CS 7641 Assignment #3: Unsupervised Learning and Dimensionality Reduction

William Koppelman
wkoppelman3@gatech.edu

I. Introduction

This assignment will analyze the use of unsupervised learning, specifically clustering, on our datasets that we've been using in previous assignments. It will also look at dimensionality reduction and clustering on this reduced data. Finally we will compare the results of our neural network model from assignment 1 using the data from these unsupervised learning methods as input.

II. Datasets

A. Contraceptive Method Choice

This dataset is a subset of a 1987 contraceptive prevalence survey taken in Indonesia. It is a survey of married women who were not pregnant. The goal of the survey was to predict current contraceptive use (no use, short-term use, or long-term use methods) based on demographic and socio-economic characteristics of the women and their husband. This is a ternary classification dataset with 9 attributes and 1,473 instances. More information can be found about the dataset in the README.

This dataset is interesting because it could be a binary classification but they chose to make it ternary. This will be interesting to analyze in the clustering of the data. In addition, from the supervised learning assignment it appeared that short-term use behaved similar to no use when I had expected it would be closer to long-term use. This can be explored with clustering as well.

It also includes numerical as well as ordinal and categorical data. Once the qualitative data is one-hot encoded we have 14 attributes. There appears to be some overlap in the attributes that could be taken advantage of in dimensionality reduction.

B. Red Wine Quality

This dataset is a collection of physicochemical attributes on the ordinal sensory output of quality. It only looked at Portuguese "Vinho Verde" red wine. There are 11 attributes and 1,599 instances in this dataset. The quality score ranges from 0 to 10 but only scores between 3 and 8 are present in the data. More information can be found about the dataset in the README.

This dataset should be interesting to use with clustering to see if the quality scores are distinct or, more likely, similar scores can be clustered together. We can also explore if binary classification could be appropriate for this dataset.

III. Clustering

Clustering is an unsupervised learning technique for data classification. Using unlabeled data, it groups them together based upon shared characteristics.

The k-means method clusters the data, using a distance metric, into k different clusters. For this assignment I chose to use the Euclidean distance metric. I implemented the clustering using the scikit learn package in Python 3.x.

Expectation maximization (EM) iterates using random components. The expectation step calculates the probability of a data point being generated by each of these components. While the maximization step maximizes the probability of the data given those assignments. For this assignment I used the Gaussian Mixture Model (GMM) estimator in the scikit learn package of Python. This implements EM using Gaussian expectation models.

To determine the appropriate number of clusters I used a number of different metrics:

- Maximize the **accuracy** of clusters having instances with the same labels.
- Maximize the **adjusted mutual information (AMI)** score, which measures the similarity between clusters (accounting for chance).
- Maximize the **silhouette score (SS)**, which compares the intra- and inter-cluster distances.
- With the **sum of squared errors (SSE)** in the k-means models we look for the elbow in the graph to reduce error while also keeping the model simple and not overfitting.
- Minimize the **Bayesian information criterion (BIC)** in the GMM models to prevent overfitting.

A. Contraceptive Method Choice

Looking at the metrics for k-means (Fig. 1) you can clearly see that the max of accuracy, AMI, and SS all occur at six clusters. This also appears to be where the elbow in the SSE curve is as well. Fig. 2 shows a silhouette plot of six k-means clusters for the contraceptive use dataset. We can see from the height of each bar the size of each cluster and the width of each

