1. The problems/data from assignment #1

1.1 Wine Recognition (Data link)

The data of Wine Recognition is the results of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in the three types of wine.

With 13 different statistics, we would like to predict which type a glass of wine belong to.

The data contains 13 numeric features and each data point belongs to one of 3 classes.

Following features are included:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

1.2 Breast Cancer Wisconsin (Diagnostic) (Data link)

To diagnosis Breast Cancer with machine learning algorithm is a promising trend. A good machine learning model would help doctors to diagnosis Breast Cancer much faster and more accurate.

In the data, features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The data contains 30 numeric features and each data point belongs to one of two types of breast cancer (i.e., WDBC-Malignant and WDBC-Benign).

Following features are included:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension (“coastline approximation” - 1)

1.3. Metrics

For this project, since we have two multi-class classification problems, we will be using the overall accuracy as the main performance metrics.

Overall accuracy can be calculated as # of correct prediction / # of total prediction

For clustering algorithms, we measure their performance with three scores.

1. homogeneity_score: A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.

2. silhouette_score: The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b).

3. completeness_score: A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.

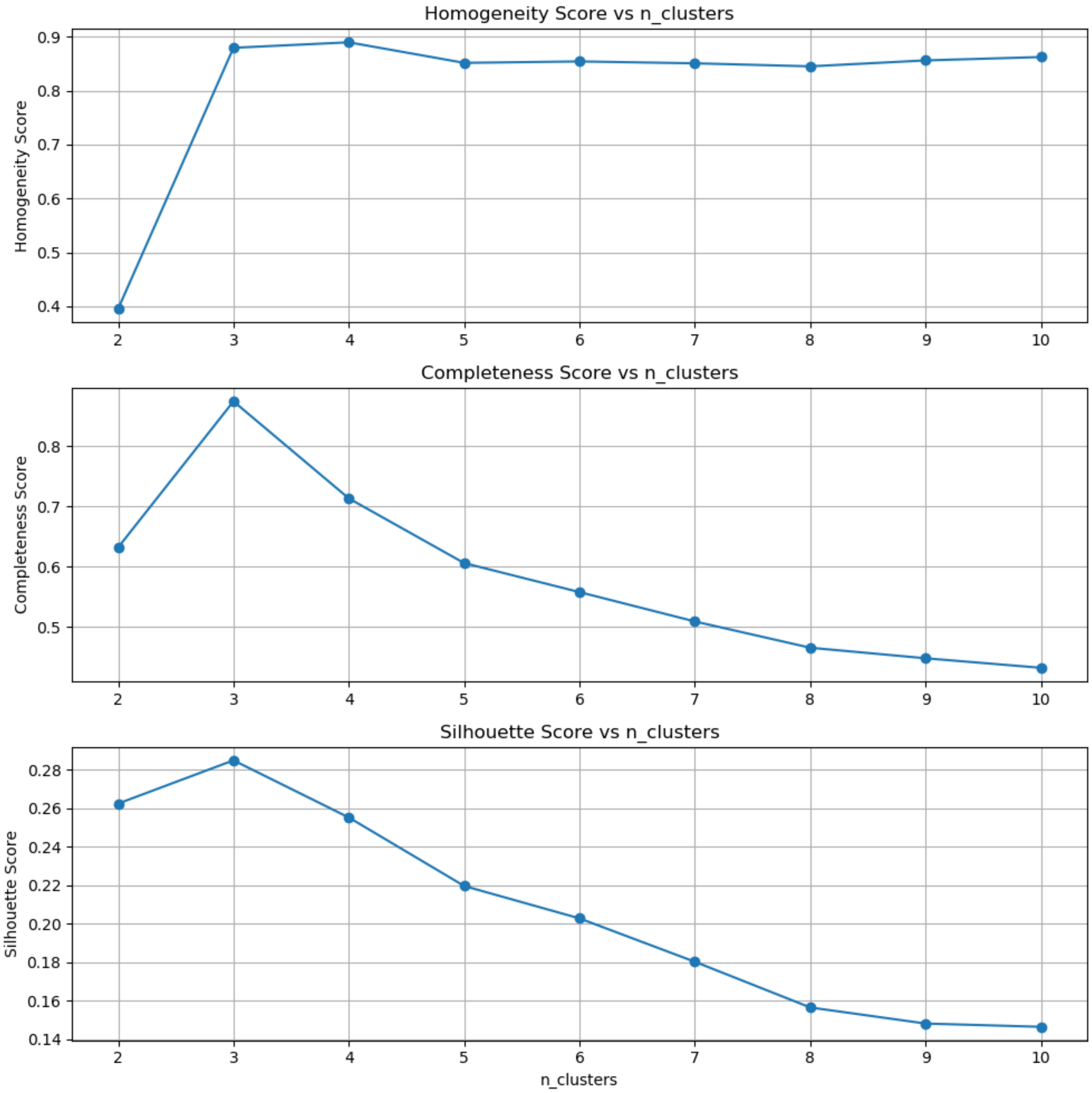
2. Clustering

2.1 Wine Recognition

We have preprocessed the data with standardization before clustering. And at each point in the plots we have rerun the algo 10 times with different initial states and calculated the average.

2.1.1 Kmeans

Blow is the performance plot against various number of clusters.

