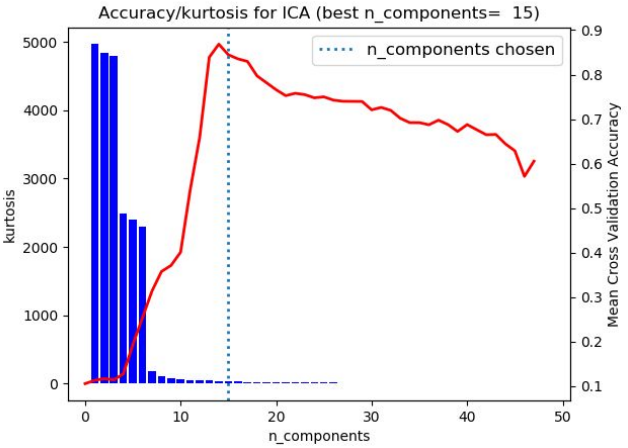
Factor Analysis:
Factor analysis is a linear Generative model with gaussian latent variables performs linear transformation to convert the current set of variables to lower dimensional latent variables with added gaussian noise. Latent factors are distributed according to zero mean and unit variance and noise also has zero mean and arbitrary diagonal covariance.





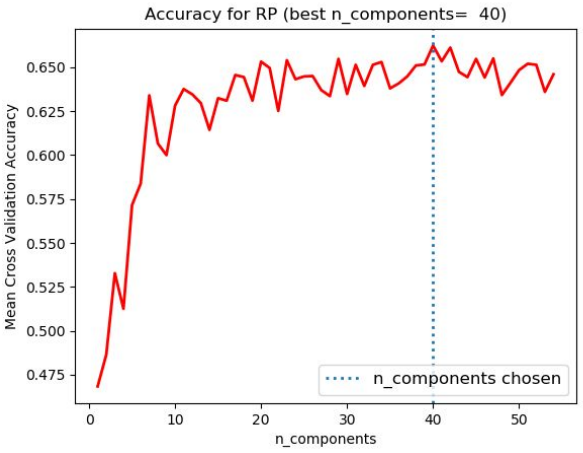
Now the datasets are dimensionally reduced with all the four methods and then clustering is performed using both the clustering methods on the dimensionally reduced dataset.
The above plots shows the accuracy of a decision tree classifier and explained variance after reducing using PCA with number of dimensions varying for both the datasets (Covtype on the left and Sensor on the right). It can be seen that maximum variance is present among only first few orthogonal projections, representing that most of the dimensions are redundant in terms of information and they can be reduced to a smaller more informative feature set. There is not much variation in the accuracy after we increase the dimension greater than the dimension of that more informative reduced feature set except small variation due to noise. Number of components are chosen based on highest accuracy.

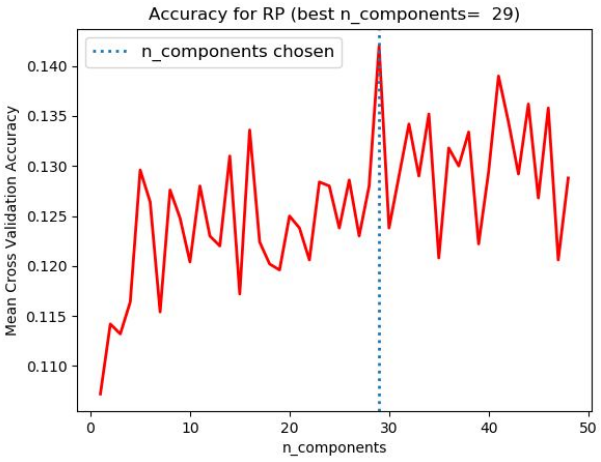




These are the same graphs for both the datasets (covtype on the left and sensor on the right) except that Independent component analysis is used for reducing the dimensions and kurtosis is plotted instead of explained variance. Kurtosis is representation of 4th moment of the data, or the skewness or the measure of 'tailedness' or independence. Again same phenomenon is observed as in PCA very few components are independent in the whole set of attributes. Interestingly, accuracy goes on to decrease for the sensor dataset if we take dimension of the transformed dataset closer to the dimension of the original dataset. This is because these components have zero kurtosis values.