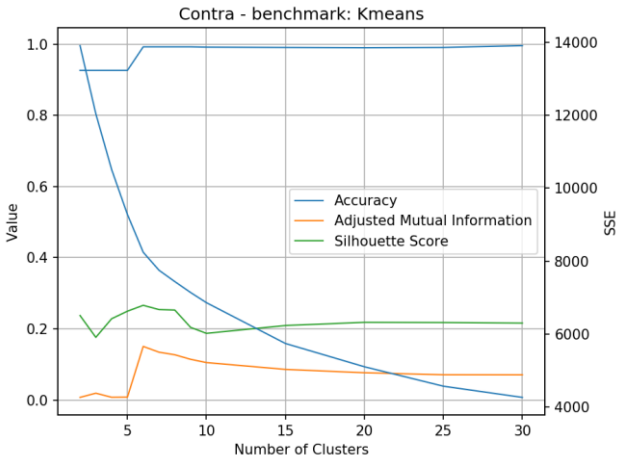

Fig. 1. Contraceptive use k-means metrics

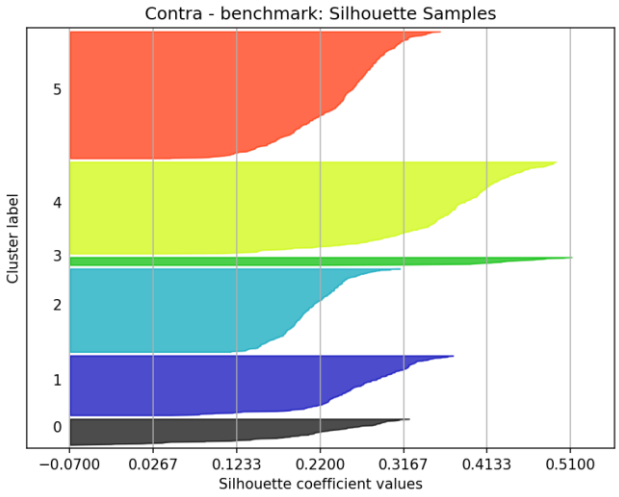

Fig. 2. Contraceptive use k-means 6 cluster silhouette plot

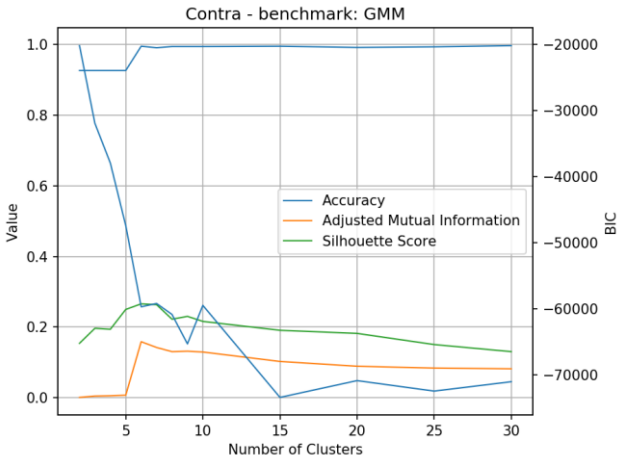

Fig. 3. Contraceptive use GMM metrics

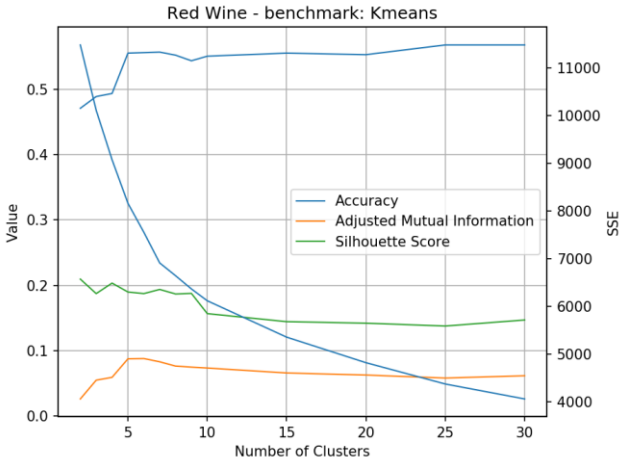

Fig. 4. Red wine rating k-means metrics

bar shows the SS of each cluster. This shows that the data has a few smaller clusters that accurately group portions of the data (0 and 3) while the other clusters are larger in size.

Similarly for GMM (Fig. 3) the max of accuracy, AMI, and SS occur at six clusters. For BIC the minimum is at 15 clusters but 6 clusters is also an elbow in its curve.

The original classification into three classes does not appear to appropriately cover the nuances of the data. Contraceptive use is an extremely complex decision, especially in religious societies. It is possible that the women are taking different routes in their decision making process to arrive at similar decisions.

For example, a young, progressive woman from a religious family might reason that short-term contraceptive use is an appropriate trade-off in her circumstances. Another person might prefer long-term contraceptive use but her standards of living only afford her short-term use. Both of these women would be classified as short-term contraceptive use but could be clustered very differently.

B. Red Wine Quality

The clustering metrics for the red wine quality dataset are more difficult to interpret. Fig. 4 shows the metric for k-means. The accuracy shows a sharp increase at 5 clusters, which it holds on to until 7 clusters. The SSE also shows an elbow between 5 and 7 clusters. The AMI reaches its maximum at 5 and 6 clusters. However, the SS has its maximum at 2 and 4 clusters. Taking all of this into account I would select 5 clusters to use.

The silhouette plot for the red wine dataset with 5 k-means clusters is shown in Fig. 5. It appears that clusters 0-3 are similar in size while cluster 4 captures a small part of the population very well.

Fig. 6 is equally as confusing for the metrics of GMM. The accuracy suggests 3 to 7 clusters, the AMI 3 to 4 clusters, the SS 2 clusters, and the BIC 5 or 7 clusters. Taking into account our k-means metrics and our domain knowledge that quality scores of 3 to 8 are present in the data, we will again choose 5 clusters for the data.

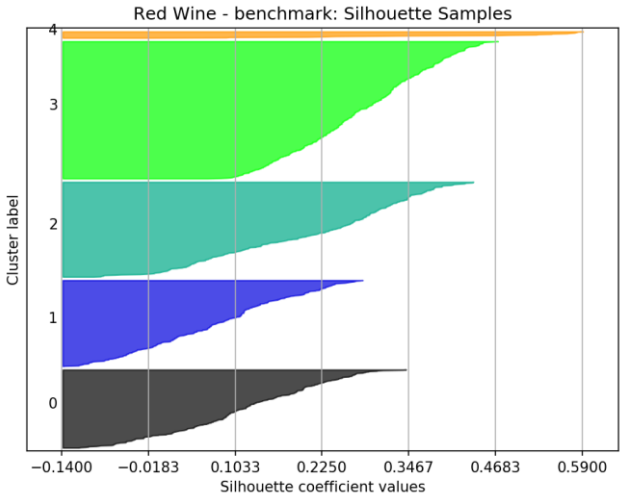

Fig. 5. Red wine quality k-means 5 cluster silhouette plot

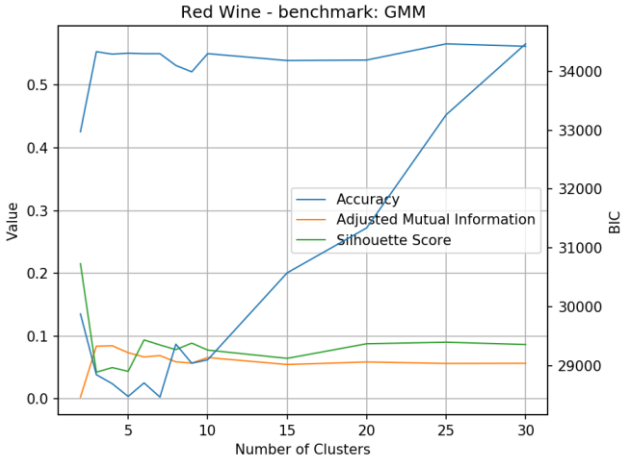

Fig. 6. Red wine quality GMM metrics

Quality scores on a 1 to 10 scale are very subjective and allow for scores to blend with their neighbors. The 5 clusters selected for each model fits with generally accepted practices that humans can only provide distinctive ratings with 4 to 7 alternatives (Lozano, García-Cueto and Muñiz, 2008). This is why most shopping sites have you rate products out of 4 or 5 stars. It is also interesting to note that the silhouette scores for both models recommends a simple binary classification of whether or not the wine was good.

IV. Dimensionality Recduction with Clustering

I applied dimensionality reduction (DR) using scikit learn in Python 3.x. I performed Principal Component Analysis, Independent Component Analysis, Random Projections, and for my final DR I chose Information Gain (using a random forest).

I then plotted the components for each and used the elbow method to choose the appropriate number of components for each DR method.