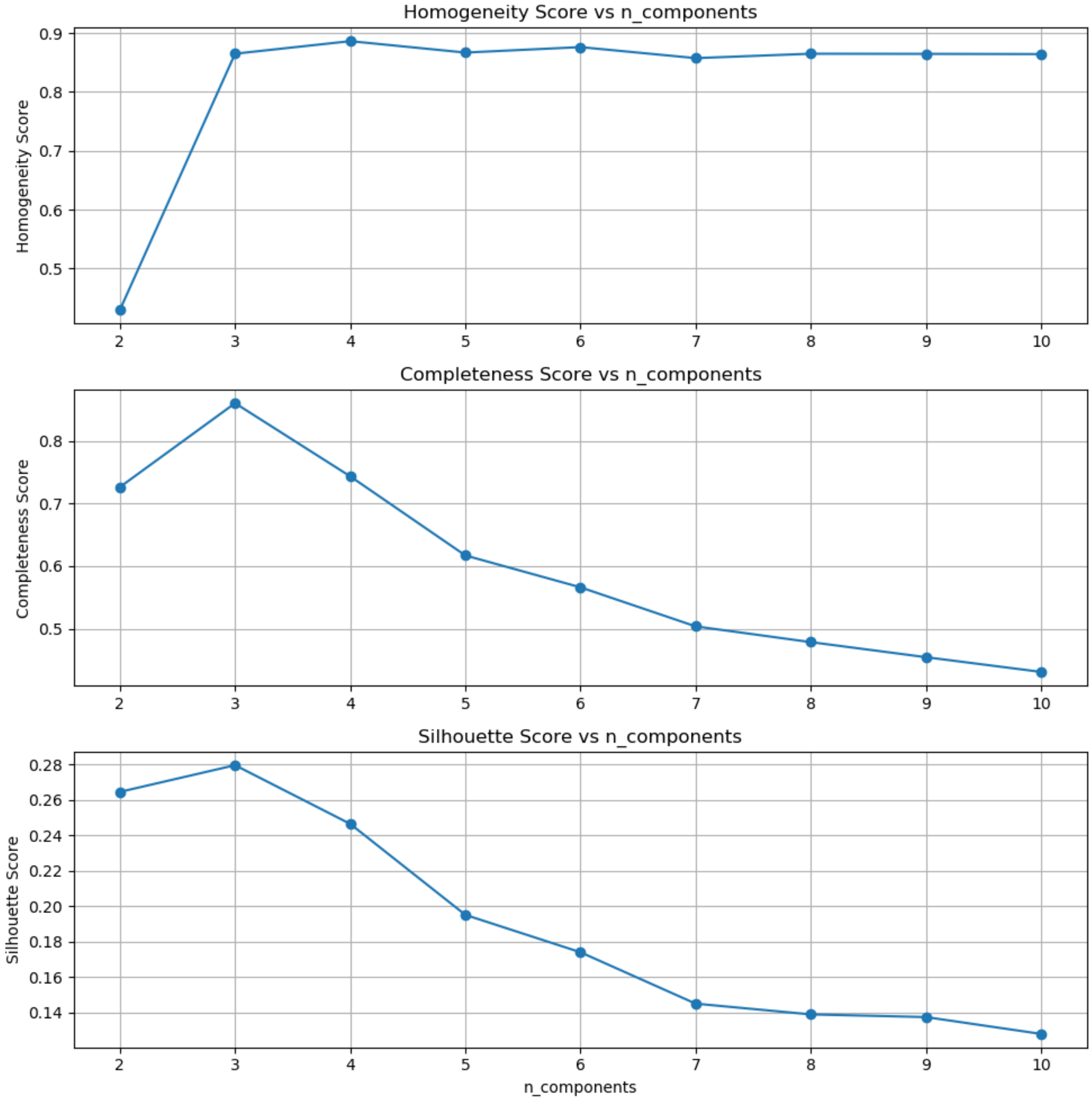


As we can see, kmeans perform best when we choose the # of clusters to be 3. It is as expected since the ground truth for this data set contains 3 labels/classes.

2.1.2 EM

Here we used Gaussian Mixture model which is trained by EM. We have set each component has its own general covariance matrix, and the max number of EM iterations to perform is 10000.

