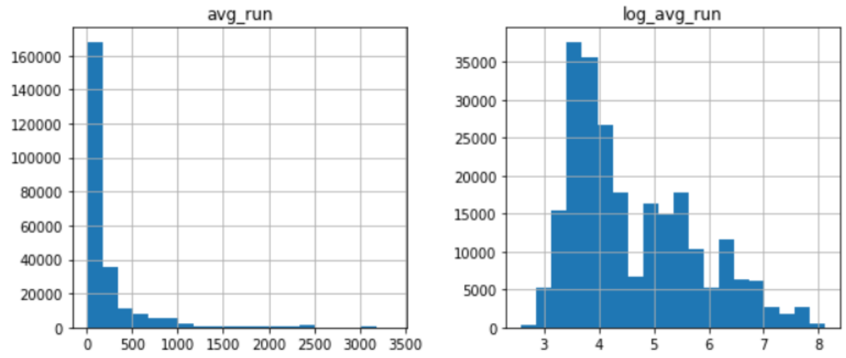
Dimensionality Reduction with K-means, ANN and EM

Data Description & Preparation

Dataset 1 - SGEMM GPU kernel performance Data

This data set measures the running time of a matrix-matrix product A*B = C, where all matrices have size 2048 x 2048, using a parameterizable SGEMM GPU kernel with 241600 possible parameter combinations. For each tested combination, 4 runs were performed.

- The dataset has 241600 observations with 18 features.
- None of the column contains any missing value.
- The target variable will be the average of the four run times.
- After computing the average, the data in the target variable is right skewed. Taking the log of the values reduces the skewness of the target variable.



- For this report, we will use all the independent variables to train our models.
- The data is divided into train test datasets in 70:30 ratio. The data is then scaled appropriately.

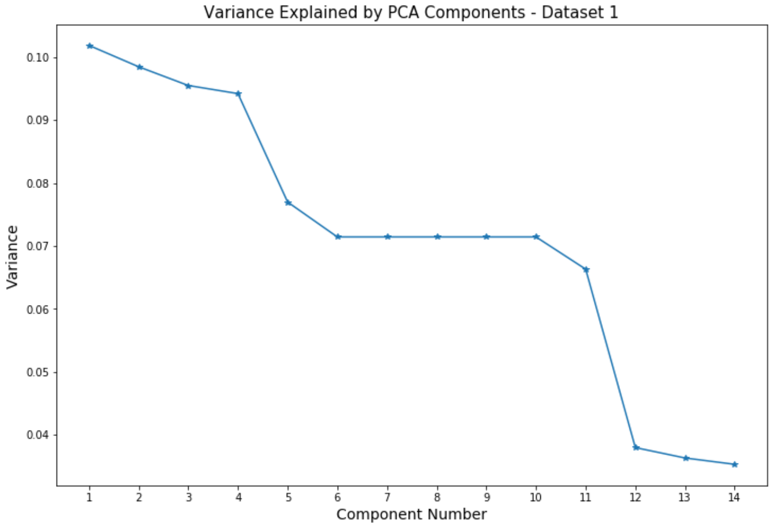
Dataset 2 - Letter Recognition Data

The objective is to identify each of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images are based on 20 different fonts and each letter within these 20 fonts is randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus is converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

- The dataset has 20000 observations with 17 features.
- None of the column contains any missing value.
- The target variable will be column “Letter”. The rest of the 16 variables, of datatype Integer, will be used as predictors.
- For this report, we will use all the independent variables to train our models.
- The data is divided into train test datasets in 80:20 ratio. The data is then scaled appropriately.

Dataset 1

Dimensionality Reduction

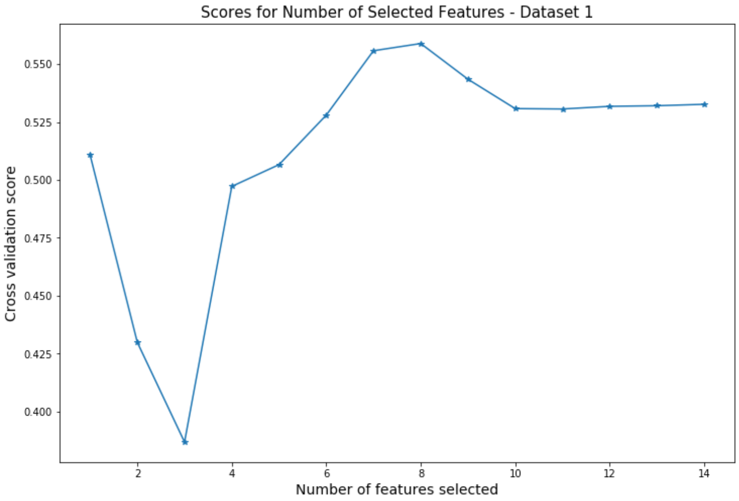


PCA - Principal component analysis transforms the features into a new domain such that the features are ordered in decreasing order of variance and all of them are orthogonal to each other.

