Machine Learning, Assignment 3 Report

Seyed Aghigh

Gatech Email: saghigh3@gatech.edu

1 Introduction

In this assignment the following steps are taken on two data sets; ***Wine data*** [1], and ***Bank data*** [2]. Detail of the Wine and Bank data is available in section 2 and section 7.

(i) **kmean on Original Data**: Perform kmeans on original dataset
(ii) **EM on Original Data**: Perform Expected Maximization (EM) on original data
(iii) **PCA on Data**: Implementing Principle Component Analysis (PCA) algorithm and determine transformed feature space.
  (a) **NN on PCA**: Perform Neural Network (NN) model on projected PCA feature space,
  (b) **keams on PCA**: Perform kmeans on PCA data
  (c) **EM on PCA**: Perform EM on PCA data,
    i. **Add Cluster Labels to PCA Data**: Add cluster labels gained from EM algorithm to original data
    ii. **NN on Label Added PCA Data**: Perform NN on cluster label added data
(iv) **ICA On Data**: Calculate Independent Component Analysis (ICA) on data
  (a) **kmeans on ICA** Perform kmeans on ICA data,
  (b) **EM on ICA**: Perform EM on ICA data
    i. **Add Cluster Labels to ICA Data**: Add cluster labels gained from kmeans algorithm to the ICA data
    ii. **NN on Label Added ICA Data**: Perform NN on cluster label added the data
(v) **RCA On Data**: Implement Randomized Component Analysis (RCA) on data
  (a) **kmeans on RCA**: Perform kmeans on RCA data
  (b) **EM on RCA**: Perform EM on RCA data
    i. **Add Cluster Labels to RCA Data**: Add cluster labels gained from EM algorithm to the RCA data
    ii. **NN on Label Added RCA Data**: Perform NN on cluster label added data
(vi) **MI Feature Selection on Data**: Perform Mutual Information (MI) supervised feature selection.
  (a) **NN on MI Selected Features**: Perform NN on the selected feature from previous step.
  (b) **kmeans on MI Selected Features**: Perform kmeans on MI selected features.

Number of Clusters in kmeans is derived using both Silhouette index and Elbow method. It is tried to pick the one that both methods agree on. **Number of Clusters in EM** is selected based on **Bayesian Information Criterion (BIC)** score. The BIC score find the cut-off between number of model parameters and total number of datapoints versus maximum likelihood function. The Objective is to find optimal number of clusters that minimize the BIC score. while doing so we also consider for covariance type between clusters and features by factoring the **Covariance_type** hyperparameter. For Neural Network (NN), **Multi Layered Perceptron (MLP)** model is considered.

Code Availability All implemented functions and Read-me file are available in the following link
https://drive.google.com/drive/folders/1D7QkFHHz6Si8ANdlx2vFFZhVMaUIxZXK?usp=sharing

2 Original Wine Data

Wine data is a highly imbalanced supervised multi class data which includes 4898 datapoints, 12 continuous features, and 7 class labels.

[1] White Wine Quality
[2] Bank Personal Loan Modeling Link

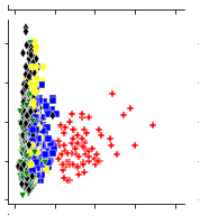
For clustering evaluation pairwise scatter plot are used. This visualization helps us to spot decisive features that creates distinguishable patterns among different clusters. The following pairwise features are considered decisive after observing and analysing clustering algorithm results. These feature are selected separately for kmeans and EM on Wine data. We call them ***kmeans feature group*** and ***EM feature groups***.

kmeans feature group
- Chlorides vs Sulphates
- Citric acid vs Chlorides
- Fixed Acidity vs Density

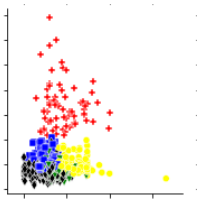
EM feature group
- Chlorides vs Free Sulfur Dioxides
- Residual Sugar vs Chlorides
- Volatile Acidity vs Chlorides

2.1 kmeans on Original Wine Data

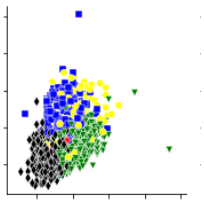
Here, 5 number of clusters is chosen. Figure 1 shows the most interesting pattern that have been found among kmeans feature group. In all the three plots (Figure 1a, Figure 1c, and Figure 1b) red, yellow and blue clusters are distinguishable. Red cluster contains wine with considerably higher chlorides. whereas, yellow cluster shows higher citric acid concentration and blue cluster contains wine with higher sulphur dioxide. In Figure 1c the black cluster shows wine with relatively lower fixed acidity level. Tabel shows the statistics for different clusters. Per cluster, mean of continues variables are considered to report.



(a) Chlorides vs Sulphates



(b) Citric Acid vs Chlorides

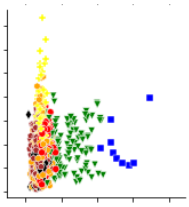


(c) Fixed Acidity vs Density

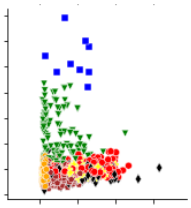
Figure 1: kmeans results on Wine data

2.2 EM on Original Wine Data

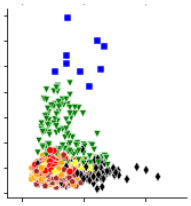
Number of clusters at this section is 7. Figure 2 shows the result for EM feature groups. A clear pattern was observed between Chlorides vs the rest of the features, in which blue cluster always have higher chlorides followed by green cluster. In Figure 2a and Figure 2b. In addition to blue and green points, the black cluster is also distinguishable in terms of higher Volatile Acidity, see Figure 2c. Tabel shows the statistics for different clusters. Per cluster, mean of continues variables are considered to report.



(a) Chlorides vs Free Sulfur Dioxide



(b) Residual Sugar vs Chlorides



(c) Volatile Acidity vs Chlorides

Figure 2: EM results on Wine data.

3 PCA on Wine Data

Figure 3 shows in order to explain at least 95% original Wine data variances, 7 PCA components are needed

3.1 NN on PCA Wine Data

After dimensionality reduction using PCA, NN model is run over new feature space, Table 1. Figure4 shows the training/validation performance metrics are as good as running neural net on original dataset, while running with fewer features helps convergence to happen faster.

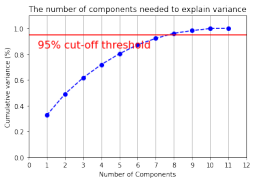


Figure 3: cut-off threshold for PCA Wine data

Test Data Classification Report.
Got 644 / 1225 with accuracy of 52.57 Average Precision Score: 0.68 Average recall Score: 0.68 Average f1-Score: 0.34
Train Data Classification Report.
Got 1741 / 3122 with accuracy of 55.77 Average Precision Score: 0.69 Average recall Score: 0.69 Average f1-Score: 0.37

Table 1: Train & Test Classification Report.

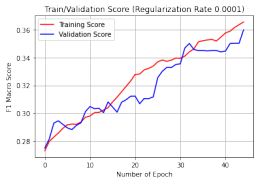
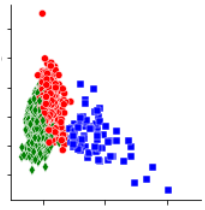


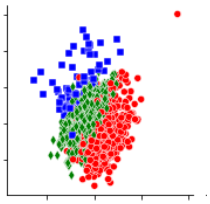
Figure 4: Learning curve for NN on PCA transformed Wine data..

3.2 kmeans on PCA Wine Data

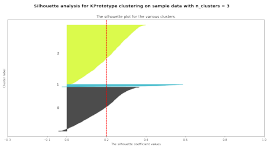
Silhouette index in Figure 5c points to 3 number of clusters. Silhouette index for each sample shows ratio of intra cluster distance and distance to nearest cluster and ranges from [-1, 1]. With 1 being easily separable and 0 being on the boundary of two clusters. Figure 5a shows the three distinguish clusters. As we project the original dataset to new reduced feature space, the interpret ability of the clusters diminishes.