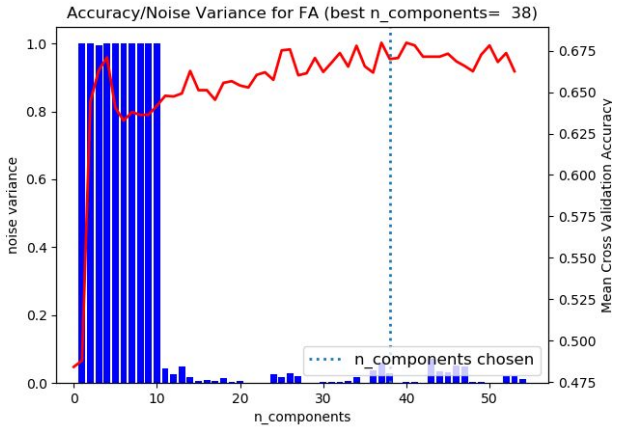
The plots below shows the accuracy for Decision Tree classifier with varying number of components of the transformed dataset obtained using Random Projection algorithm. Accuracy for sensor dataset is extremely low and random with the reduced dataset. While for covtype dataset accuracy is stabilizes

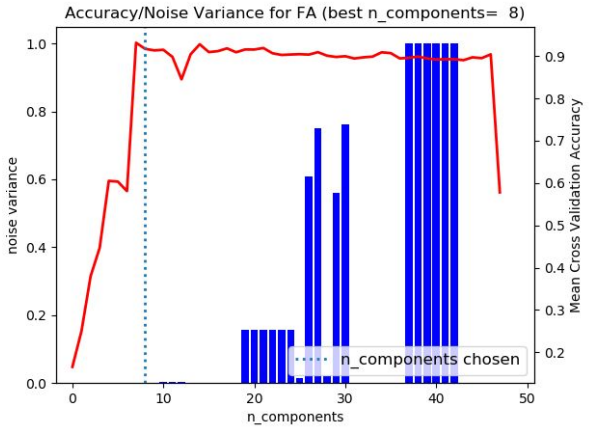




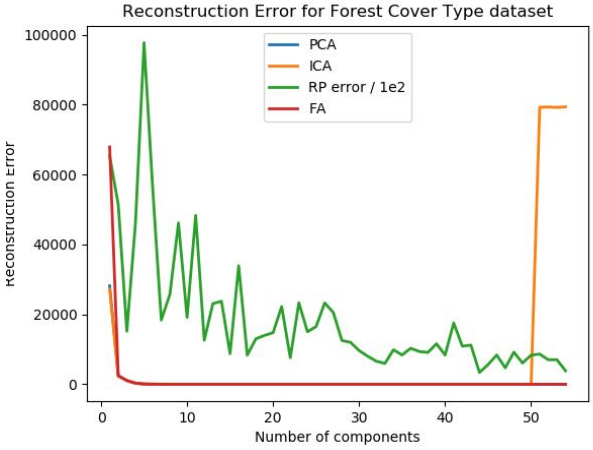
after around 20 components.

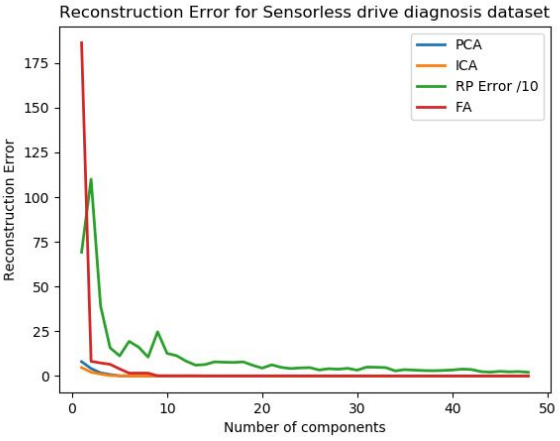
And finally the graphs below are for noise variance of different components for factor analysis method. Noise variation is high for initial components for covtype dataset and high for later dimensions for sensor dataset.