82.4% of the variance is explained by the first 10 components. Therefore, we select only 10 features with the PCA.

ICA – Independent component analysis assumes that the subcomponents are non-Gaussian signals and that they are statistically independent from each other. It attempts to decompose a multivariate signal into independent non-Gaussian signals. Here we transform the features into 10 new features in the new domain.

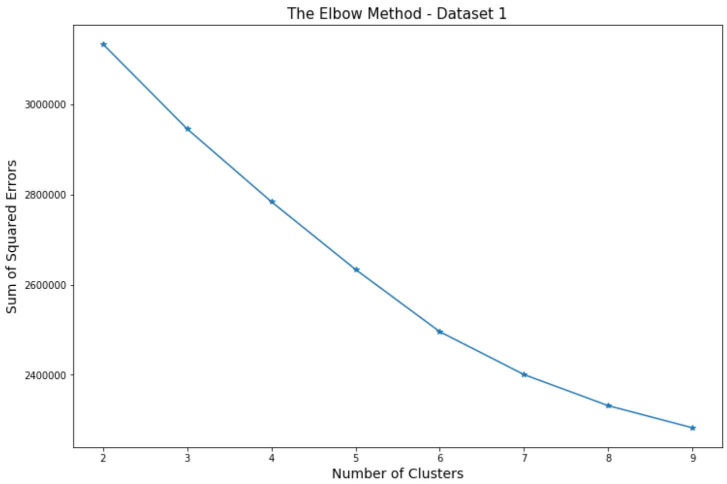
Random Projections – This is like PCA, but instead of choosing the dimensions in the order of variance, it chooses random dimensions and projects the features into those dimensions. Here we transform the features into 10 new features in the new domain using random projections.

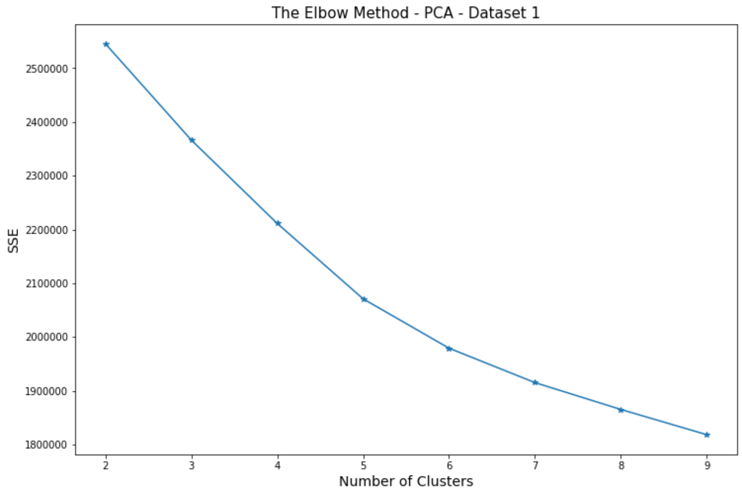


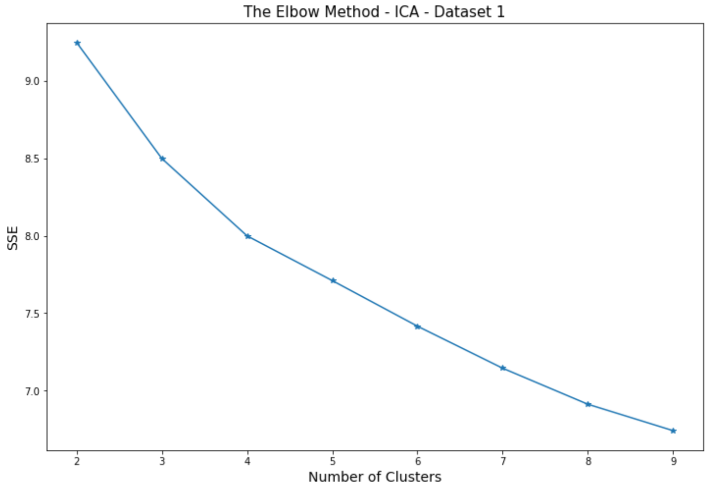
Feature Selection – We have done feature selection using a Decision Tree classifier with criterion 'gini' and maximum depth 15. We got 8 optimal features, which were at the index 0,1,3,4,8,10,12 and 13.

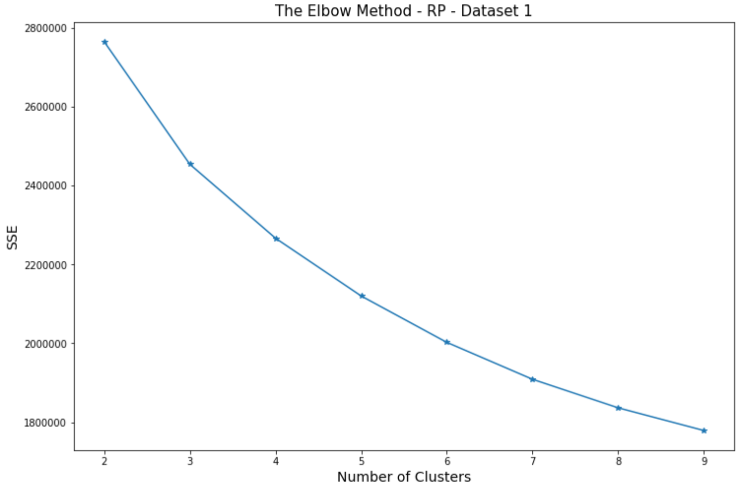
In the graph 8 features have the highest cross validation score.