(a) Component 1 vs Component 7



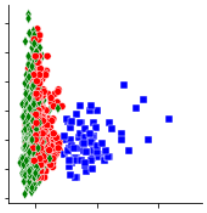
(b) Component 2 vs Component 3



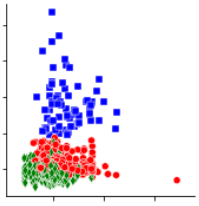
(c) Silhouette Index

Figure 5: kmeans results on PCA wine data

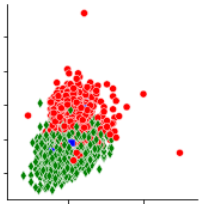
For improving interpretability of PCA clustering, the PCA cluster labels are projected to original data set and pairwise scatter plot are shown in Figure 6. Figure 6a shows blue cluster include wines with higher chloride. whereas, Figure 6c clearly shows the red cluster contains wines with higher density versus blue cluster.



(a) Chloride vs Sulphate



(b) Citric Acid vs Chlorides

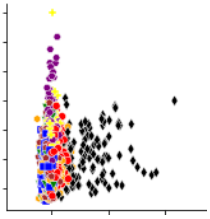


(c) Fixed Acidity vs Density

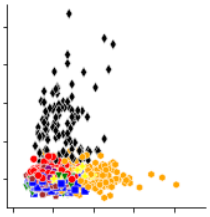
Figure 6: kmeans results on PCA Wine data.

3.3 EM on PCA Wine Data

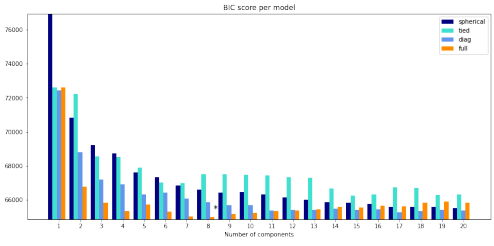
Figure 7c depicts BIC score analysis for determining optimal number of cluster for EM algorithm. 8 number of clusters with **full** covariance type is chosen for this section. EM feature group is plotted in Figure 7. As it seen in Figure 7b, again black cluster shows wines with higher chlorides and yellow cluster contains wines with higher volatile acidity. Comparing these figures with the Figure 2, shows how EM and PCA are pointing to almost same direction.



(a) Chlorides vs Free Sulphur Dioxides



(b) Volatile Acidity vs Chlorides



(c) BIC score for EM algorithm on PCA data

Figure 7: EM results on PCA Wine data.

.

3.4 NN On Label Added To PCA Bank Data

Among possible combination, the PCA Wine data with labels gained by EM model performed the best. Comparing Table 1 and Table 2 indicates the average precision and recall score improved over all classes but overall f1 macros score decreased which could indicate by including the EM cluster labels in feature space precision and recall score for some classes improved significantly but some other class performed poorly and as a result f1 macro score declined. The classification report and learning curve is shown in Table 2 and Figure8 respectively.

Test Data Classification Report.
Got 461 / 1225 with accuracy of 37.63 Average Precision Score: 0.67 Average recall Score: 0.67 Average f1-Score: 0.18
Train Data Classification Report.
Got 1514 / 3122 with accuracy of 48.49 Average Precision Score: 0.74 Average recall Score: 0.74 Average f1-Score: 0.24

Table 2: Train & Test Classification Report.

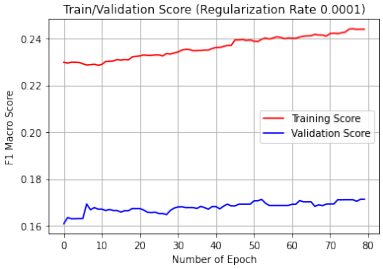


Figure 8: Learning curve for NN performance on PCA transformed Wine data with adding EM labels.

4 ICA on Wine Data

Figure 10 shows when the number of components is equal to two, the sum of mutual information between components are minimum. Therefore, considering Mutual Information and kurtosis metrics, 2 number of components is selected, and original feature space is projected to two dimensional space.

4.1 NN on ICA Wine Data

Table 4 and Figure**??** illustrates the classification report and the learning curve. NN on original Wine data still performs better.

Test Data Classification Report.
Got 550 / 1225 with accuracy of 44.90 Average Precision Score: 0.88 Average recall Score: 0.88 Average f1-Score: 0.13
Train Data Classification Report.
Got 1401 / 3122 with accuracy of 44.88 Average Precision Score: 0.88 Average recall Score: 0.88 Average f1-Score: 0.14

Table 3: Train & Test Classification Report.

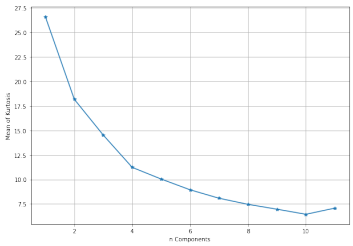


Figure 9: Mean of kurtosis score Versus number of Components data.

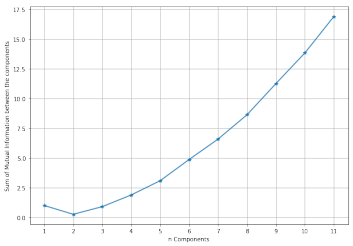
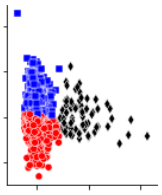


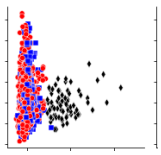
Figure 10: Sum of Mutual Information between components components.

4.2 kmeans on ICA Wine Data

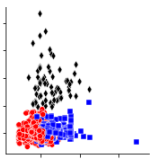
Using elbow and silhouette method, 3 number of clusters is selected. Figure 11a shows the result of kmeans with ICA projected components. All clusters clearly can be recognized with respect to the components. Figure 11 shows the kmeans feature group. Again chlorides, acidity and density features are coming out as a decisive factors.



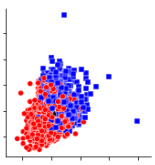
(a) Component 1 vs Component 2



(b) Chlorides vs Sulphates



(c) Citric Acid vs Chlorides

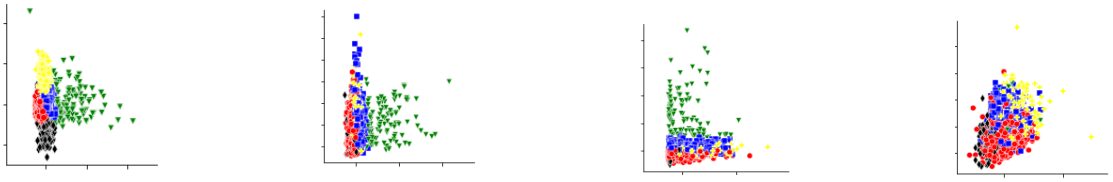


(d) Fixed Acidity vs Density

Figure 11: kmeans results on ICA Wine data

4.3 EM on ICA Wine Data

Figure 37 shows 5 as an optimal number of clusters for this section. Figure 12a shows the resulting clustering in ICA space using EM algorithm, Figure 12 shows the result EM feature group. Almost, the same pattern was observed.



(a) Component 1 vs Component 2 **(b)** Chloride vs Free Sulphur Dioxides **(c)** Residual Sugar vs Chlorides **(d)** Fixed Acidity and Density

Figure 12: EM results on ICA Wine data

4.4 NN on Label Added ICA Wine Data

The labels of the kmeans is fed to the data as a new feature, and then a simple MLP NN model is run. Table 4 and Figure 13 shows the result. Still NN on original Wine data performs the best. Again by adding labels the average precision and recall score improved while the overall f1 macro score declined which again, means by adding the cluster labels some classes are easily distinguishable while it makes some other classes to be invisible.

Test Data Classification Report.
Got 548 / 1225 with accuracy of 44.73 Average Precision Score: 0.81 Average recall Score: 0.81 Average f1-Score: 0.13
Train Data Classification Report.
Got 1411 / 3122 with accuracy of 45.20 Average Precision Score: 0.82 Average recall Score: 0.82 Average f1-Score: 0.15

Table 4: Train & Test Classification Report.

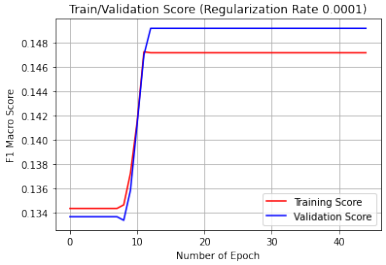


Figure 13: Learning curve for NN performance on ICA transformed Wine data with added kmeans label.

5 RCA on Wine Data

Based on reconstruction error in Table 6,8 is selected as the number of components in Randomized Component Analysis. Selecting the number of components in this case is to some extent subjective, because as the number of components increases the reconstruction error decreases as well, which result in subjective decision making between reducing the number of features while the reconstruction error remains smaller.