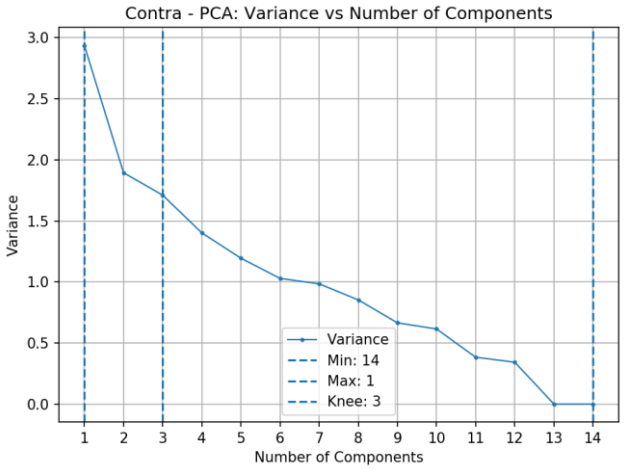

Fig. 7. Contraceptive use PCA component variance

Finally, I performed clustering, using k-means and GMM, similar to above.

A. *Principal Component Analysis*

Principal Component Analysis (PCA) finds orthogonal eigenvectors whose eigenvalues explain the most variance in the dataset. We are looking to use the fewest components that explains the most variance in the data. I chose to do this visually by determining the elbow/knee in the variance graph.

As seen in Fig. 7 for the contraceptive use data three components were chosen. By only choosing this first three components and removing the rest, we hope to remove some of the noise from our dataset.

After applying this DR to the contraceptive data it was then clustered using k-means and GMM. The same metrics as above were then plotted in Figs. 8 and 9. For k-means it appears that four clusters is the appropriate choice which is less than the data prior to DR. For GMM 6 appears to be the best choice, which is the same as before.

Comparing the k-means metrics with the original data, we see that we now have a lower SSE but the accuracy is negatively affected. The SS indicates that we have better clustering of the data. Similarly for GMM we have a higher SS indicating better clustering but a lower accuracy.

Fig. 10 shows the PCA for the red wine quality data. Here, three components were chosen as well. Looking at the metrics for k-means clustering using the PCA with three components, it appears that four clusters best captures the data (Fig. 11), while six appears to be best for GMM (Fig. 12). Both of these are within our domain knowledge of the rating scale and similar to the non-DR clustering.

The k-means metrics show a lower SSE, a better clustering from the SS and in this case a higher accuracy. The GMM shows a less complex model (BIC) a much higher SS for better clustering and a similar accuracy score.

PCA appears to cluster the data together better but does not necessarily improve the accuracy.

B. Independent Component Analysis

Independent Component Analysis (ICA) maximizes the independence of its components in the dataset. It is typically used for separating signals. We are looking to use the fewest components that have the highest kurtosis in the data. I again chose the elbow/knee method.

As seen in Fig. 13 for the contraceptive data seven components were chosen. After applying this DR it was then clustered. For k-means it appears that seven clusters is the appropriate choice (Fig. 14) while for GMM nine appears to be the best choice (Fig. 15). Both of these are more clusters than the non-DR data.

The k-means metrics show a much lower SSE, by a couple of orders of magnitude. The accuracy is lower but the clustering is similar to the original data. For GMM we have a higher SS indicating better clustering and a similar accuracy.