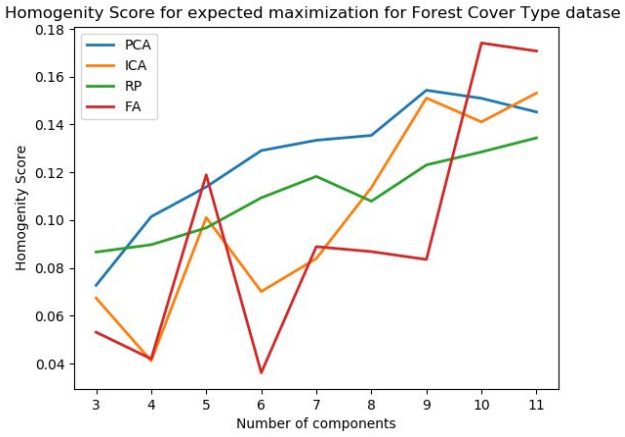


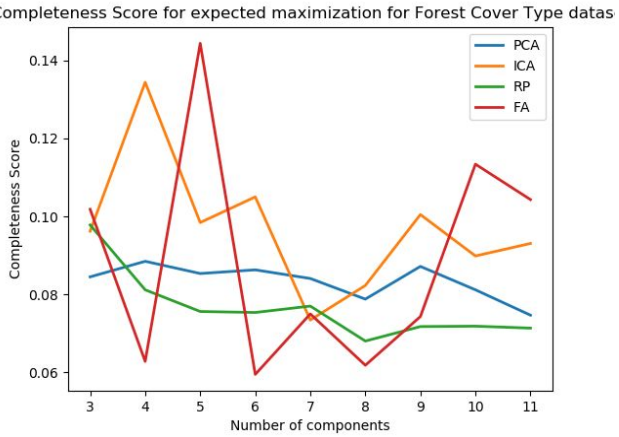
The plots below represents reconstruction error (reconstruction error is the error difference between projection of reduced dataset on components in original signal space and initial original dataset) for covtype and sensor dataset. It can be seen that as more number of components are added the reconstruction error decreases. Since, ICA has zero kurtosis for higher dimensional reduced dataset, reconstruction error suddenly increases.

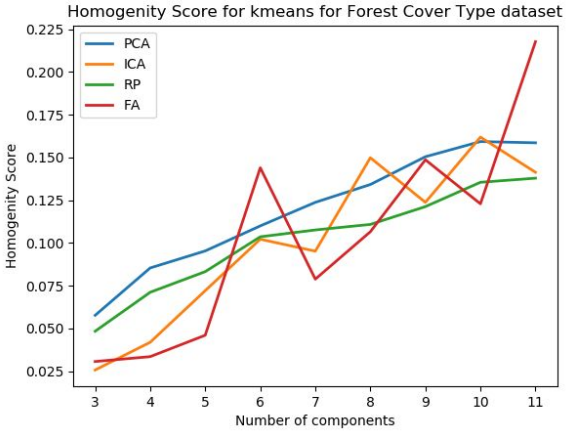


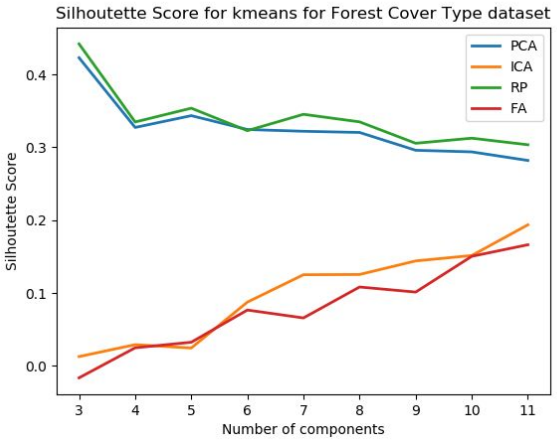


Clustering: Now for each dataset clustering experiments are performed and performance evaluation metrics and accuracies are evaluated. In this experiment number of components for dimensionality reduction is chosen from the optimal number of components calculated based on decision tree accuracy in the previous part of this experiment.