5.1 NN on RCA Wine Data

Table 5 and Figure14 show the classification report and learning curve of the result. Agian, NN on original Wine data performs better in terms of f1 macro score.

Test Data Classification Report.
Got 567 / 1225 with accuracy of 46.29 Average Precision Score: 0.77 Average recall Score: 0.77 Average f1-Score: 0.20
Train Data Classification Report.
Got 1474 / 3122 with accuracy of 47.21 Average Precision Score: 0.77 Average recall Score: 0.77 Average f1-Score: 0.21

Table 5: Train & Test Classification Report.

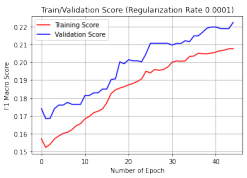


Figure 14: Learning curve for NN performance on RCA transformed Wine data with added kmeans label.

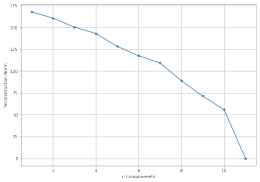
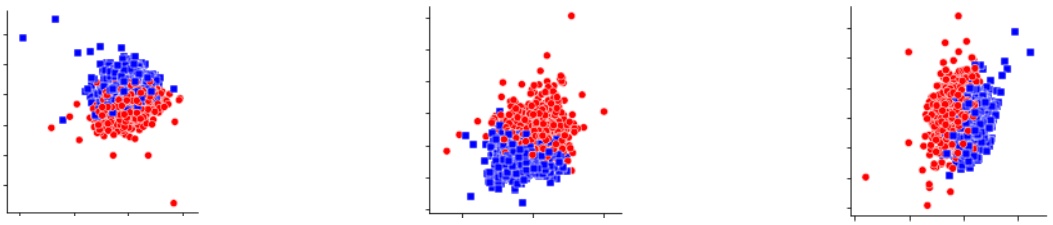


Table 6: fig Reconstruction error for RCA Wine data

5.2 kmeans on RCA Wine Data

Two is chosen as the number of clusters. Clusters clearly distinguishable with respect to each component, Figure 15. No clear pattern was observed among the kmeans feature group.

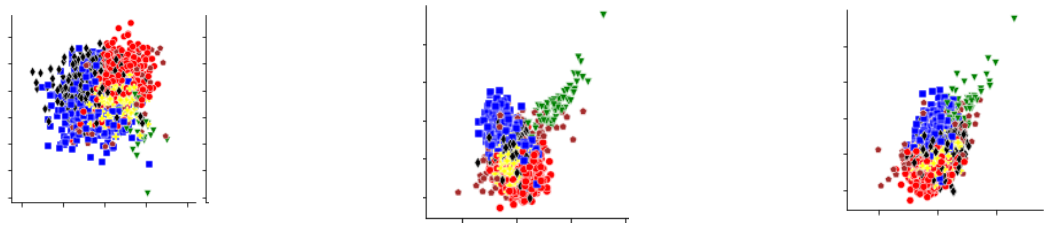


(a) Component 1 vs Component 3 **(b)** Component 2 vs Component 4 **(c)** Component 3 vs Component 6

Figure 15: kmeans result on RCA Wine data

5.3 EM on RCA Wine Data

Figure 37 shows 5 as an optimal number of clusters. In each component, the green clusters followed by the blue ones, are slightly recognizable, Figure 16. No clear pattern was observed among the kmeans feature group.



(a) Component 2 vs Component 6 **(b)** Component 4 vs Component 8 **(c)** Component 7 vs Component 8

Figure 16: EM result on RCA Wine data.

5.4 NN With Cluster Labels Added To The RCA Wine Data

A simple MLP NN with adding labels gained from EM is perform. The result is shown in Table 17. Using RCA generated labels the NN result is very close to original wine feature space, which surprisingly indicates even-though the directions of eight components have been selected randomly but still it outperforms other NN results.

Test Data Classification Report.
Got 641 / 1225 with accuracy of 52.33 Average Precision Score: 0.68 Average recall Score: 0.68 Average f1-Score: 0.31
Train Data Classification Report.
Got 1773 / 3122 with accuracy of 56.79 Average Precision Score: 0.72 Average recall Score: 0.72 Average f1-Score: 0.35

Table 7: Train & Test Classification Report.

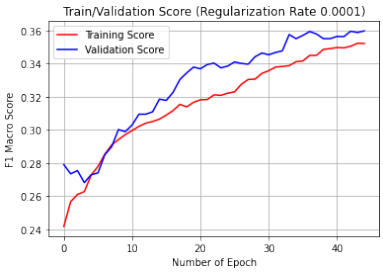


Figure 17: Learning curve for NN performance on RCA transformed Wine data with added EM label.

6 MI Feature Selection On Wine Data

Mutual information (MI) supervised feature selection is method in which the MI pairwise association between the target value and the features are calculated. Features that have low score of MI will be dropped from the dataset. The top MI scored features are Residual Sugar, Density, Alcohol, Figure 19. Interestingly, we saw that in unsupervised clustering setting, features such as chlorides, citric acid and density and fixed acidity are decisive factor that clearly distinguishes clusters. However in supervised feature selection setting we figured those features are not considered decisive and important, this indicates in feature selection applications the objective at hand determines which features are important.

Test Data Classification Report.
Got 580 / 1225 with accuracy of 47.35 Average Precision Score: 0.64 Average recall Score: 0.64 Average f1-Score: 0.28
Train Data Classification Report.
Got 1587 / 3122 with accuracy of 50.83 Average Precision Score: 0.66 Average recall Score: 0.66 Average f1-Score: 0.32

Table 8: Train & Test Classification Report.

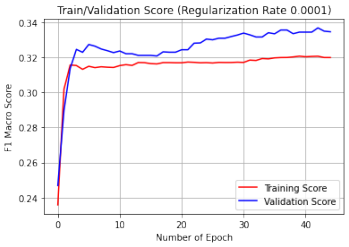


Figure 18: Learning curve for NN performance on RCA transformed Wine data.

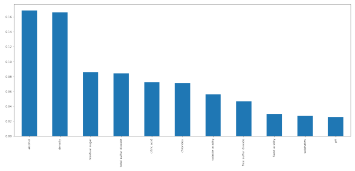
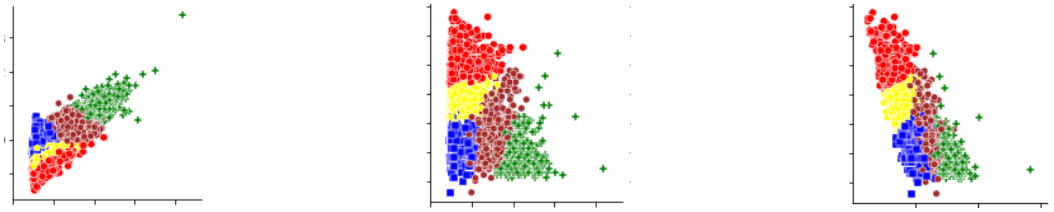


Figure 19: MI features' scores

6.1 kmeans on MI Selected Feature On Wine Data

Figure 20 Here, the number of clusters are 5. Surprisingly, the all the 5 clusters perfectly have clear boundary with respect to each coordinate. In other words, each cluster is meaningful with respect to all features, and clusters are clearly distinguishable.



(a) Residual Sugar vs Density **(b)** Residual Sugar Alcohol **(c)** Density vs Alcohol

Figure 20: kmeans result on MI selected feature Wine data.