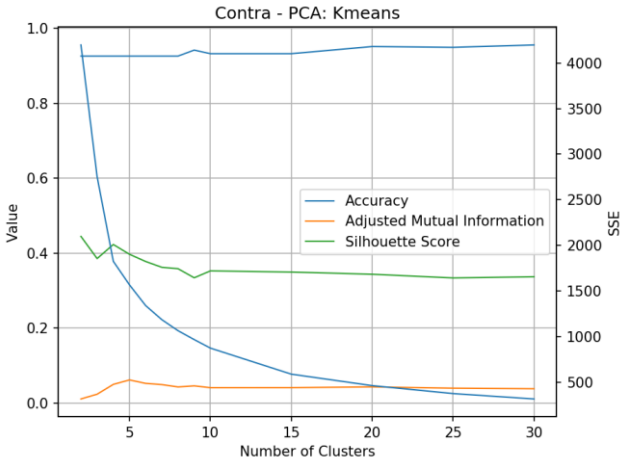

Fig. 8. Contraceptive use 3 component PCA with k-means clustering

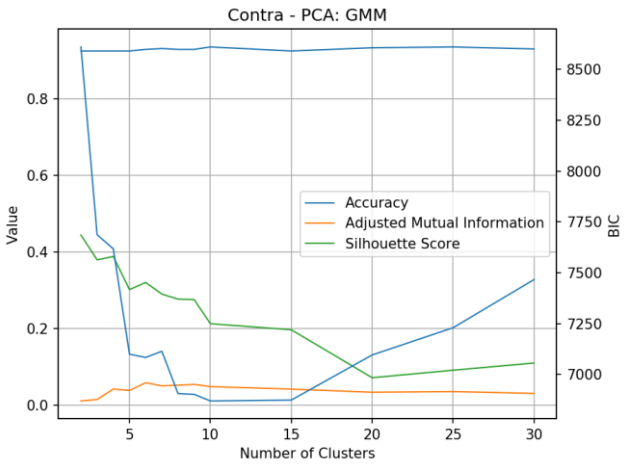

Fig. 9. Contraceptive use 3 component PCA with GMM clustering

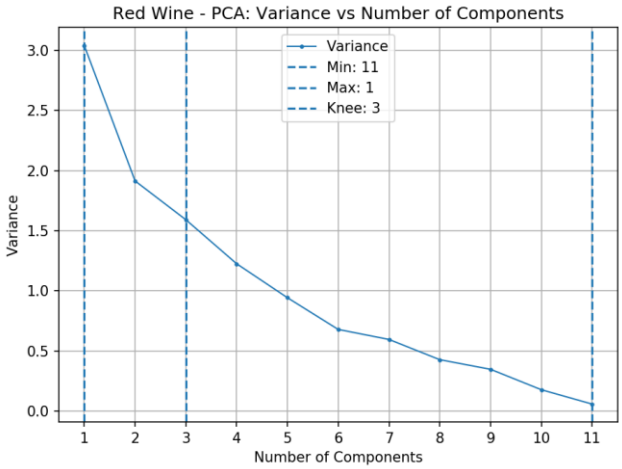

Fig. 10. Red wine quality PCA component variance

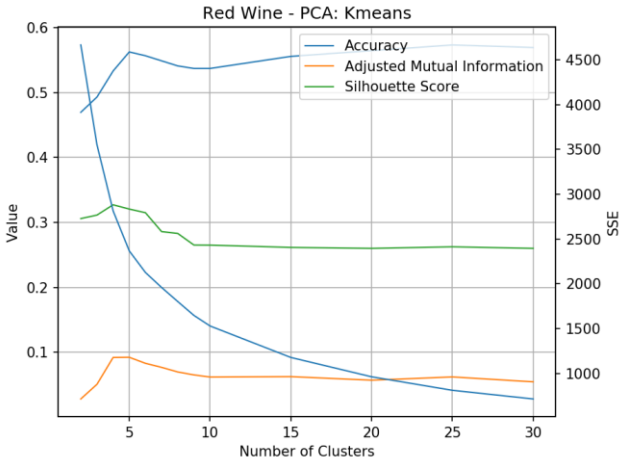

Fig. 11. Red wine quality 3 component PCA with k-means clustering

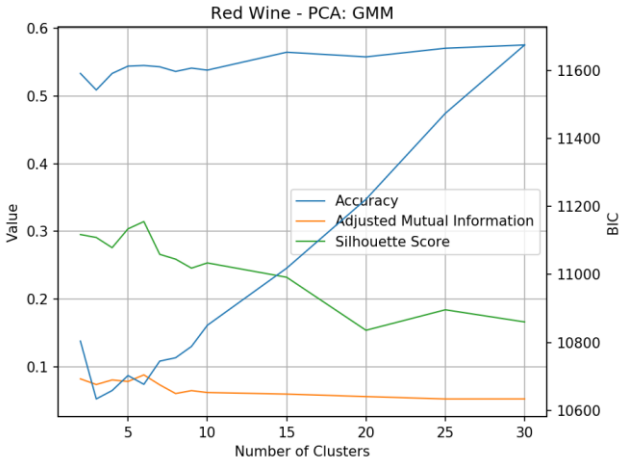

Fig. 12. Red wine quality 3 component PCA with GMM clustering

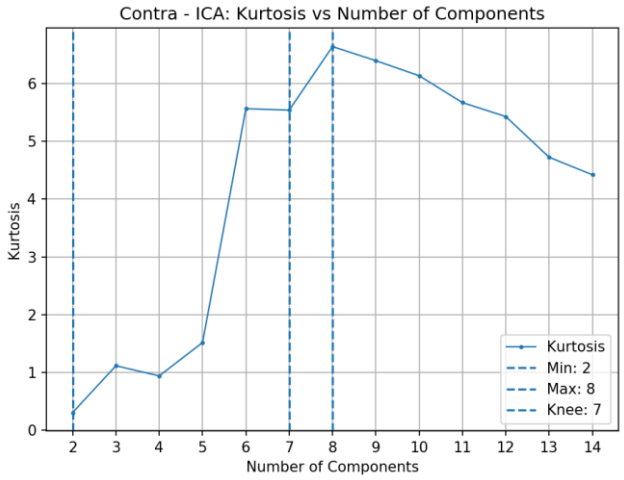

Fig. 13. Contraceptive use ICA components kurtosis

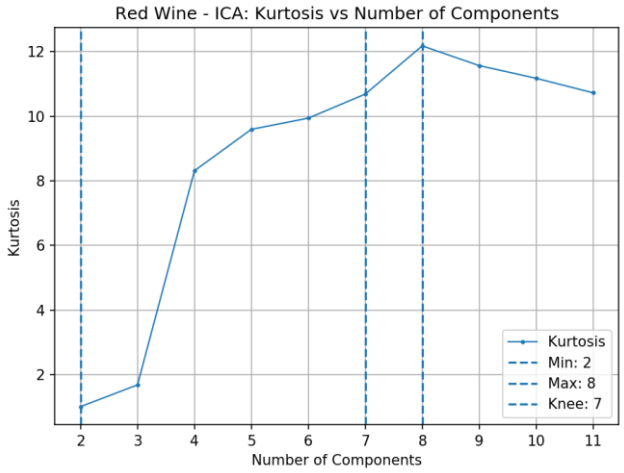

Fig. 16. Red wine quality ICA components kurtosis

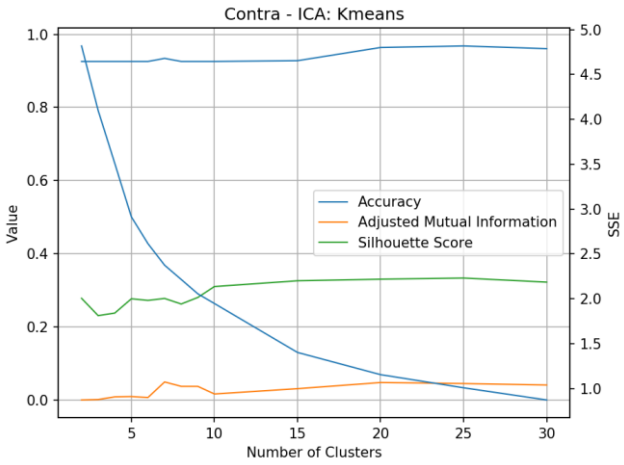

Fig. 14. Contraceptive use 7 component ICA with k-means clustering

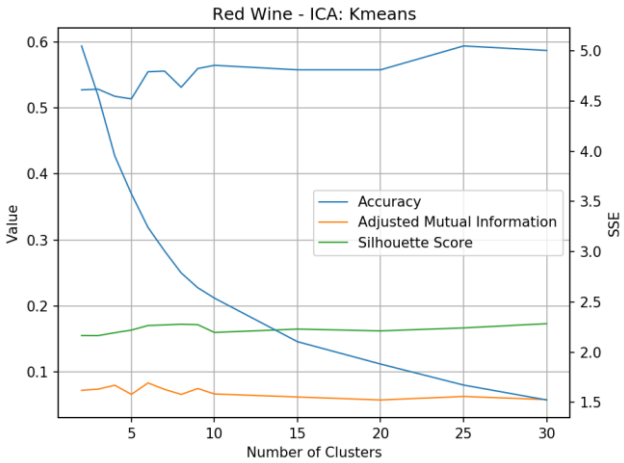

Fig. 17. Red wine quality 7 component ICA with k-means clustering