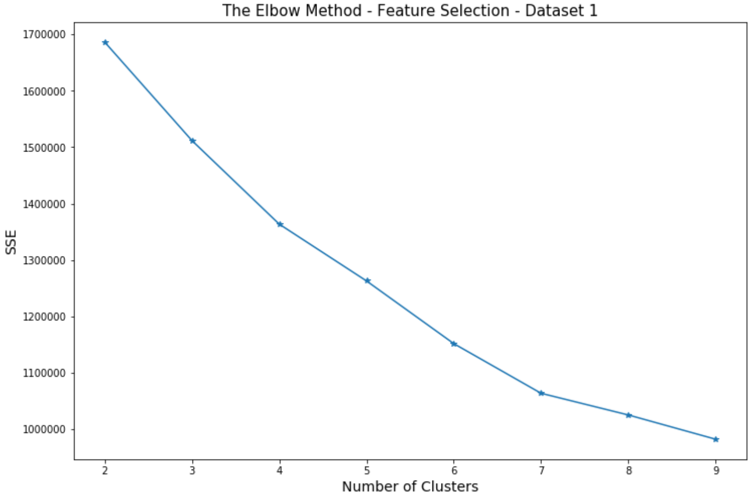
K-means











We choose the optimal number of K using these elbow graphs. We choose the k where we observe an elbow in the graph. From all the graphs we get that the optimal number of clusters from the data is 7.

From the table below, we observe that the features generated with ICA gives the smallest sum of squared errors.

The scores shown in the table below are clustering performance evaluation scores:

Adjusted Rand Index: It is a function that measures the similarity of the two assignments, ignoring permutations and with chance normalization. The bounded range is [-1, 1]. Negative values are bad, similar clusterings have a positive ARI, 1.0 is the perfect match score.

Adjusted mutual info score: Mutual Information is a function that measures the agreement of the two assignments, ignoring permutations. Values close to zero indicate two label assignments that are largely independent, while values close to one indicate significant agreement.

Homogeneity: each cluster contains only members of a single class.

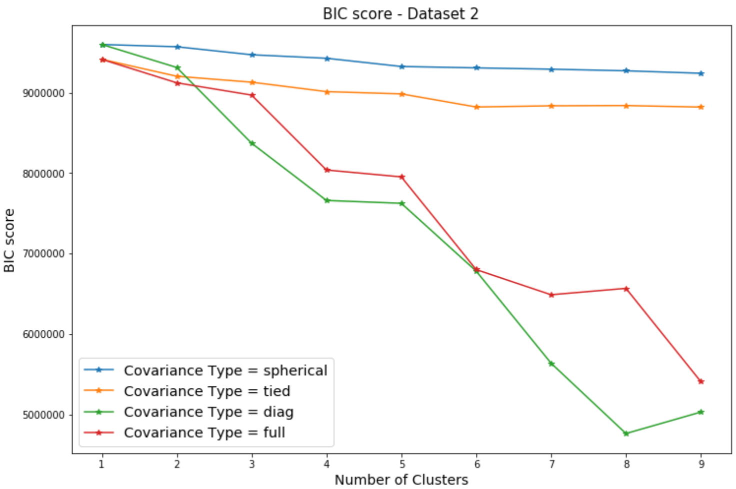
Completeness: all members of a given class are assigned to the same cluster.

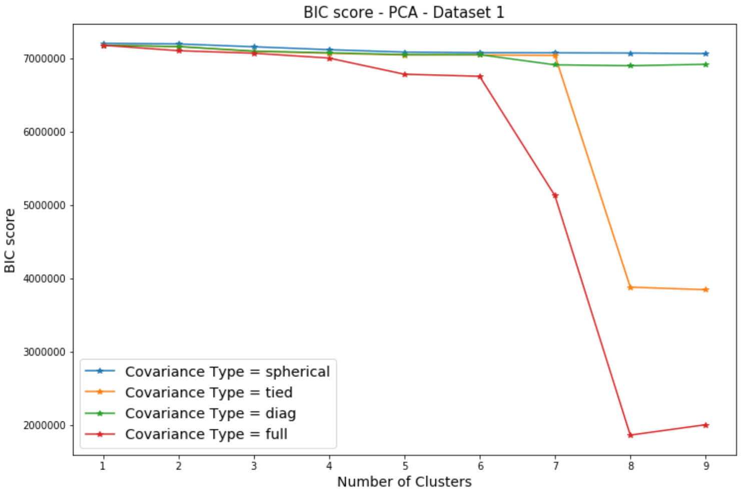
	Without Dimensionality Reduction	With PCA	With ICA	With Random Projection	With Feature Selection
Optimal Clusters	7	7	7	7	7
Sum of squared errors	2400000	1900000	7.2	1900000	1000000
Scores with the original labels					
Adjusted rand score	0.01188	0.01465	0.0099	0.010628	0.01419
Adjusted mutual info score	0.02465	0.02344	0.01952	0.014807	0.02565
Homogeneity score	0.0453	0.04432	0.03669	0.0279	0.04829
Completeness score	0.0169	0.01594	0.0133	0.01008	0.01747
Scores with the clusters of kmeans without dimensionality reduction					
Adjusted rand score	NA	0.4673	0.3919	0.2667	0.30496
Adjusted mutual info score	NA	0.58577	0.54204	0.344746	0.49367
Homogeneity score	NA	0.5974	0.55055	0.3507	0.50199
Completeness score	NA	0.57458	0.5338	0.339	0.48566