We test various number of Gaussian component, below is the performance plot.



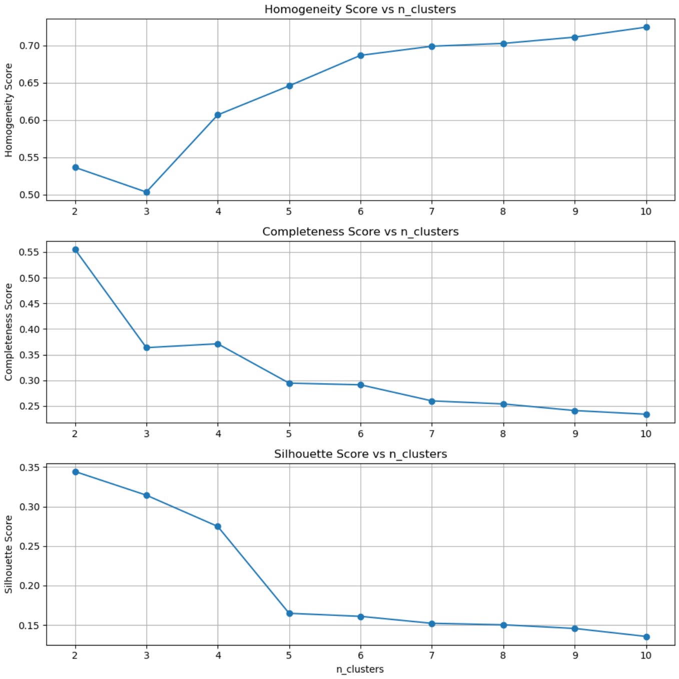
As we can see, EM perform best when we choose the # of components to be 3. It is as expected since the ground truth for this data set contains 3 labels/classes.

There is not much difference between kmeans and EM, both are able to correctly label around 90% data when we set n_cluster/n_component to 3.