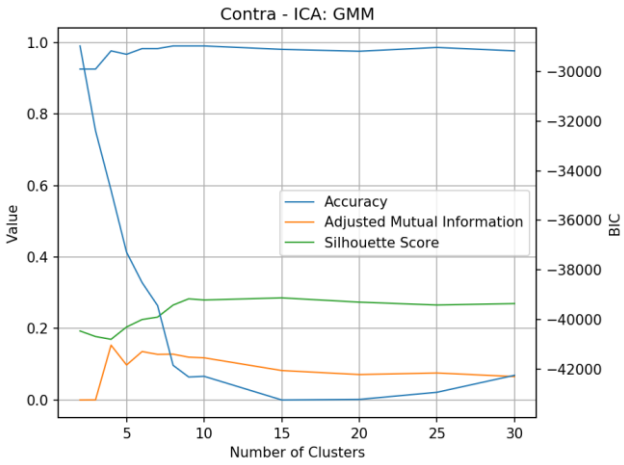

Fig. 15. Contraceptive use 7 component ICA with GMM clustering

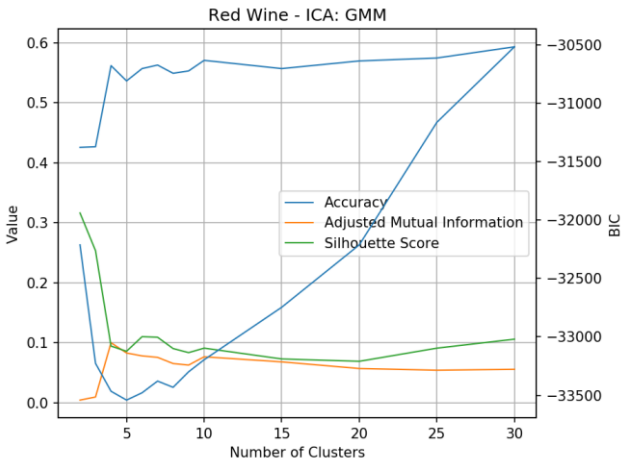

Fig. 18. Red wine quality 7 component ICA with GMM clustering

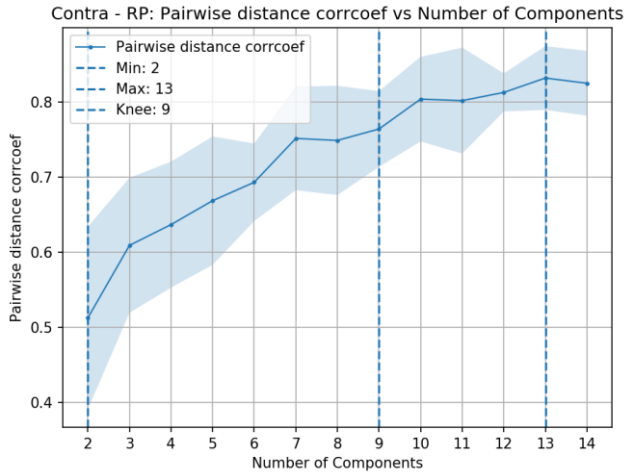

Fig. 19. Contraceptive use RP components pairwise distance correlation

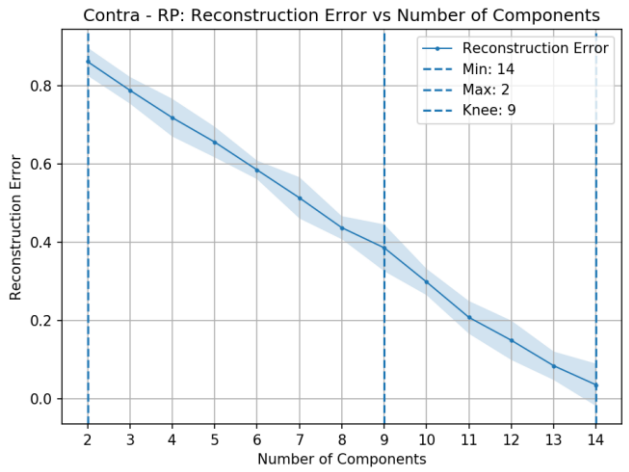

Fig. 20. Contraceptive use RP components reconstruction error

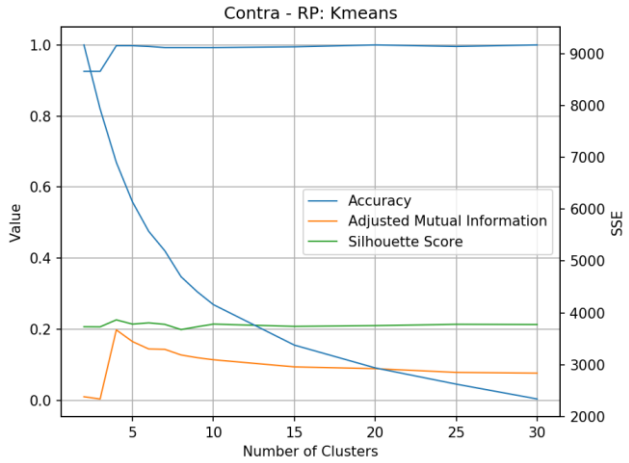

Fig. 21. Contraceptive use 9 component RP with k-means clustering

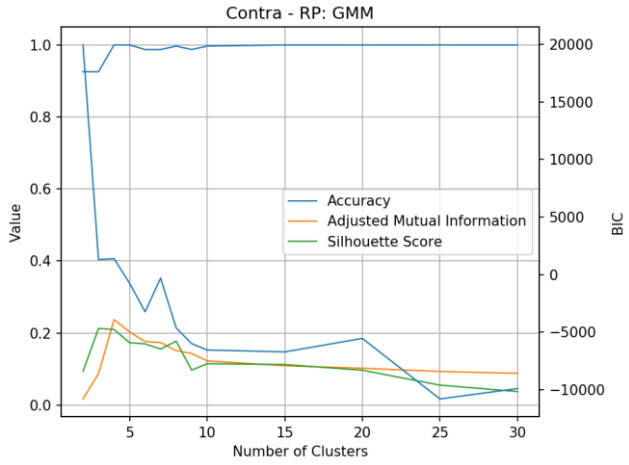

Fig. 22. Contraceptive use 9 component RP with GMM clustering

Fig. 16 shows the ICA for the red wine quality data. Here, seven components were chosen as well. Looking at the metrics for k-means clustering using DR, it appears that six clusters best captures the data (Fig. 17), while four appears to be best for GMM (Fig. 18). Again, these are similar to what we saw before.