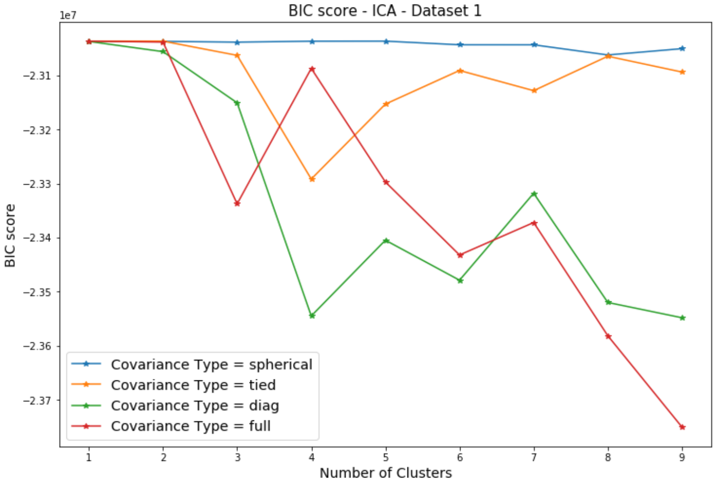
We see that the clusters created without are not very similar to the original clusters as all the scores when compared with the original class labels are less than 0.1

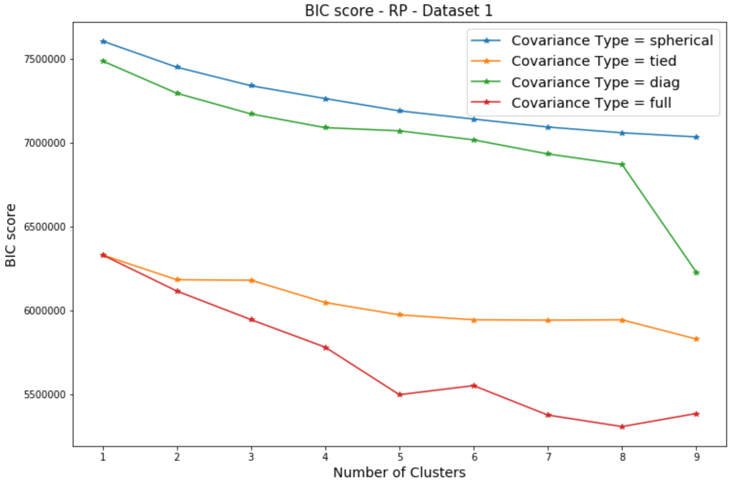
We see that the clusters created with PCA, ICA and Feature selection are the most similar to the clusters created without dimensionality reduction. The scores lie between 0.5 and 0.6. The scores got with Random projections are close to 0.3 which is less.

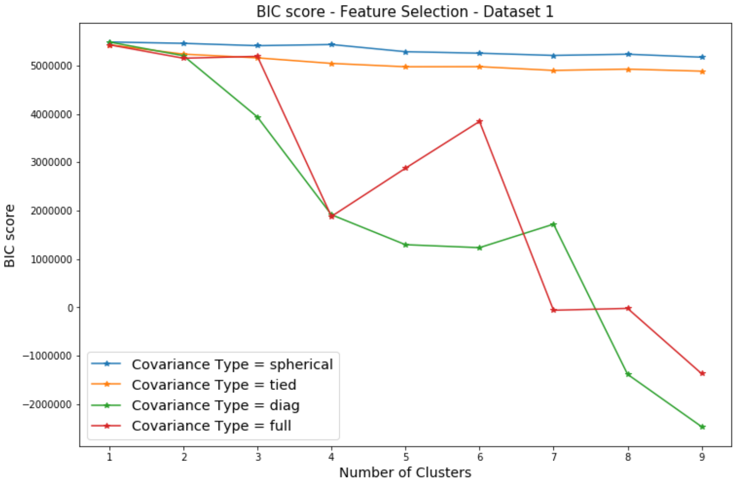
Expectation Maximization











The model selection can be performed with Gaussian Mixture Models using information-theoretic criteria (BIC). The lowest BIC is preferred.