7 Original Bank Data

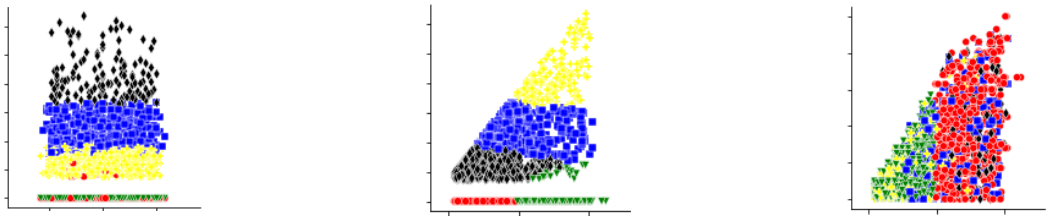
Bank data is a highly imbalanced supervised binary data which includes 5000 datapoints and 14 continuous and categorical variables. All dummy variables are dropped from the data; "Personal Loan", "ID'", "ZIP Code", "Age", "Securities Account", "CD Account", "Online", "CreditCard".

kmeans feature group
- Experience vs Mortgage
- Income vs Mortgage
- CCAvg vs Income

EM feature group
- Experience vs Income
- Income vs Mortgage
- CCAvg vs Mortgage

7.1 kmeans on Original Bank Data

Here, 5 is the number of clusters. Figure 21 shows the most interesting pattern that have been found among the kmeans feature group. In each plot, one of the black, blue, and yellow clusters are distinguishable. Points in black cluster have considerably higher Experience, followed by Blue cluster. While, yellow cluster has lower Experience, they are significantly have higher Income. Blue clusters fairly have lower and higher experience in compare with the black and yellow clusters respectively. They also have the same trend for income, as they have higher income than the black cluster and lower income than the yellow one. In summary, black cluster have high experience, and low income. Whereas, the yellow cluster has the highest income and the lowest experience. Cluster Blue is in the middle. Blues are in the middle

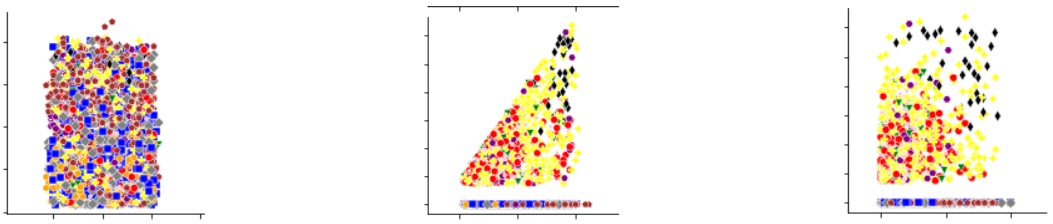


(a) Experience vs Mortgane **(b)** Income vs Mortgage **(c)** Income vs CCAvg

Figure 21: kmeans result on original bank data.

8 EM on Original Bank Data

In this section 10 is selected as the number of clusters. Figure 22 shows the EM feature group.



(a) Experience vs Income **(b)** Income vs Mortgage **(c)** CCAvg vs Mortgage

Figure 22: EM results on original Bank data

9 PCA on Bank Data

Considering 95% cut-off threshold needed components for explained variance, 5 components are selected for the Bank data.

9.1 NN on PCA Bank Data

Here, Table 9, Figure23, Figure24 gives a summary of the NN performance.

Test Data Classification Report.
Got 1003 / 1250 with accuracy of 80.24 Average Precision Score: 0.48 Average recall Score: 0.48 Average f1-Score: 0.46
Train Data Classification Report.
Got 3023 / 3187 with accuracy of 94.85 Average Precision Score: 0.94 Average recall Score: 0.94 Average f1-Score: 0.78

Table 9: Train & Test Classification Report.

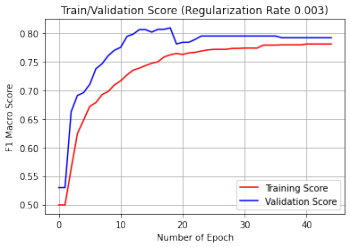


Figure 23: Learning curve for NN performance on PCA Bank data.

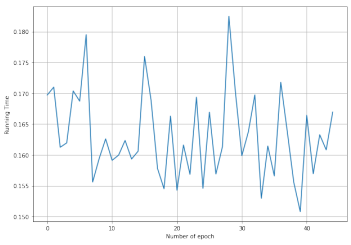
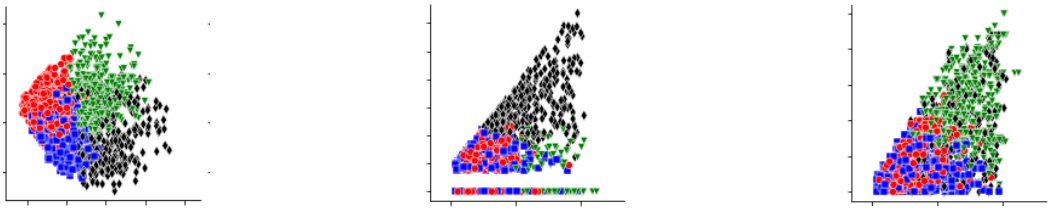


Figure 24: Running Time vs Number of Epoch

9.2 kmeans on PCA Bank Data

Number of clusters is 4. Clusters with respect to the PCA components are well-formed. Figure 25a is one of the plotting result. Among the three EM feature group, the Figure 25b and Figure 25c are the most exciting ones. As it seen from Figure 25b, black clusters tend to have both higher Income and Mortgage. This attitude distinguish these points form the rest. The high difference of this group with the rest is their high income. Similarly, in Figure 25c is shown that the green points have significantly more Income and CCAvg. Whereas, this trend is inverse for the red and blue points.



(a) Component 1 vs Component 5 **(b)** Income vs Mortgage **(c)** Income vs CCAvg

Figure 25: kmeans result on PCA Bank data

9.3 EM on PCA Bank Data

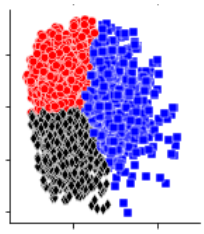
Number of clusters is 9. The pattern found is highly similar to Figure 22, therefore, plots are not depicted here.

9.4 ICA Bank Data

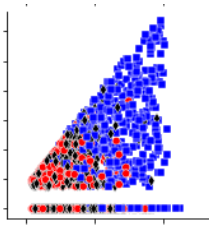
Considering MI and Kurtosis metrics, 2 number of components is selected.

9.5 kmeans on ICA Bank Data

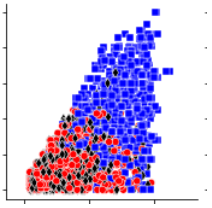
Number of selected clusters 3. While clusters are perfectly have pattern in terms of ICA components (Figure 26a), as it seen from Figure 26b and Figure 26c, some clusters are slightly meaningful (see Figure 27). Surprisingly, The blue cluster have considerably high both Income, Mortgage, and CCAvg. While, the red points have lowest of those.



(a) Component 1 vs Component 2



(b) Income vs mortgage

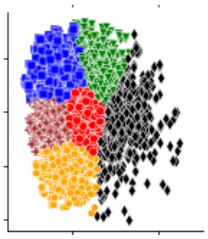


(c) Income vs CCAvg

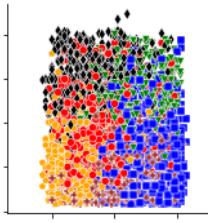
Figure 26: kmeans result on ICA Bank data

9.6 EM on ICA Bank Data

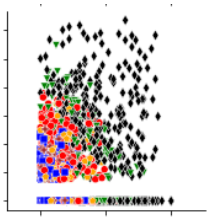
Number of selected components is 6, and Figure 27 shows the most exciting plots. Very interesting, all the clusters have clear boundary making them separate of each other. It can be said that the explained variance catched by first and second component is high enough to separate the clusters. By mapping the cluster labels into the EM feature group, however, those clear pattern diminishes. With all this regard, there exists interesting patter. Specially, it is clear that black points have higher CCAvg, and also they have higher Mortgage in compare with other clusters, Figure 27b and Figure 27c.



(a) Component 1 vs Component 2



(b) Experience vs Income



(c) CCAvg vs Mortgage

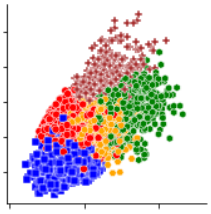
Figure 27: EM result on ICA Bank data.

10 RCA on Bank Data

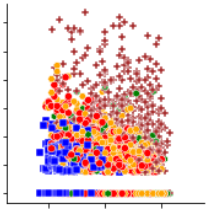
Based on reconstruction error, 5 components is selected as the number of components in Randomized Component Analysis (RCA).

10.1 kmeans on RCA Bank Data

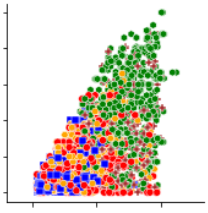
In this section, 5 is selected as the number of clusters. Figure 28 shows the result. Figure 28 shows the clustering result for kmeans on RCA Bank data. Similar to Figure 27a, as it seen from Figure 28a, clusters clearly have pattern in comparing Component 3 and Component 5. It can be said, the explained variable catches by these two components are high enough to produce a clear clusters. Mapping these labels into the kmeans feature group, it turns out the brown clusters tend to be significantly high Experience Figure 28b. Similarly, according to Figure 28c, the green clusters tend to have higher income and CCAvg in compare with the other clusters.