The k-means metrics show a much lower SSE, a better clustering from the SS and similar accuracy. The GMM has a less complex model as shown by the BIC and a similar clustering and accuracy as the original data.

In ICA we can see its limits in DR. We have many more components in our ICA DR. It is mainly used to separate signals. It also appears that ICA performed better on the clustering than PCA. This implies that our data might have some overlapping attributes.

C. *Randomized Projections*

Randomized Projections (RP) is a simple yet powerful method of DR. The projections were chosen such that the error is controlled and they preserve distances. To implement this I used scikit learn's SparseRandomProjection method. To determine the appropriate number of components I looked to maximize the pairwise distance correlation, minimize the reconstruction error, and minimize the number of components needed by using the elbow/knee method on the graphs.

As seen in Figs. 19 and 20 for the contraceptive data nine components were chosen. It is a larger number of components to capture the properties of our data than the other methods, because of the randomness of the projections. We can also see from the confidence region around each line that the pairwise distance would vary greatly from run to run but the reconstruction error varied a little less.

After applying this DR to the contraceptive data it was then clustered using k-means and GMM. The same metrics as above were then plotted in Figs. 21 and 22. For k-means and GMM it appears that four clusters is the appropriate. This is fewer than was needed with non-DR data.

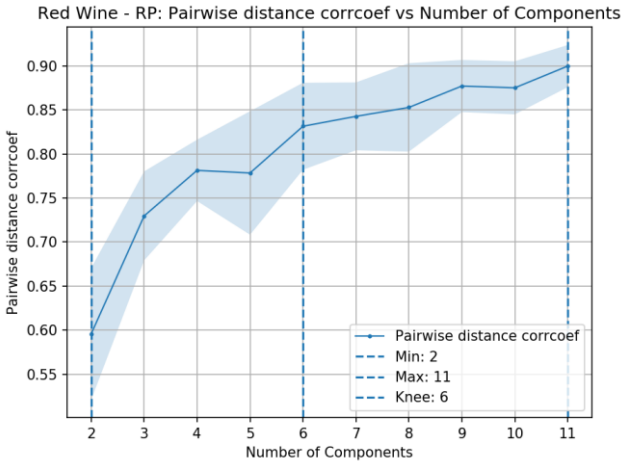

Fig. 23. Red wine quality RP components pairwise distance correlation

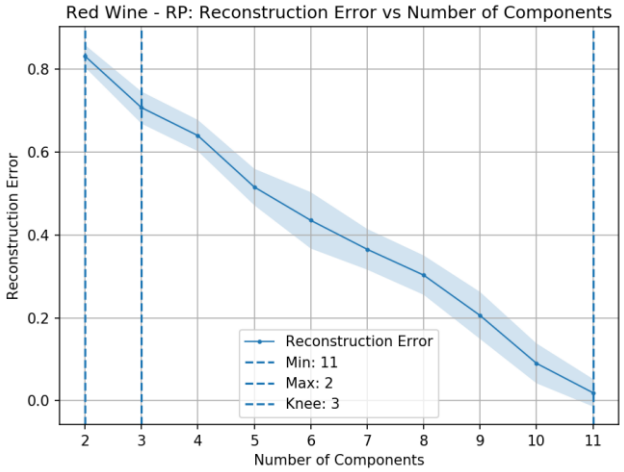

Fig. 24. Red wine quality RP components reconstruction error

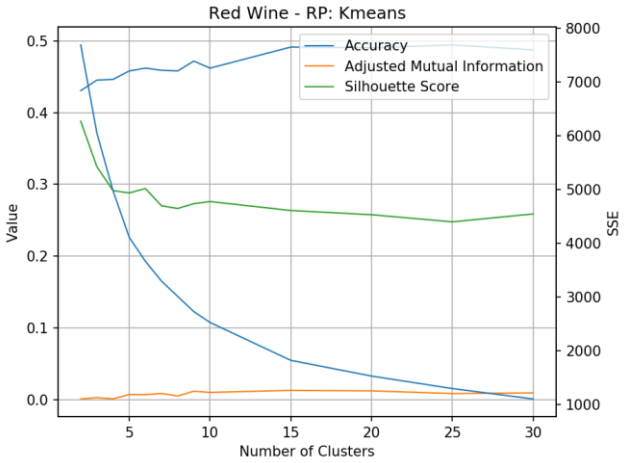

Fig. 25. Red wine quality 3 component RP with k-means clustering

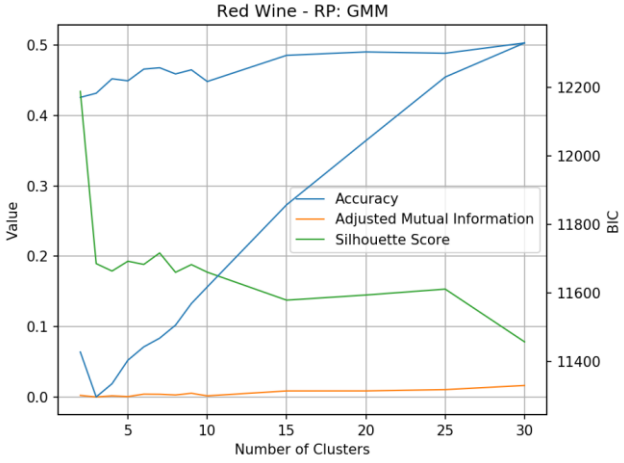

Fig. 26. Red wine quality 3 component RP with GMM clustering

Comparing the k-means metrics with the original data, we see that we now have a lower SSE and a higher accuracy. In addition, the SS shows the same quality of clustering was achieved using fewer clusters. Similarly for GMM we have a higher accuracy with similar quality of clustering achieved in fewer clusters.

Figs. 23 and 24 shows the RP for the red wine quality data. Here it was more difficult to determine the appropriate number of components. I decided on three since that is what was suggested by the reconstruction error, it was less complex, and the pairwise distance correlation also had a bit of a 'knee' there. This small changes could also be affected by the error or confidence region around each line.