From the table below we can see that the model with ICA features has the lowest BIC.

The optimal clusters given are either 8 or 9.

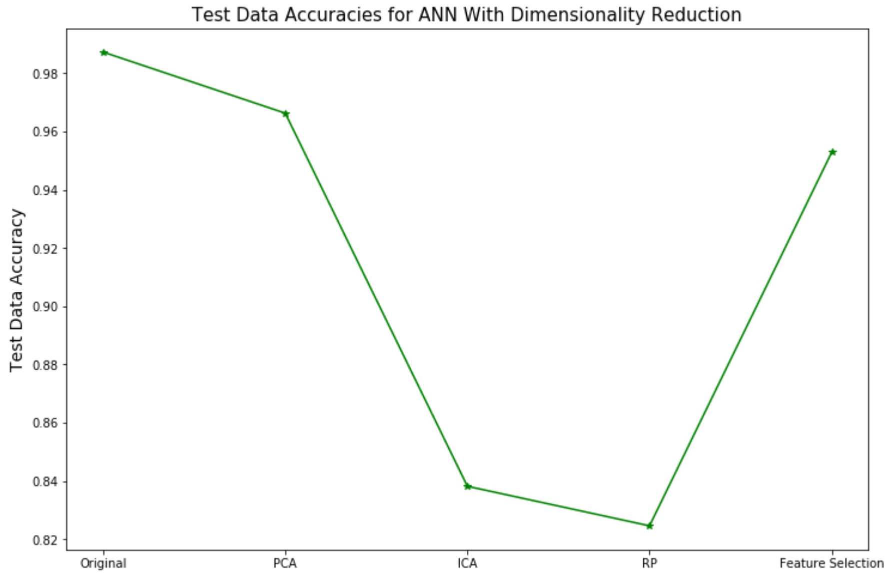
We see that the clusters created with EM are not very similar to the original clusters as all the scores when compared with the original class labels are less than 0.1

The clusters created with random projections are the most similar to the cluster created without dimensionality reduction. They have the highest scores.

	Without Dimensionality Reduction	With PCA	With ICA	With Random Projection	With Feature Selection
Optimal Clusters	8	8	9	8	9
Covariance Type	Diag	full	full	Full	Diag
BIC	4763408	1861753	-23750083	5402378.8	-1923193
Scores with the original labels					
Adjusted rand score	0.025	0.009	0.0085	0.00649	0.008
Adjusted mutual info score	0.04	0.02	0.01	0.014	0.028
Homogeneity score	0.08	0.035	0.018	0.027	0.0529
Completeness score	0.027	0.015	0.0077	0.0936	0.0197
Scores with the clusters of EM without dimensionality reductio					
Adjusted rand score	NA	0.09	0.01668	0.454	0.0306
Adjusted mutual info score	NA	0.243	0.06449	0.609	0.106
Homogeneity score	NA	0.218	0.058	0.603	0.1012
Completeness score	NA	0.274	0.0725	0.616	0.1114

ANN

We get the highest test data accuracy from the ANN run without any dimensionality reduction. The accuracy obtained from PCA and feature selection is also comparable. The ICA and RP gives a little worse performance.

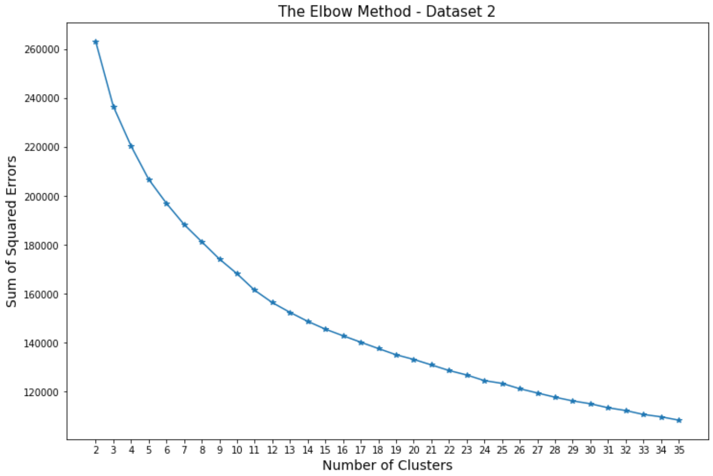


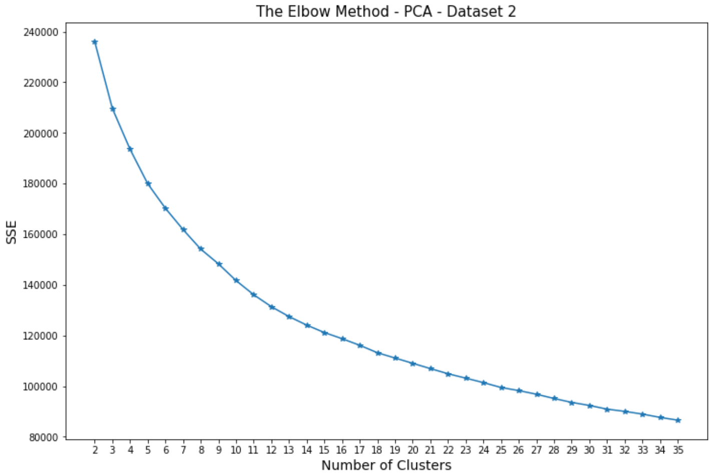
ANN with Clustering Result as Independent Feature