(a) Component 3 vs Component 5



(b) Experience vs Mortgage

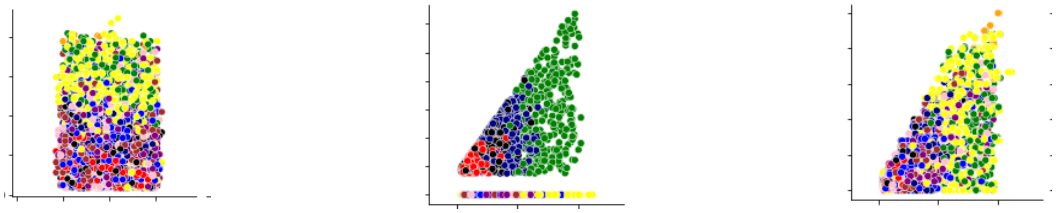


(c) Income vs CCAvg

Figure 28: kmeans result on RCA Bank data.

10.2 EM on RCA Bank Data

Number of clusters for this section is 10. Figure 29 shows the EM feature group. As it seen from Figure 29b, green points have remarkably higher income and Mortgage. The dark blue points are placed in the next position. And the red ones has the least Income and Mortgage. Likewise, in Figure 29c is shown that the yellow points have high Income as well.

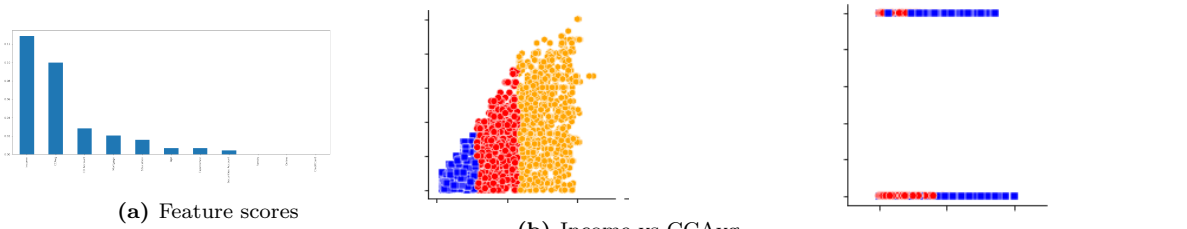


(a) Experience vs Income **(b)** Income vs Mortgage **(c)** Income vs CCAvg

Figure 29: EM result on RCA Bank data

11 MI Feature Selection On Bank Data

MI supervised feature selection is performed to find the selected feature on the original Bank data. 3 features are selected; CD Account, CCAvg, and Income (seeFigure 30b). In Figure 30b and Figure 30c, clearly the three clusters are recognizable. The yellow cluster are the dominant points with respect to the Income and CCAvg. Similarly, the blue have higher CD Account, followed by the red points.

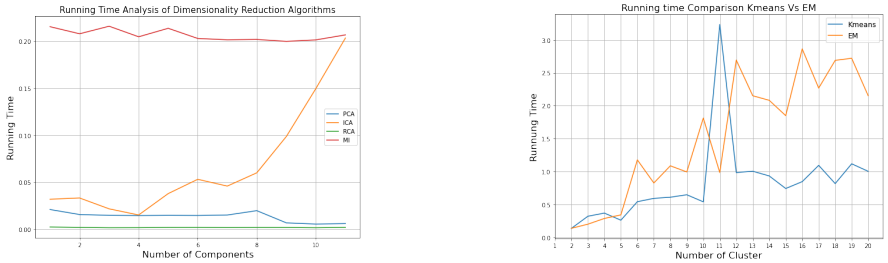


(a) Feature scores **(b)** Income vs CCAvg **(c)** CCAvg vs CD Account

Figure 30: kmeans result on MI selected features

12 Conclusion Wine & Bank Data

Figure 31 shows the running time of aforementioned algorithm. In Figure 31a, the running time of PCA, ICA, RCA, and MI on Wine data is illustrated. As it seen, MI regardless of the number of components, has the higher running time. While, in ICA as the number of components increases, the running time tend to higher exponentially. On the other hand, RCA has the least followed by PCA. Considering the clustering pattern generated by these method on Win data, Figure 1 and Figure 2 and Figure 6 and Figure 7 and Figure 11 and Figure 20, it can be said that both kmeans on PCA Wine and kmeans on MI Wine data produce the most clear clusters. Resulting clustering are clearly distinguishable, each has a specific pattern with respect to specific features. In summary, clusters have clear description. Figure 31b shows the running time for kmeans and EM on Wine data. kmeans is more time efficient, and as the number of clusters increases the EM spend more time. With similar explanation, among Figure 21 and Figure 22 and Figure 25 and Figure 26 and Figure 28 and Figure 29 and Figure 30, it can be concluded that kmeans on original Bank data or kmeans on ICA Bank data produces the best clustering outcome.

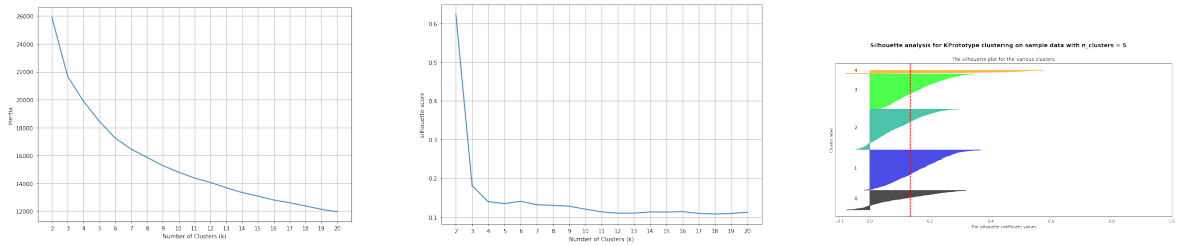


(a) Running Time of PCA, ICA RCA, and MI on Wine data. **(b)** Running time of kmeans and EM on original Wine data.

Figure 31: Running time comparison.

13 Supplementary Data

13.1 Number of Clusters for Kmeans on Original Wine Data



(a) Elbow method for kmeans on original Wine data

(b) Silhouette method for kmeans on original Wine data

(c)

Figure 32: kmeans clustering cross validation for number of clusters in original wine data

13.2 Number of Clusters for EM on Original Wine Data

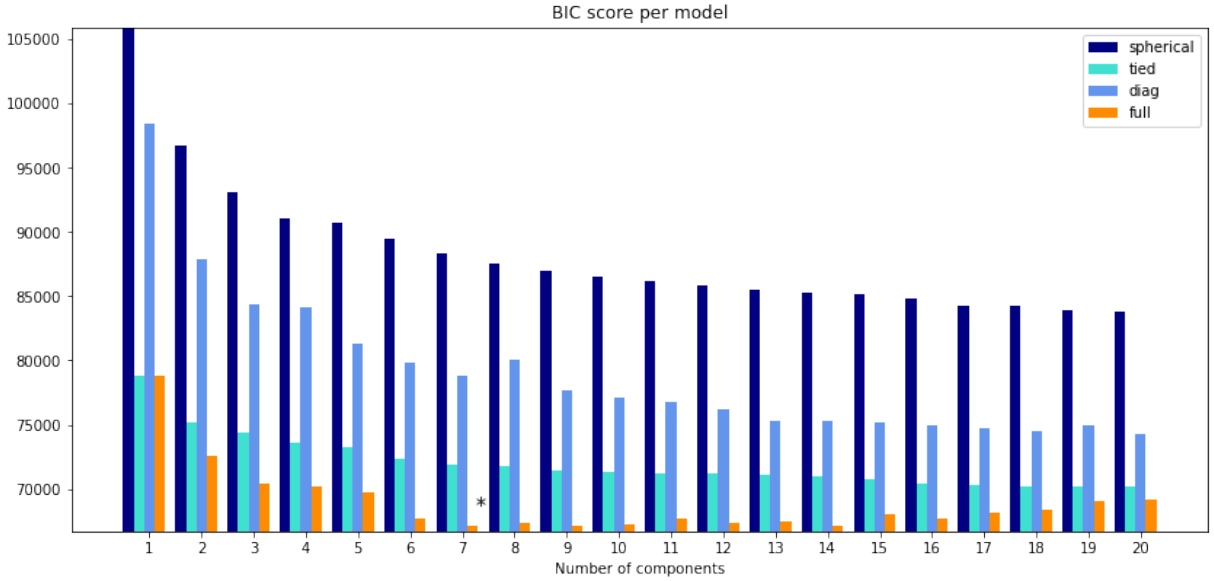
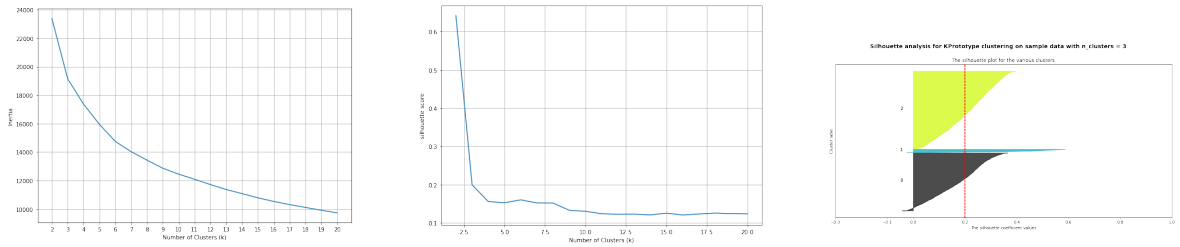


Figure 33: BIC score clustering cross validation for number of clusters in original wine data.