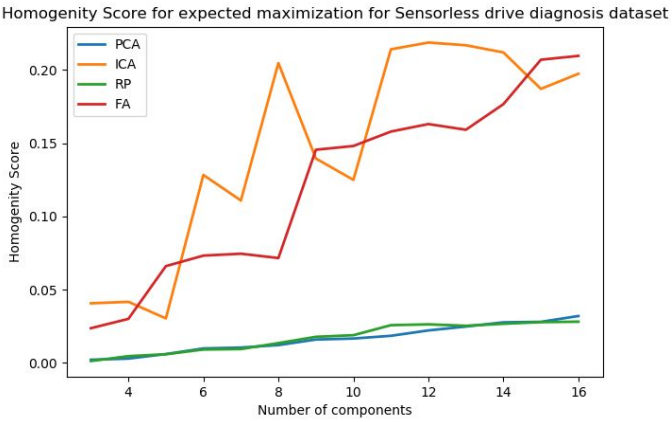


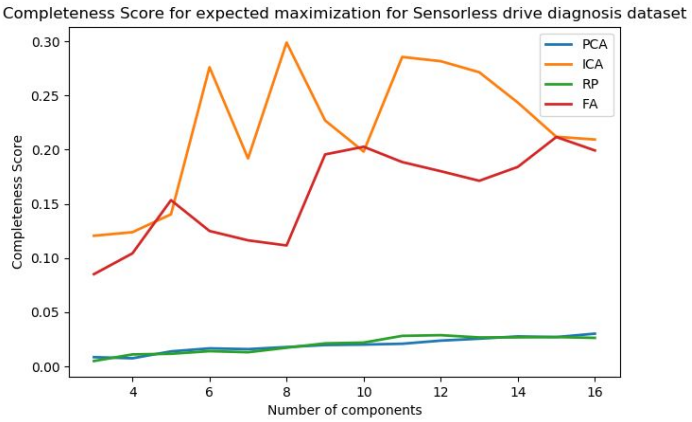


These are the plots for homogeneity, completeness and silhouette score for different clustering algorithms on the reduced forest covertype dataset. Accuracies with number of components equal to actual number of classes in the dataset is presented in the table below. It can be clearly seen that accuracies are much higher than corresponding accuracies in full dimension as in experiment 1 with Factor analysis using KM having the highest accuracy. Thus, it can be said that this dataset is much more clusterable in reduced dimensions than in higher dimensions.

	PCA		ICA		RP		FA	
	EM	**KM**	**EM**	**KM**	**EM**	**KM**	**EM**	**KM**
Covtype	18.525	9.9	24.85	9.325	6.75	14.35	31.4	44.074
Sensor	9.0	8.25	16.5	7.475	8.85	8.674	20.325	18.375

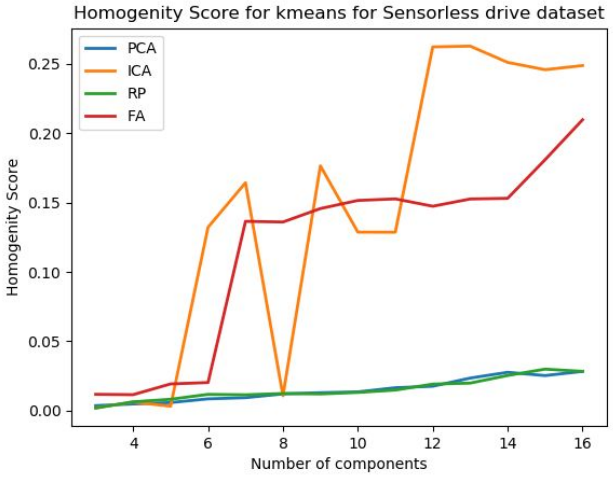
The plots below are the performance evaluation metrics (Homogeneity, Completeness and Silhouette) for four different dimensionality reduction algorithms varying with number of components of clusters for sensorless drive dataset. It can be seen that when accuracies of