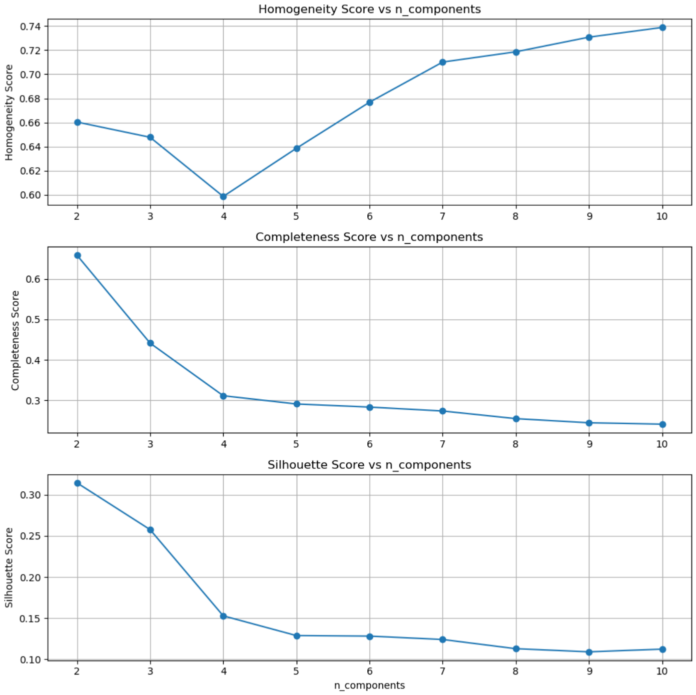
2.2 Breast Cancer Wisconsin

Here the ground truth is 2 labels/classes.

2.2.1 Kmeans



2.2.2 EM



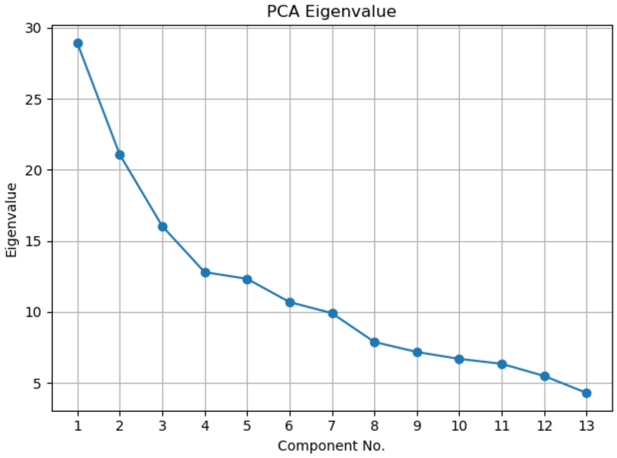
There is not much difference between kmeans and EM, both will perform great when we set n_cluster/n_component to 2.

2. Dimensionality reduction

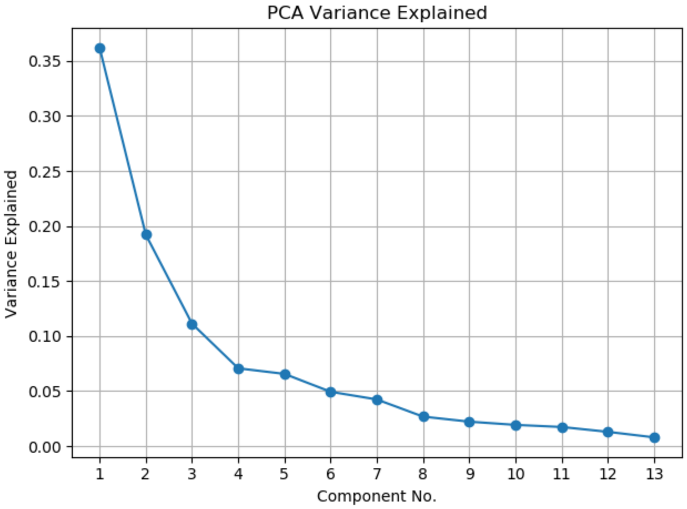
2.1 Wine Recognition

2.1.1 PCA

Below is the distribution of PCA eigenvalues.



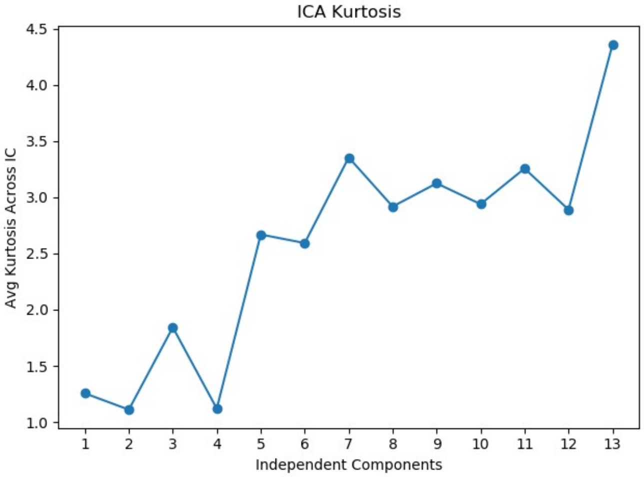
Below is the percentage of variance explained by each component.