13.3 Number of Clusters for kmeans on PCA Wine Data



(a) Elbow method for kmeans on PCA Wine data

(b) Silhouette method for kmeans on PCA Wine data

(c)

Figure 34: kmeans clustering cross validation for number of clusters in PCA wine data

13.4 Number of Clusters for EM on PCA Wine Data

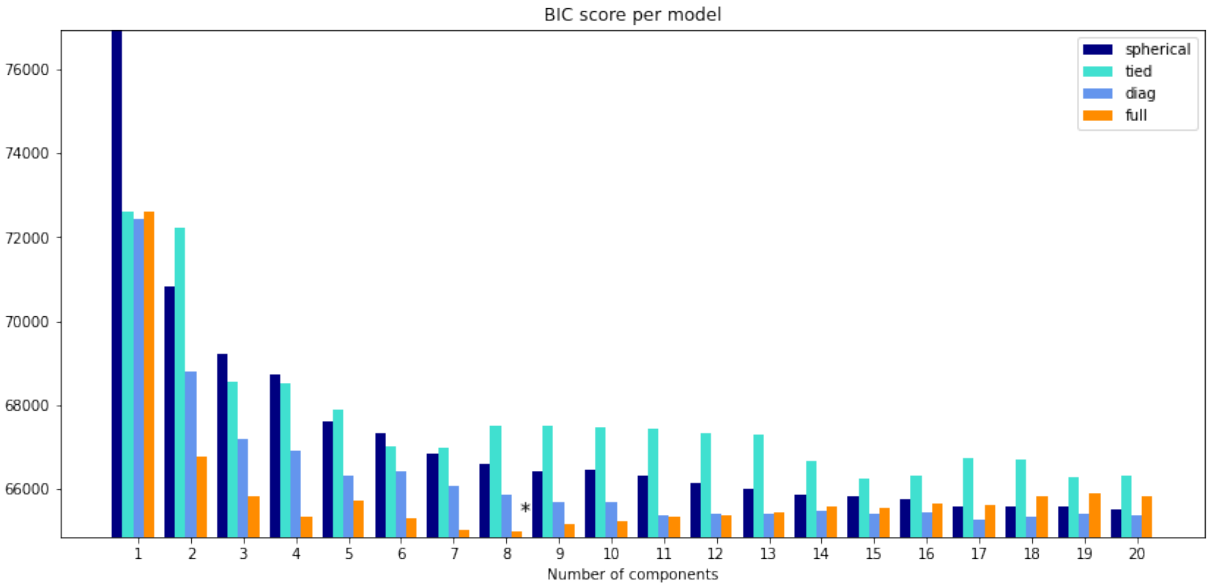
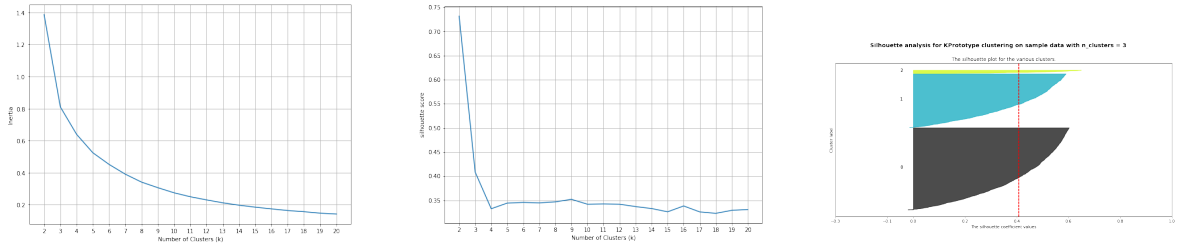


Figure 35: BIC score clustering cross validation for number of clusters in PCA wine data.

13.5 Number of Clusters for kmeans on Wine ICA Data



(a) Elbow method for kmeans on ICA Wine data

(b) Silhouette method for kmeans on ICA Wine data

(c) Silhouette Method

Figure 36: kmeans clustering cross validation for number of clusters in ICA wine data

13.6 Number of Clusters for EM on ICA Wine Data

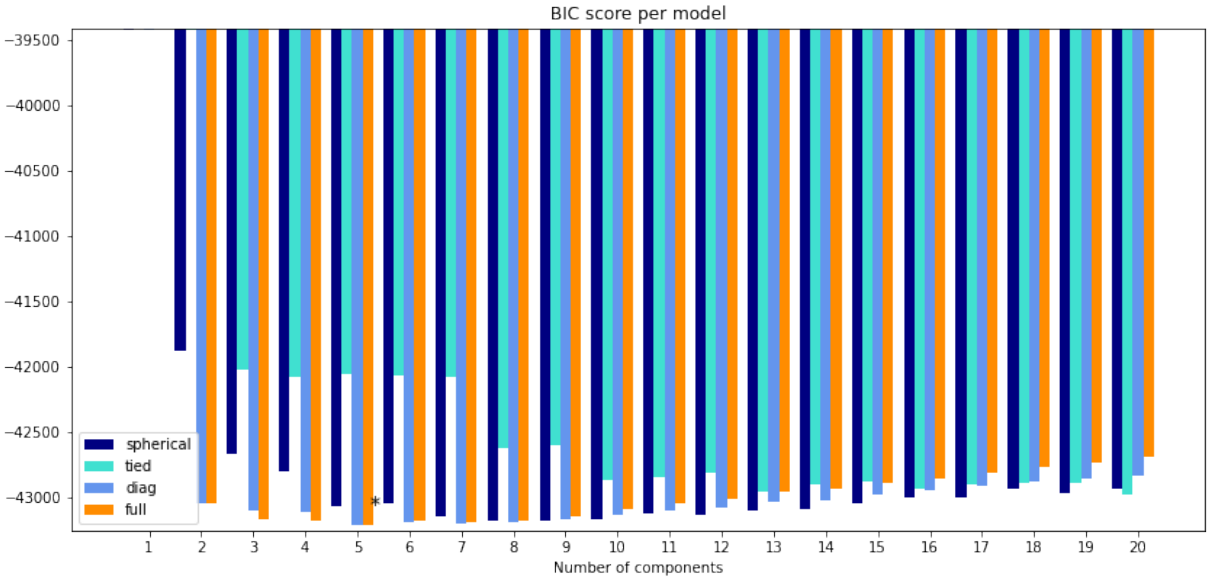
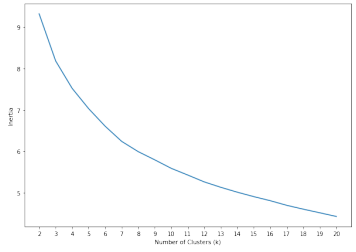
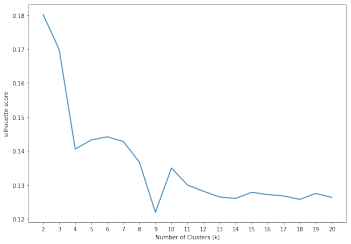


Figure 37: BIC score clustering cross validation for number of clusters in ICA wine data.