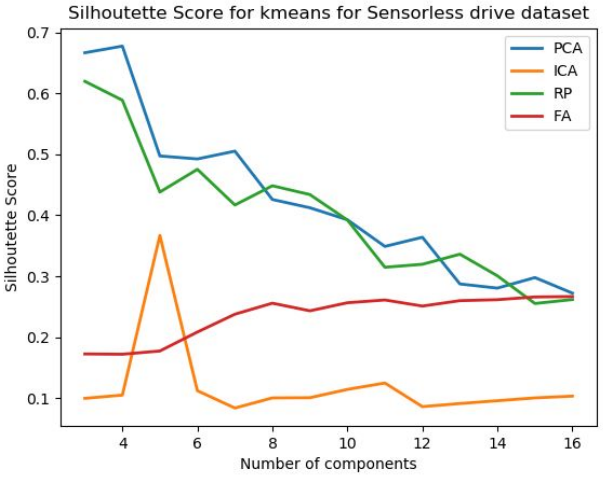




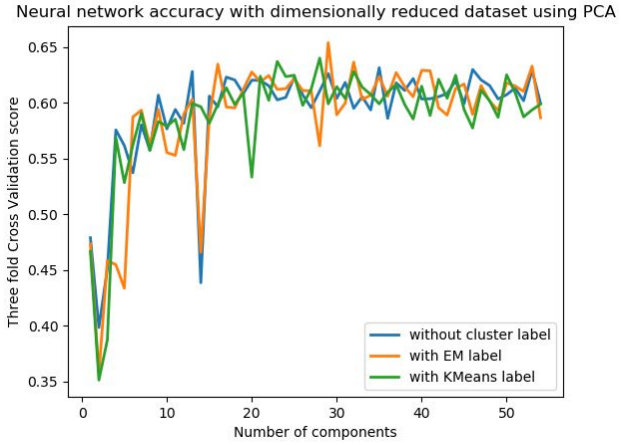
these clusters are compared with accuracies of clusters without dimensionality reduction, these clusters have higher accuracies. But they are less than corresponding accuracies of

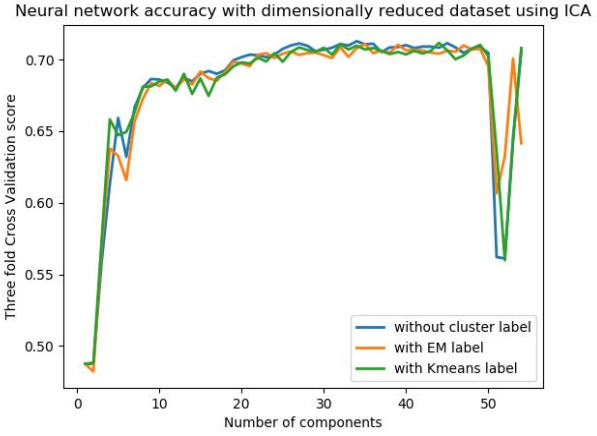
covtype dataset. Thus, it can be said that sensorless drive dataset is more clusterable in lower dimension than higher dimension but less clusterable than covtype dataset in the lower dimension.