Hence, if we need to contain 90% of variance from original data, we choose first 8 components.

2.1.2 ICA

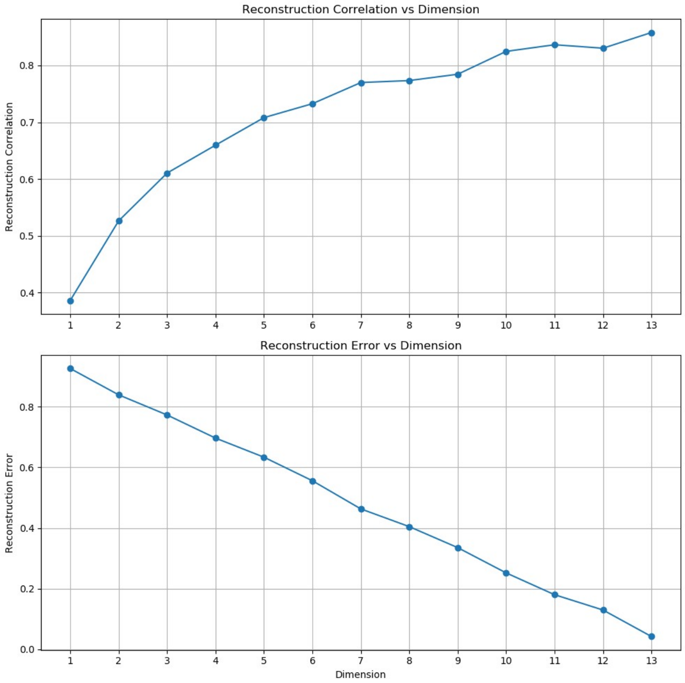
Below is the kurtosis against the number of independent components we use for ICA.



As we can see, the more components we include in ICA, the larger kurtosis we can get. With 13 dimensions, we can get the maximum kurtosis around 4.4.

2.1.3 Randomized Projections

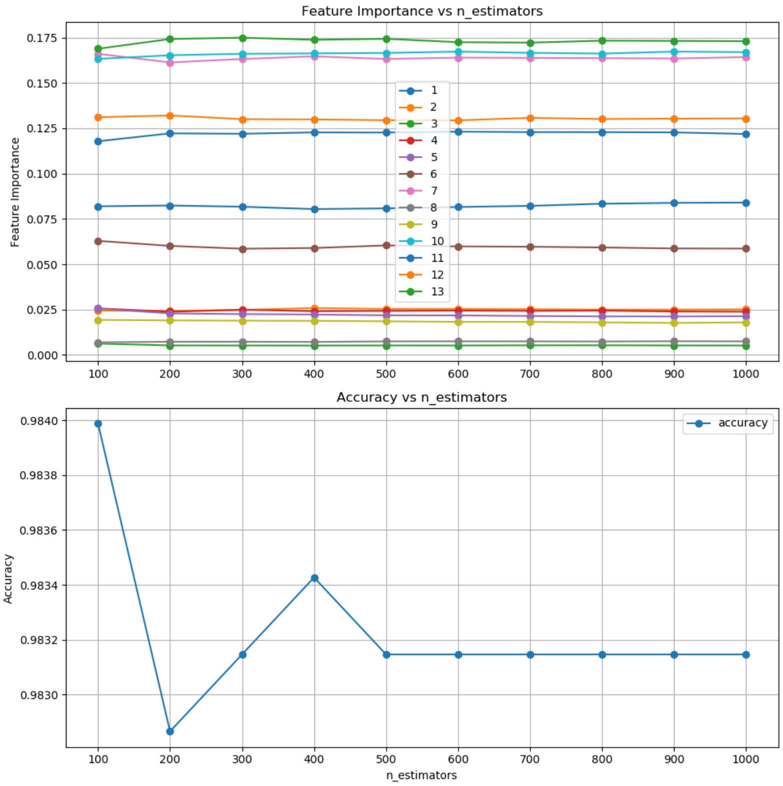
Below is the reconstruction correlation and error against # of components in randomized projection.