Looking at the metrics for k-means clustering using RP, it appears that six clusters best captures the data (Fig. 25), while seven appears to be best for GMM (Fig. 26). These are more clusters than before but still within our rating scale domain.

The k-means metrics show a lower SSE, a much better clustering from the SS, however, the accuracy has declined significantly. The GMM performs similarly with lower accuracy but better clustering.

RP appear to help improve clustering by removing components from our dataset but do not appear to increase the accuracy of our clusters.

D. Random Forest (Information Gain)

Using Random Forests (RF) as a DR method takes advantage of their greedy algorithm of information gain. Inherently the algorithm searches for the attribute and split which provides the most information gain. We can use this to select the most important features and reduce our dimensions. I chose to do this visually by determining the elbow/knee in a graph of the feature importances, pulled from the scikit learn RandomForestClassifier method.

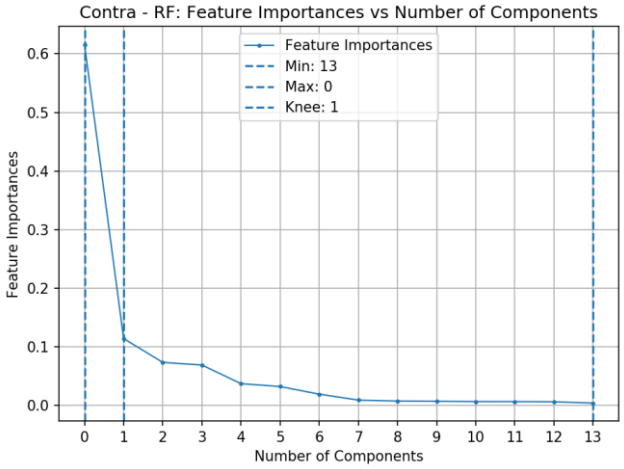

Fig. 27. Contraceptive use RF components feature importances

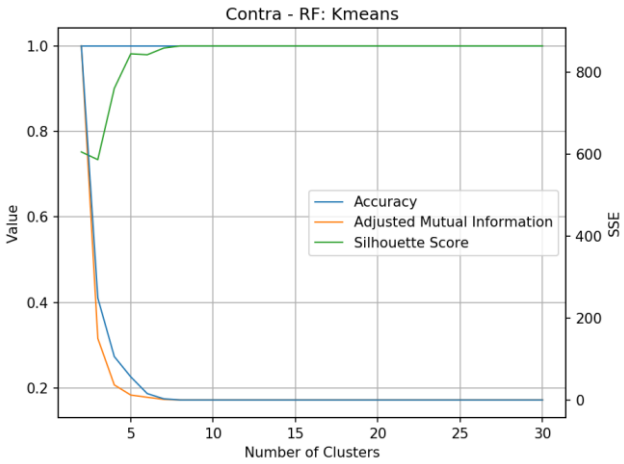

Fig. 28. Contraceptive use 2 component RF with k-means clustering

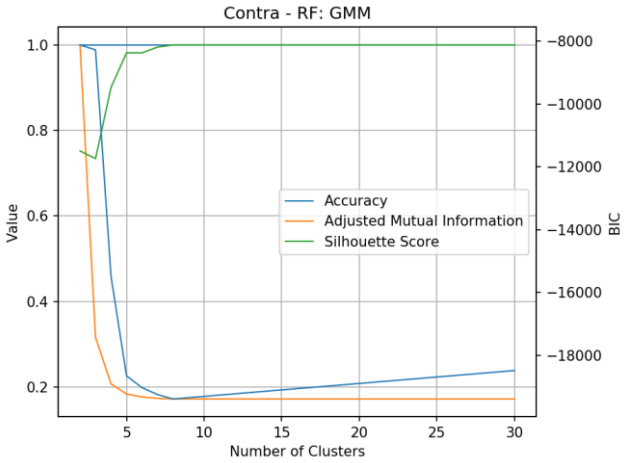

Fig. 29. Contraceptive use 2 component RF with GMM clustering

As seen in Fig. 27 for the contraceptive data two components were chosen (Note: the number of components axis was mistakenly labeled zero based). The first two components appear to be the most important.

After applying this DR to the contraceptive data it was then clustered using k-means and GMM. The same metrics as above were then plotted in Figs. 28 and 29. For k-means and GMM it appears that five clusters is the appropriate choice.

Comparing both the k-means and GMM metrics with the original data, we see that we now have a lower SSE, a higher accuracy, and a better clustering of the data.

Fig. 30 shows the RF for the red wine quality data. Here, four components were chosen. Looking at the metrics for k-means and GMM clustering, it appears that four clusters best captures the data (Figs. 31 and 32).

Both k-means and GMM metrics show a lower SSE, a better clustering from the SS and a higher accuracy.

RF/Information Gain for both datasets had fewer clusters than the non-DR data. RF appear to be the best performing DR of the four explored. The collection of greedy decision trees appears to encapsulate the most important information in the data.

V. Neural Network Analysis

A. *Neural Network using Dimensionality Reduction*

For the neural network analysis I chose to work with the contraceptive use dataset. The original neural network from assignment 1 had a training accuracy of 0.561 and a training time of 11.2811 seconds (Table I). This will be the standard each will be compared to.

For each DR method, I originally used the same neural network parameters as assignment 1, however, the training accuracy was always 0.333, or exactly a random choice with the ternary labels of the dataset. Therefore, I chose to optimize a neural network for each DR method.