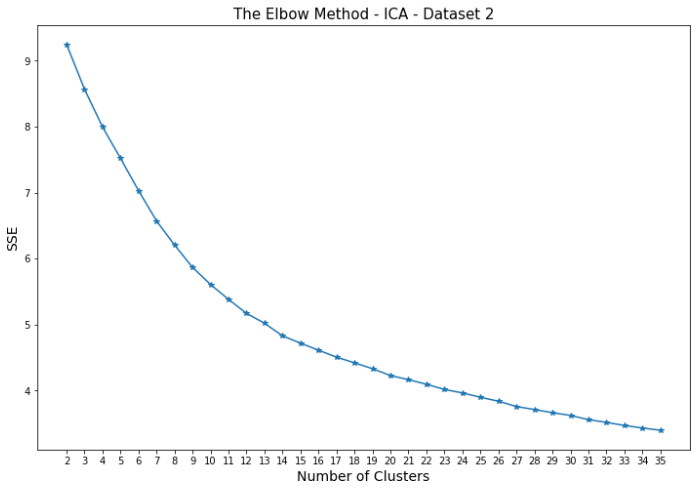
This gives us 58.36% test data accuracy. We used the output of kmeans run with 7 clusters as independent feature.

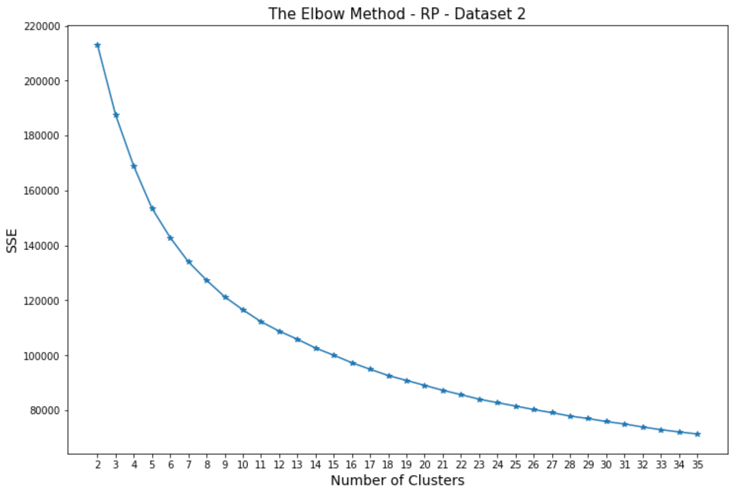
Dataset 2

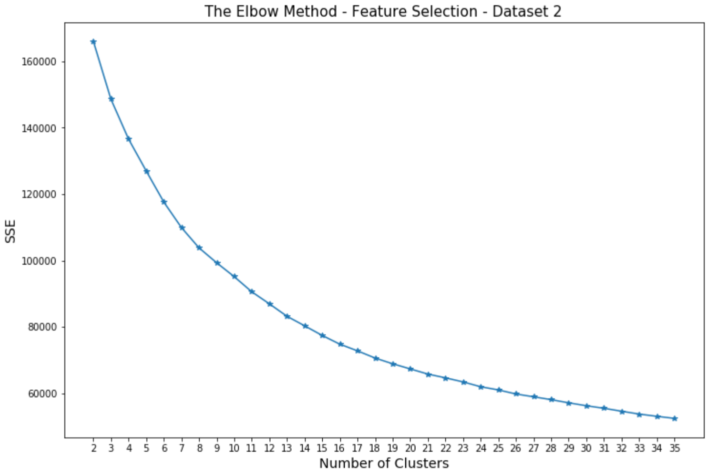
Kmeans











We choose the optimal number of K using these elbow graphs. We choose the k where we observe an elbow in the graph. From all the graphs we get that the optimal number of clusters from the data is 26.

The lowest SSE is obtained with ICA.