As we can see, the more components we include in RP, the larger correlation (smaller error) we can get. With all 13 dimensions, we can get the maximum correlation around 0.9 and error around 0.05.

2.1.4 Random Forest

Here we use random forest to calculate importance of each feature and only keeps the most important ones. Hence it can be viewed as a dimension reduction algo.

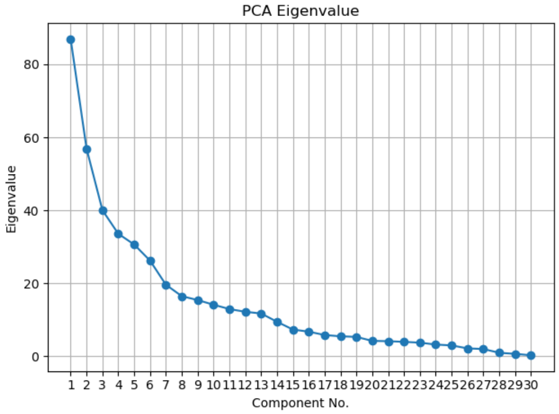


As we can see, RF model achieved 98% accuracy and selected feature 3, 7, 10 to be most important.

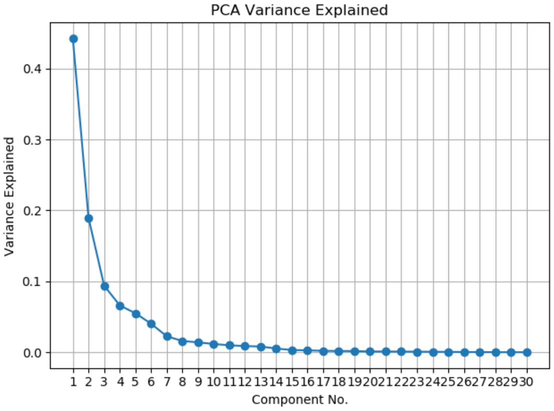
2.2 Breast Cancer Wisconsin

2.2.1 PCA

Below is the distribution of PCA eigenvalues.



Below is the percentage of variance explained by each component.

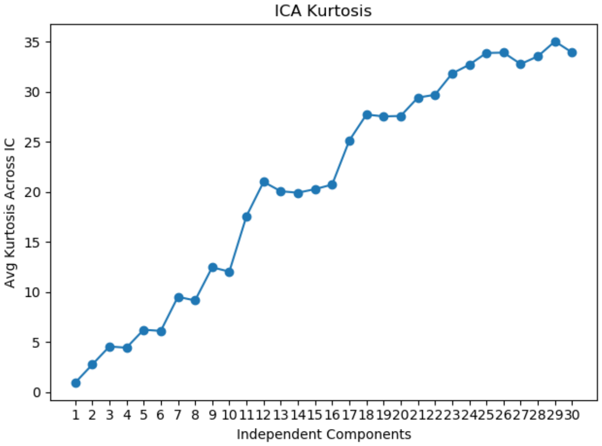


Hence, if we need to contain 90% of variance from original data, we choose first 7 components.