I used a grid search with cross validation on each DR method with the following parameters:

- Alpha: log scale from 1e-5 to 1e-1,
- Hidden layers: 1 or 2,
- Nodes per layer: 10, 20, 30, or 40,
- DR components: 2 to max

TABLE I. Contraceptive Use Neural Networks with Dimensionality Reduction

Neural Network	Training Accuracy	Training Time	Alpha	Hidden Layer Sizes	Number of Components
Original	0.561	11.2811	0.001	(10, 10)	14
PCA	0.998	0.1376	0.1	(30,)	10
ICA	0.683	0.1024	0.1	(10,)	12
RP	0.999	0.1498	0.1	(30,)	7
RF	1.000	0.3860	0.1	(30,)	2

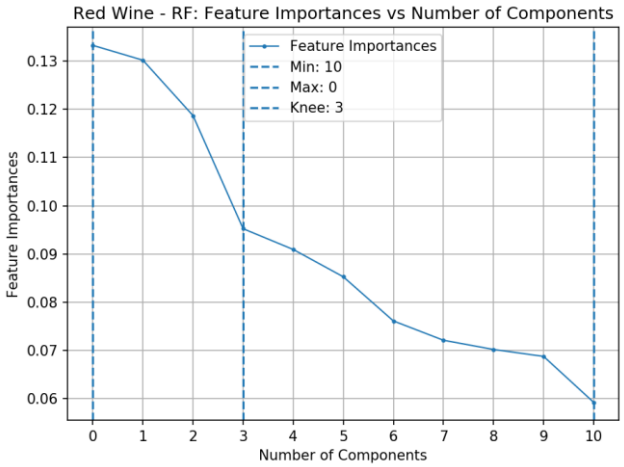

Fig. 30. Red wine quality RF components feature importances

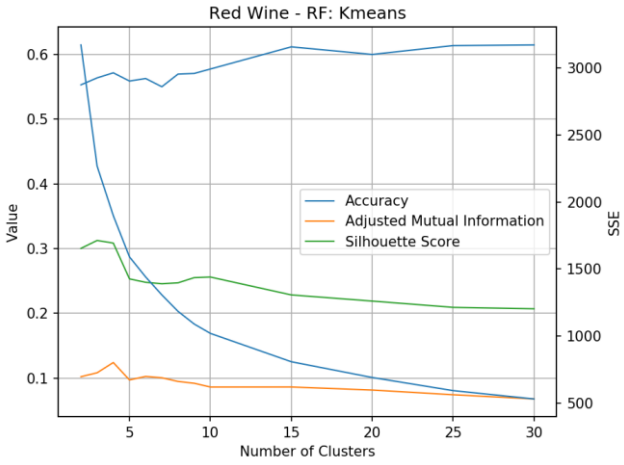

Fig. 31. Red wine quality 4 component RF with k-means clustering

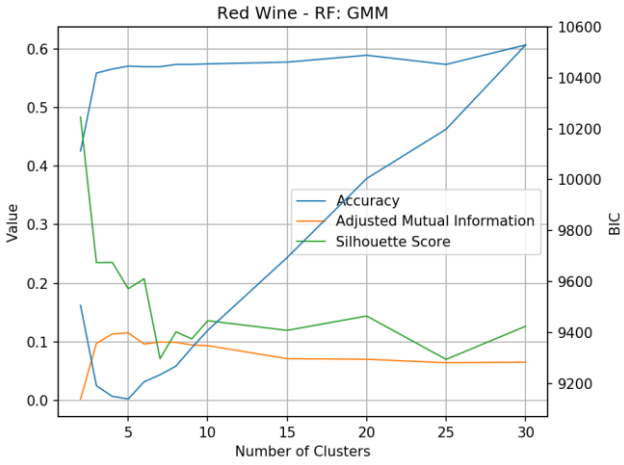

Fig. 32. Red wine quality 4 component RF with GMM clustering

Table I shows that all DR methods had a higher accuracy than the original model and were a couple of orders of magnitude faster. Unsurprisingly, ICA needed the most components and RF needed the fewest. Since, as discussed already, ICA does not have a goal of DR and RF maximize the information gain. Also, similar to the clustering, ICA had the lowest accuracy.

Based upon speed and accuracy, PCA and RP appear to be the best choices, with RF close behind. This shows how powerful RP can be, which is not intuitive.

B. Neural Network using Clustering

For this part of the assignment I first used the four DR methods as before on the contraceptive use dataset. I then clustered the data using both clustering methods and fit a neural network using these clusters as attributes.

Again I used a grid search with cross validation on each DR method and also used it to search for the best number of clusters. Here were the parameters:

- Alpha: log scale from 1e-5 to 1e-1,
- Hidden layers: 1 or 2,
- Nodes per layer: 10, 20, 30, or 40,
- Clusters: [2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30]

I compared the results to the original neural network from assignment 1 and the results are shown in Table II. Both PCA and RF are completely accurate, fast, and work for multiple parameters. RP is close behind in accuracy and time but uses specific parameters that work. ICA is again out-performed. The ICA with k-means actually performs worse than the original neural network.

The large increase in accuracy makes me question the validity of these findings and more specifically, whether or not we are still answering the original question or something different.

TABLE II. CONTRACEPTIVE USE NEURAL NETWORKS WITH DIMENSIONALITY REDUCTION CLUSTERS AS ATTRIBUTES

Neural Network	Training Accuracy	Training Time	Alpha	Hidden Layer Sizes	Number of Clusters
Original	0.561	11.2811	0.001	(10, 10)	14
PCA (k-means)	1.000	0.0560	Many	Many	Many
PCA (GMM)	1.000	0.0407	Many	Many	2/3/4/25
ICA (k-means)	0.428	0.3050	0.1	(40, 40)	25
ICA (GMM)	0.815	0.2460	0.1	(30,)	15
RP (k-means)	0.987	0.3502	0.0001	(40, 40)	20
RP (GMM)	0.988	0.1798	Many	(30, 20)	8
RF (k-means)	1.000	0.0560	Many	Many	Many
RF (GMM)	1.000	0.0407	Many	Many	2/3/4/25

VI. Discussion

RF/Information Gain appears to be the best DR method for clustering while ICA performed the worst and did not reduce the number of components as much as the others. DR reduced the number of clusters need for the contraceptive use dataset and brought it into line with our ternary labels. The red wine dataset had a similar number of clusters, also appropriate for our number of labels.

For neural networks RF, PCA, and RP all performed well and much quicker than the original network. However, more exploration needs to be done to completely understand the relationship between this DR/Clustered neural network and the original one from assignment 1. Again, ICA did not perform as well. I believe that ICA has an appropriate use for signal separation, but that is not useful for my data in these circumstances.

DR appears to be very useful in clustering and simplifying large neural networks. Another option to be explored in the future would be adding the clusters from the data as additional attributes to the dataset. This may affect the correlation and independence of the data, but could also prove useful.

VII. References

[1] Lozano, L., García-Cueto, E. and Muñiz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology*, 4(2), pp.73-79.