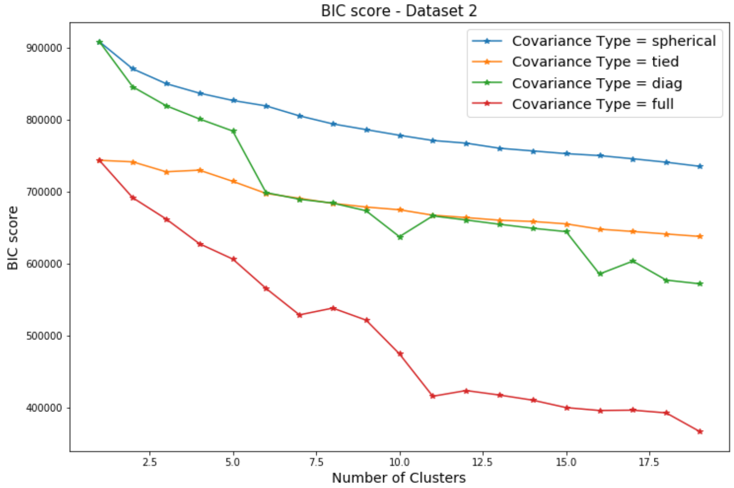
The clusters are somewhat similar to the original clusters because the scores are around 0.3 – 0.4.

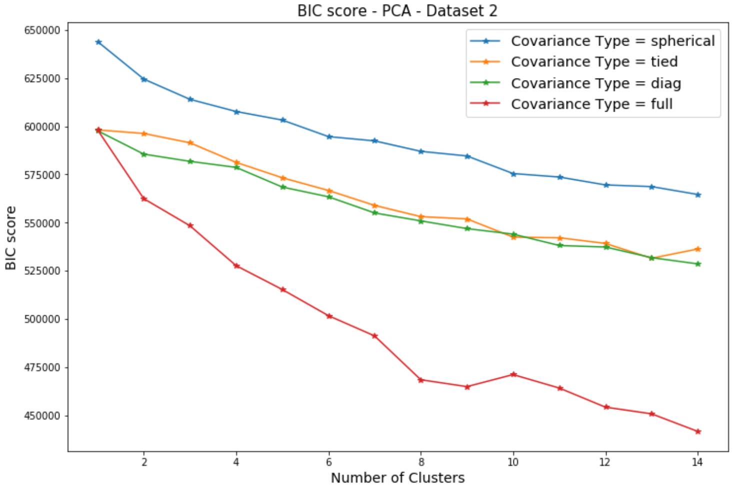
The clusters made with dimensionality reduction are similar to the clusters without it because the scores are around 0.6

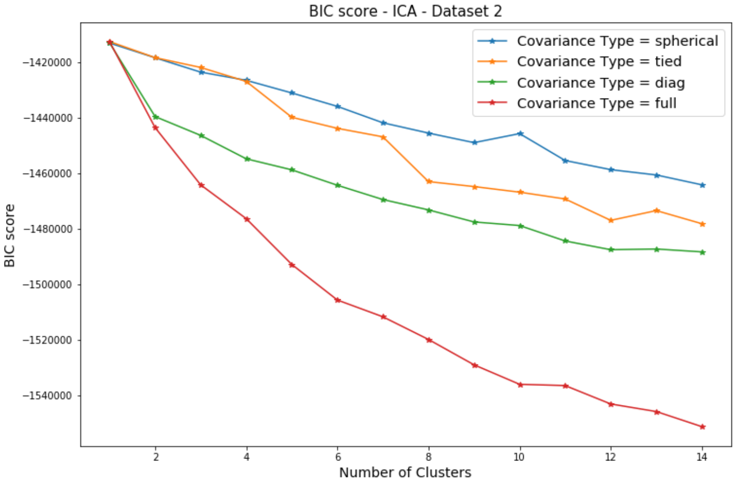
	Without Dimensionality Reduction	With PCA	With ICA	With Random Projection	With Feature Selection
Optimal Clusters	26	26	26	26	26
Sum of squared errors	120000	190000	7.2	1900000	1000000
Scores with the original labels					
Adjusted rand score	0.1514	0.146	0.18	0.108	0.25267
Adjusted mutual info score	0.367	0.36466	0.427	1.308	0.469
Homogeneity score	0.367	0.364	0.4248	0.3088	0.492
Completeness score	0.373	0.37	0.4359	0.3149	0.506

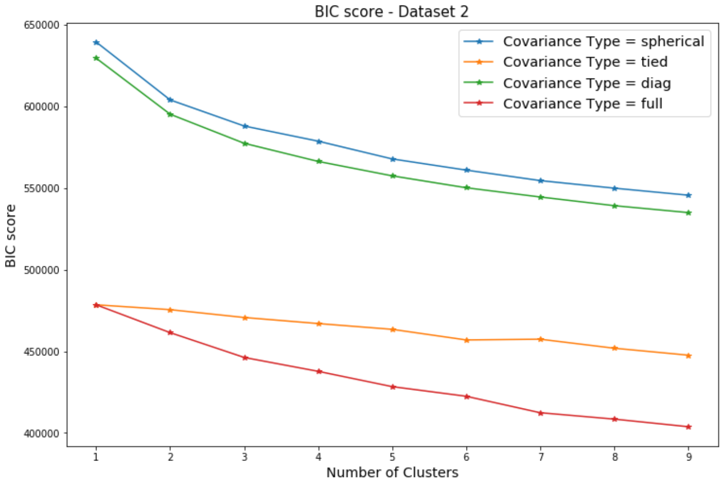
Scores with the clusters of kmeans without dimensionality reduction					
Adjusted rand score	NA	0.596	0.3856	0.217	0.3515
Adjusted mutual info score	NA	0.748	0.624	0.46	0.602
Homogeneity score	NA	0.749	0.623	0.462	0.6012
Completeness score	NA	0.75	0.6286	0.463	0.607