2.2.2 ICA

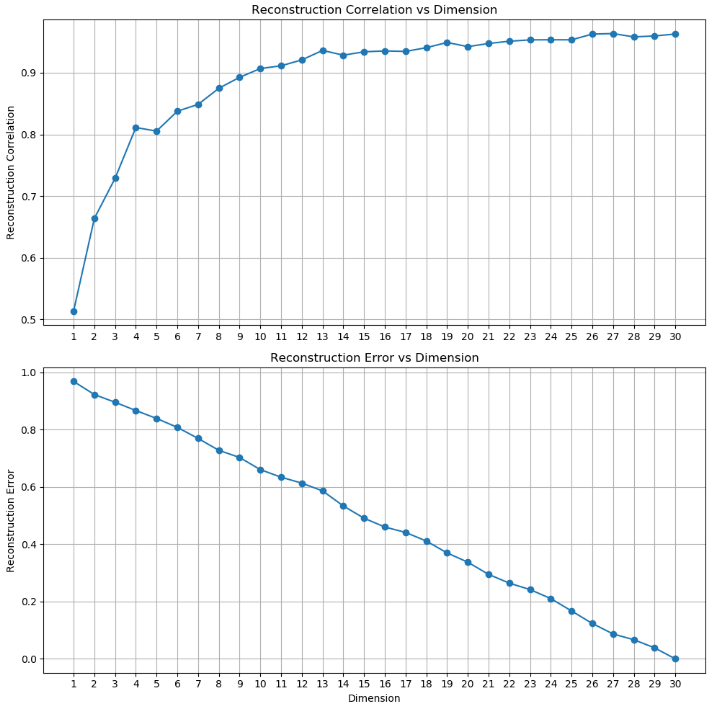
Below is the kurtosis against the number of independent components we use for ICA.

As we can see, the more components we include in ICA, the larger kurtosis we can get. With 29 dimensions, we can get the maximum kurtosis around 35.



2.2.3 Randomized Projections

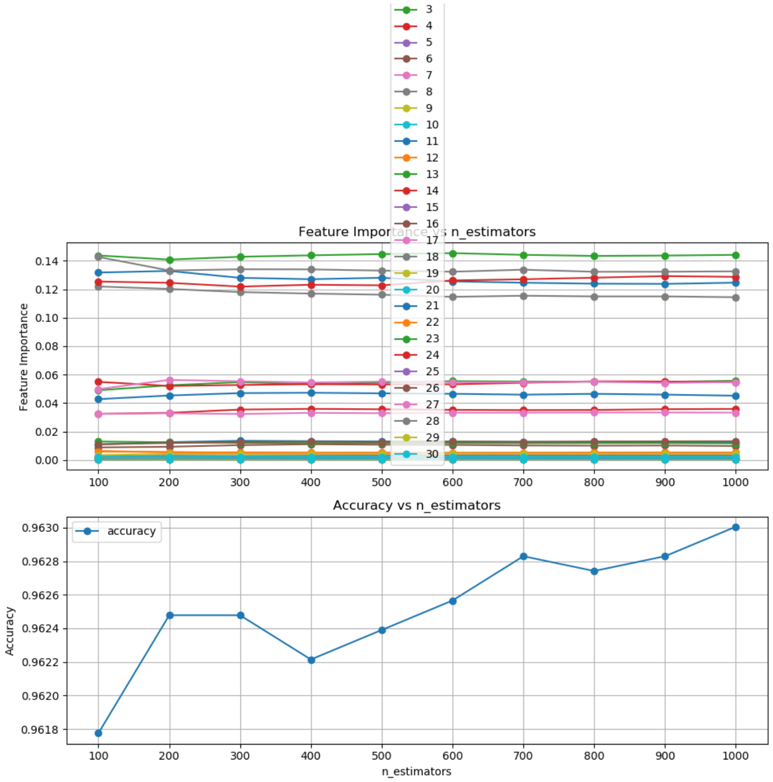
Below is the reconstruction correlation and error against # of components in randomized projection.



As we can see, the more components we include in RP, the larger correlation (smaller error) we can get. With around 20 dimensions, we can get the maximum correlation around 0.95 and error 0.3.

2.2.4 Random Forest

Here we use random forest to calculate importance of each feature and only keeps the most important ones. Hence it can be viewed as a dimension reduction algo.



RF model achieved 96% accuracy and selected feature 3, 8, 11, 14 to be most important.