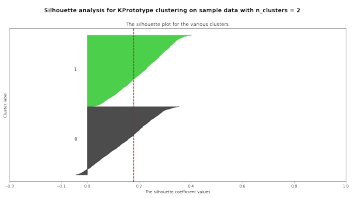
13.7 Number of Clusters for kmeans on Wine RCA Data



(a) Elbow method for kmeans on RCA Wine data



(b) Silhouette method for kmeans on RCA Wine data



(c) Silhouette Method

Figure 38: kmeans clustering cross validation for number of clusters in RCA wine data

13.8 Number of Clusters for EM on RCA Wine Data

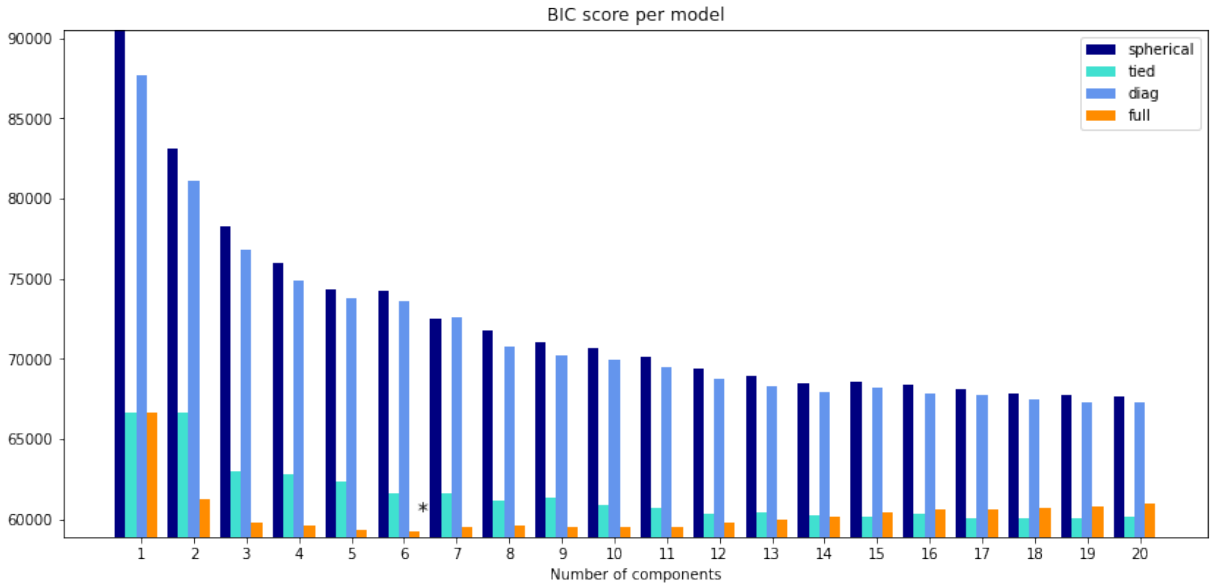
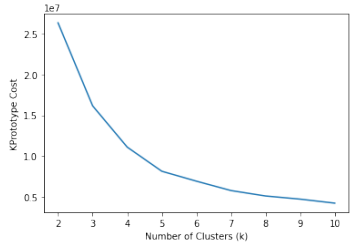
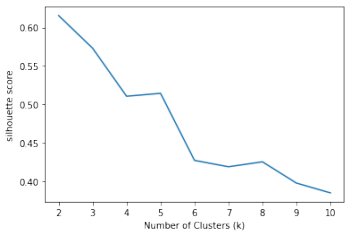


Figure 39: BIC score clustering cross validation for number of clusters in RCA wine data.

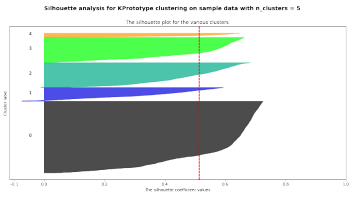
13.9 Number of Clusters for kmeans on Bank Original Data



(a) Elbow method for kmeans on Original Bank data



(b) Silhouette method for kmeans on Original Bank data



(c) Silhouette Method

Figure 40: kmeans clustering cross validation for number of clusters in Original Bank data

13.10 Number of Clusters for EM on Original Bank Data

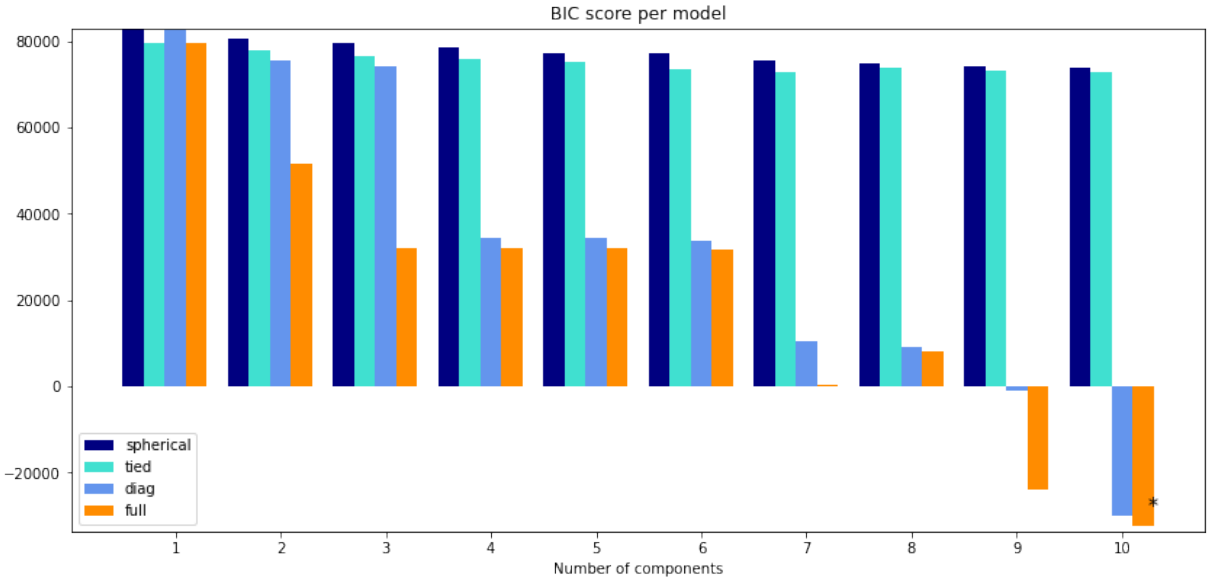
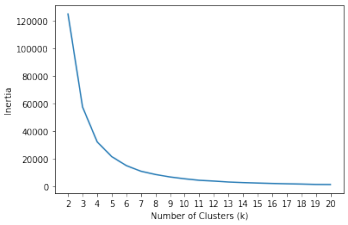
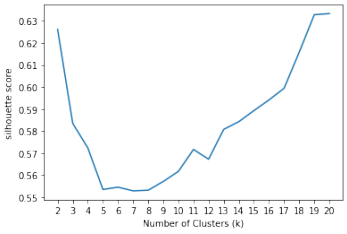


Figure 41: BIC score clustering cross validation for number of clusters in Original Bank data.

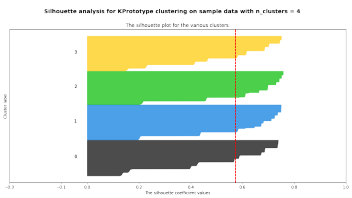
13.11 Number of Clusters for kmeans on Bank PCA Data



(a) Elbow method for kmeans on PCA Bank data



(b) Silhouette method for kmeans on PCA Bank data



(c) Silhouette Method

Figure 42: kmeans clustering cross validation for number of clusters in PCA Bank data