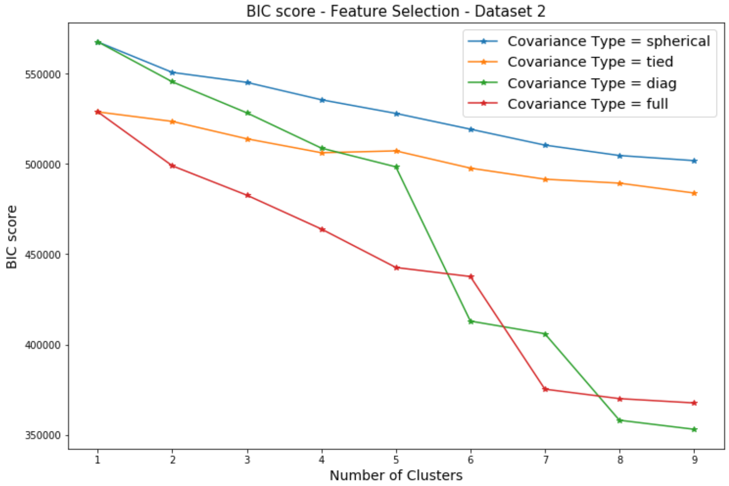
Expectation Maximization











The model selection can be performed with Gaussian Mixture Models using information-theoretic criteria (BIC). The lowest BIC is preferred.

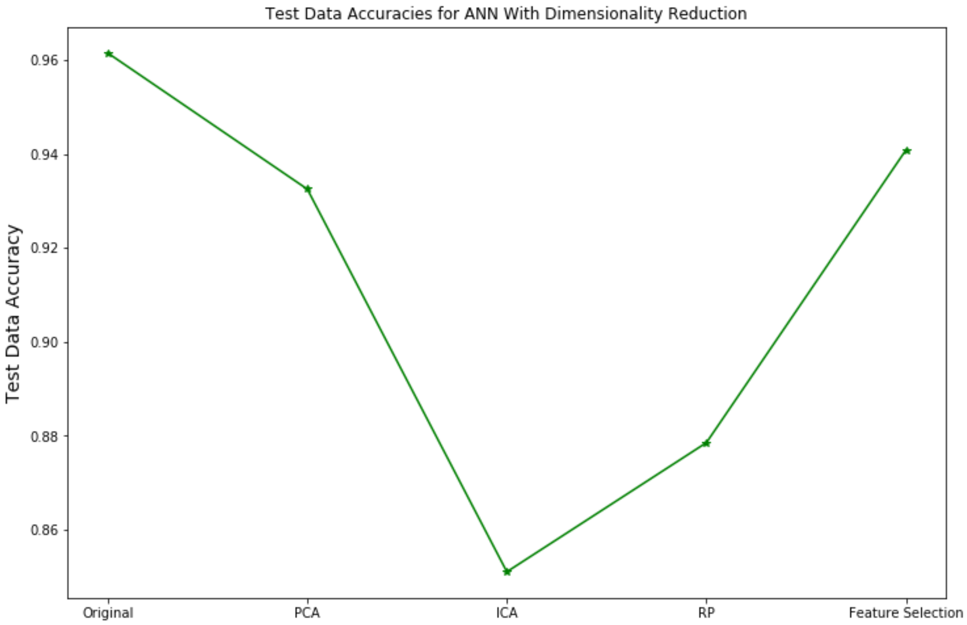
From the table below we can see that the model with ICA features has the lowest BIC.

The optimal clusters are 26.

We see that the clusters created with EM are very similar to the original clusters as all the scores when compared with the original class labels are approx. 0.5

	Without Dimensionality Reduction	With PCA	With ICA	With Random Projection	With Feature Selection
Optimal Clusters	26	26	26	26	26
Covariance Type	Full	full	full	Full	Full
BIC	327414	404378	-1586855	355941	218470
Scores with the original labels					
Adjusted rand score	0.18	0.203	0.238	0.18	0.161
Adjusted mutual info score	0.455	0.47	0.516	0.454	0.477
Homogeneity score	0.444	0.47	0.508	0.446	0.451
Completeness score	0.473	0.48	0.529	0.467	0.512
Scores with the clusters of kmeans without dimensionality reduction					
Adjusted rand score	NA	0.302	0.34	0.257	0.437
Adjusted mutual info score	NA	0.604	0.626	0.548	0.646
Homogeneity score	NA	0.618	0.635	0.554	0.628
Completeness score	NA	0.594	0.62	0.546	0.669

ANN



The accuracies of test data is given in the graph.

ANN with clustering results as features

We get accuracy of 27.8%.