3. Clustering on dimension reduced data

3.1 Wine Recognition

Here we first applied dimension reduction methods, then do the clustering. Below is the comparison on various performance metrics for clustering.

	kmeans	kmeans & PCA	kmeans & ICA	kmeans & RP	kmeans & RF
homogeneity_score	0.879	0.879	0.892	0.519	0.704
silhouette_score	0.873	0.875	0.883	0.515	0.702
completeness_score	0.285	0.284	0.276	0.189	0.264

	EM	EM & PCA	EM & ICA	EM & RP	EM & RF
homogeneity_score	0.864	0.957	0.501	0.533	0.712
silhouette_score	0.858	0.953	0.532	0.528	0.711
completeness_score	0.284	0.283	0.155	0.184	0.251

PCA tends to improve the performance of both kmeans and EM. The reason is that in PCA we preserved 90% variance in the data while reducing dimension from 13 to 8.

Randomized Projection is not able to retain most information and variance in the original data, hence tend to make the clustering performance even worse.

ICA and RF are not consistent in improving clustering performance. In some cases, it will be better to not use them.

3.2 Breast Cancer Wisconsin

Here we first applied dimension reduction methods, then do the clustering. Below is the comparison on various metrics.

	kmeans	kmeans & PCA	kmeans & ICA	kmeans & RP	kmeans & RF
homogeneity_score	0.525	0.544	0.283	0.292	0.535
silhouette_score	0.540	0.565	0.343	0.344	0.581
completeness_score	0.343	0.345	0.280	0.335	0.348

	EM	EM & PCA	EM & ICA	EM & RP	EM & RF
homogeneity_score	0.662	0.679	0.567	0.077	0.519
silhouette_score	0.660	0.671	0.557	0.084	0.519
completeness_score	0.314	0.388	0.275	0.258	0.303

Again, PCA tends to improve the performance of both kmeans and EM. The reason is that in PCA we preserved 90% variance in the data while reducing dimension from 30 to 7.

Randomized Projection is not able to retain most information and variance in the original data, hence tend to make the clustering performance worse.

ICA and RF are not consistent in improving clustering performance.

4. Neural Network with Dimension reduction

Here we applied all four Dimension reduction algos to the data with best hyper parameter as mentioned above and then trained the same neural networks as before.

Here we showed the training data dimension, training time and train/test accuracy.

4.1 Wine Recognition

	Original Data	PCA	ICA	RP	RF
Training Data Dimension	13	8	10	10	10
Training Time (s)	159.8	9.4	37.7	13.5	15.1
Train Accuracy	80.50%	97.00%	95.00%	96.00%	97.00%
Test Accuracy	73.48%	98.00%	90.00%	94.00%	98.00%

4.2 Breast Cancer Wisconsin

	Original Data	PCA	ICA	RP	RF
Training Data Dimension	30	7	10	10	10
Training Time (s)	150.4	41.4	40.1	43.2	31
Train Accuracy	68.00%	96.00%	86.00%	94.00%	96.00%
Test Accuracy	67.00%	95.00%	80.00%	93.00%	95.00%

As we can see, dimension reduction greatly improved training time, it makes sense since the data is less sparse and gradient will converge much faster.

Also, with dimension reduction, the neural networks tend to perform better, and the training process will be much shorter and more efficient.Among the four methods, PCA and RF are better, both method can reduce the dimension and pick the most significant feature while retaining most variance and important information in data.

5. Neural Network with Clustering

Here we applied both clustering algos to the data with best hyper parameter as mentioned above, added the output cluster as a new feature to the data, then trained the same neural networks as before.

5.1 Wine Recognition

	Original Data	Kmeans	EM
Training Data Dimension	13	14	14
Training Time (s)	159.8	16.4	15.7
Train Accuracy	80.50%	97.00%	97.00%
Test Accuracy	73.48%	96.00%	96.00%

5.2 Breast Cancer Wisconsin

	Original Data	Kmeans	EM
Training Data Dimension	30	31	31
Training Time (s)	150.4	37.7	39.1
Train Accuracy	80.50%	97.00%	97.00%
Test Accuracy	73.48%	95.00%	96.00%

As we can see, clustering greatly improved training time, it makes sense since we add a significant feature to the training data.

Also, with cluster as new feature, the neural networks tend to perform better, and the training process will be much shorter and more efficient.

Both Kmeans and EM can equally do a great job at improving NN performance.