13.12 Number of Clusters for EM on PCA Bank Data

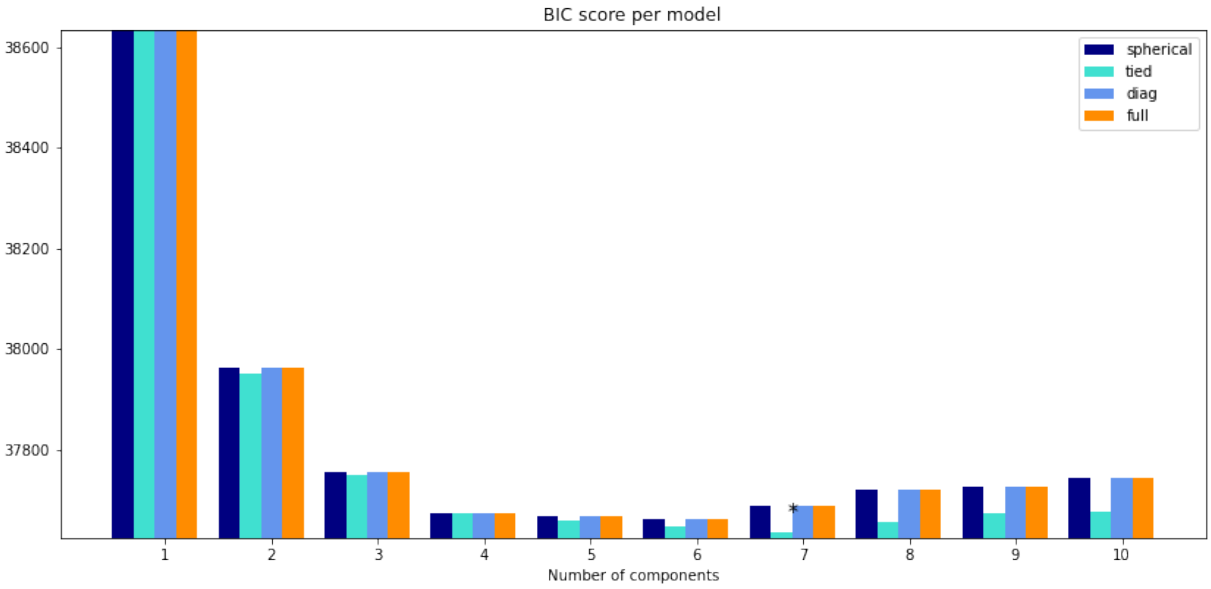
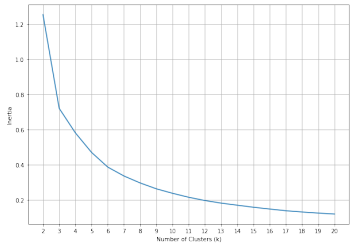
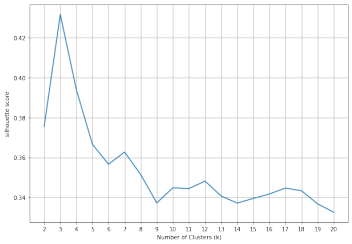


Figure 43: BIC score clustering cross validation for number of clusters in CA Bank data.

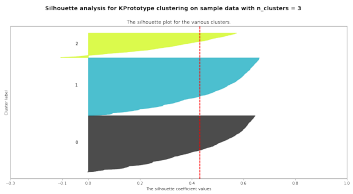
13.13 Number of Clusters for kmeans on Bank ICA Data



Elbow method for kmeans on ICA Bank data



Silhouette method for kmeans on ICA Bank data



Silhouette Method

kmeans clustering cross validation for number of clusters in ICA Bank data

13.14 Number of Clusters for EM on ICA Bank Data

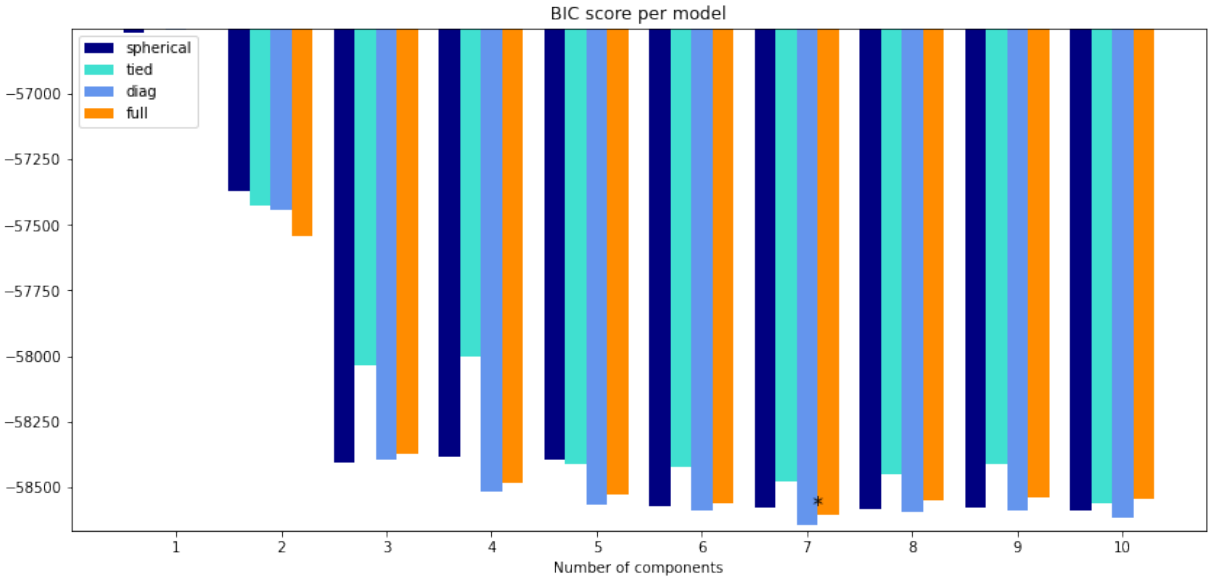
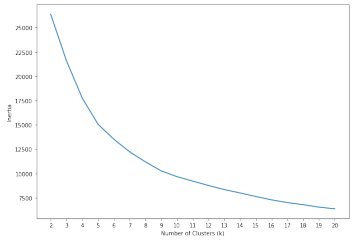
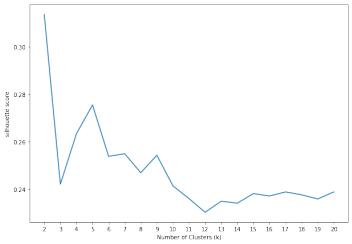


Figure 45: BIC score clustering cross validation for number of clusters in ICA Bank data.

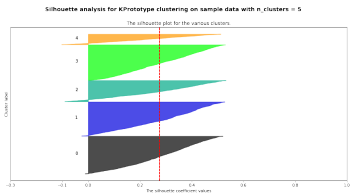
13.15 Number of Clusters for kmeans on Bank RCA Data



Elbow method for kmeans on RCA Bank data



Silhouette method for kmeans on RCA Bank data



Silhouette Method

kmeans clustering cross validation for number of clusters in RCA Bank data

13.16 Number of Clusters for EM on RCA Bank Data

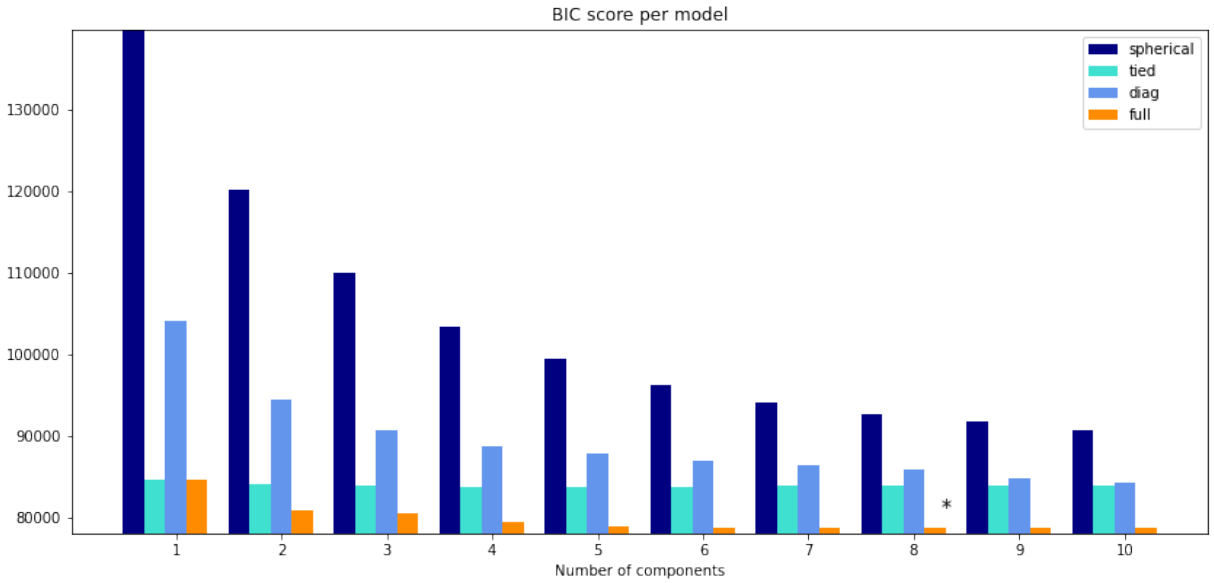


Figure 47: BIC score clustering cross validation for number of clusters in RCA Bank data.