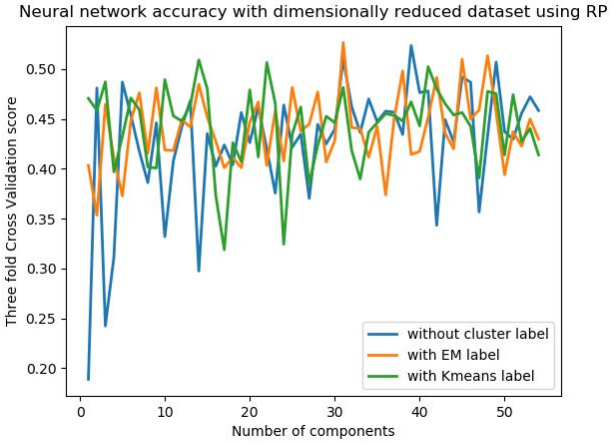


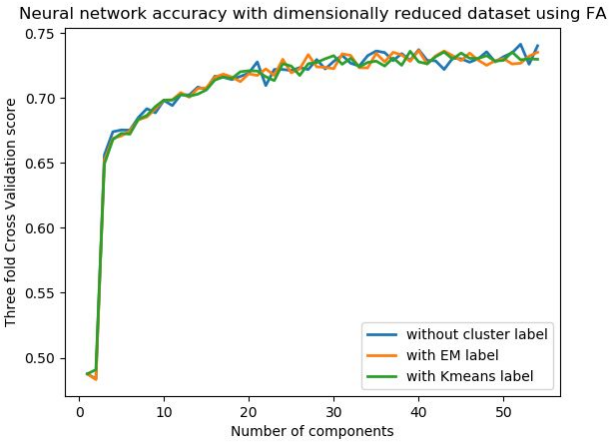


Experiment 3 and 4: Comparing performance of neural network on dimensionality reduced dataset and clustered dataset









In this experiment, the covtype dataset was dimensionally reduced using the four dimensional reduction algorithms. And a neural network with number of hidden layers as the average of number of components and number of classes was fitted and cross validation accuracy was checked. This accuracy is shown in the graph with label as 'without cluster label'. Now, instead of fitting the neural network directly on reduced dataset, it is fitted on appended dataset which has output of cluster labels appended to the reduced dataset. It is done for both Expected maximization and K-means clustering algorithm which are shown in plots with appropriate labels. Since accuracy of the clustering algorithm on the reduced dataset is not exceptionally high there is not much of the difference between accuracies without cluster label and with cluster label although mean accuracy is slightly higher. But if clustering accuracy on the reduced dataset would have been higher then since we are adding the labels of the cluster (i.e. the actual output class), neural network accuracy would have been much higher. If we run the neural network directly on the normal dataset appended with cluster labels then time taken of neural network to run as compared to time taken to run on dimensionally reduced dataset would be much higher since this dataset have all the features while dimensionally reduced dataset only few features.

Observations and Conclusions:

1. Clusters for Covtype dataset have very low accuracy with most of the accuracies less than 5%. This represents that this dataset is not much clusterable in higher full dimensions.
2. Since, this dataset is not clusterable in high full dimension, there is no elbow visible in any of the performance evaluation metrics. Thus, elbow method is not completely reliable method for finding the optimal K. Generally for finding this K, domain knowledge is used.
3. Sensor dataset had twice the accuracy for clusters in full dimensions as compared to corresponding accuracies for Covtype dataset.
4. Slight elbow was visible for approximately K in between 6 - 8.
5. For dimensionality reduction using PCA most of the variance is explained by first few components, and accuracy of the Decision Tree algorithm does not vary much after these few components except small variations due to noise.
6. Similarly, for other methods most of the variation or information is present in the first few components.
7. For random projections, decision tree accuracies for sensor dataset is low and random.
8. Reconstruction error for random projections is extremely high. Thus, dataset is not distance separable.
9. Reconstruction error decreases rapidly after first few components since most of the information was contained in the first few components.
10. Cluster accuracies are much higher for reduced dataset. Sensor dataset have lower accuracies than covtype dataset in reduced dimensions. Thus, Covtype dataset is more cluster separable than Sensor dataset in reduced dimensions.
11. Neural network accuracies with EM and K-Means labels is slightly higher when cluster labels are appended. Time taken to train neural network with less dimensions is less than time taken to train on full dataset.