A Comparison of Unsupervised Learning and Dimensionality Reduction Techniques

Chapman Siu

1 Introduction

Pima Indian Diabetes

The Pima Indian Diabetes is a useful dataset to examine the health conditions which may be leading indicators about how susceptible one is to diabetes in the future. The full data set available from the R package mlbench was used in this analysis.

Wine Quality

The Wine Quality dataset used in this analysis is a subset of the Wine Quality dataset available from the UCI repository index [3]. Here we examine a subset of the data set containing only red wines with scores 5, 6, or 7 and attempt to construct a classifier those wines since all other labels appear to be outliers.

Comparisons between Pima Indian Diabetes and Wine Quality

Almost all variables in the Pima Indian Diabetes data set are positive correlated with each other, with only a few variables only slightly negatively correlated.

Through research we can in fact determine that some variables (for example, insulin and glucose in Pima Diabetes data set and “free sulfur dioxide”, “total sulfur dioxide”) are interdependent on each other.

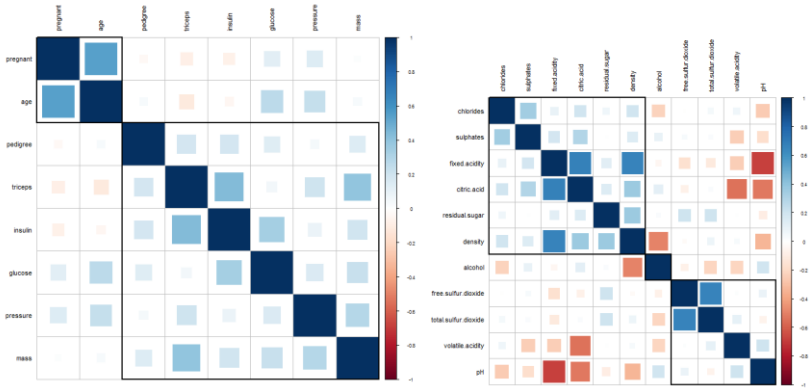


Figure 1 - The correlation matrix of the two data sets. Pima Indian Diabetes data set is on the left and Wine Quality data set is on the right. Many variables in the Pima Indian Diabetes are positively correlated with each other

2 Application of Dimension Reduction

2.1 Pima Indian Diabetes

Examining the Scree plot (on the next page) for Principal Component Analysis (PCA), we notice the "elbow" of the plot forms at six components. The biplot which shows the relationship between the first two components solidifies the close relationships between the variables as shown in the correlation matrix above.

Applying ICA to the data set and examining the kurtosis, we see that the there is one vector which appears to be Gaussian.

Applying Random projection (RP), the L2 norm of the data reconstruction increases drastically if we have less than 8 columns. This result is expected and reflects the Johnson-Lindenstrauss lemma. Since the number of components we have is very low, there is not a suitable lower bound of variables which we can reduce within a small error epsilon. From the graphs, it is clear that different reconstructions can differ by over 10^7 when the number of components is less than 8, even when just comparing the 95th and 5th percentile.

Finally, the filter method of feature selection used through deriving variable importance scores from a random forest model (RF). This filters the variables before the application of the various learning algorithms. In this instance I have chosen to filter and keep the top six variables by variable importance.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	# non-Gaussian components
Pima Indian	0.4839	1.0340	3.7070	3.9270	5.8110	9.2450	1

2.2 Wine Quality

Similar to the Pima Indian data set above, the wine quality data set can be approached in a similar way. From the Scree plot and the biplot it is clear that the various components are mostly orthogonal with each other. This is in line with the correlation plot shown in section 1. Hence for this data set, PCA will be used as a rotation rather than reducing the dimensions.

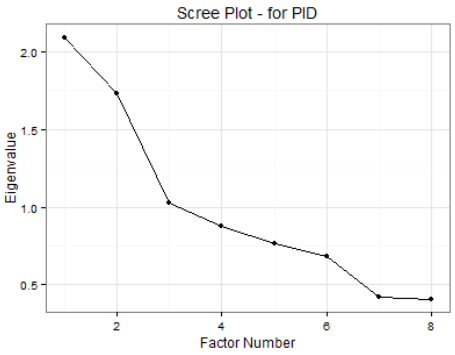
Using ICA and examining the kurtosis of the resulting components removes two components which appear to be Gaussian.

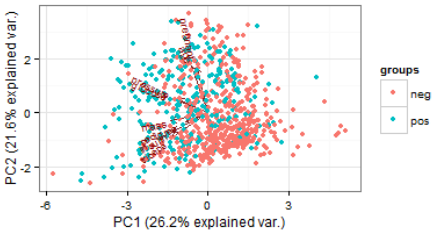
Again using random projections to reduce dimensions does not appear to be suitable, since the L2 norm reconstruction difference is too large (roughly 10^6) for components less than 11.

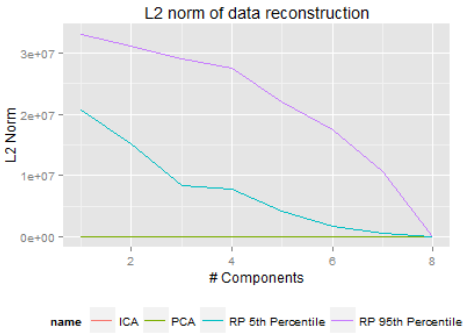
Finally using RF, shows that the mean gini decrease begins to flatten out when we get to the variable "citric acid", hence the bottom four variables will be removed in this analysis.

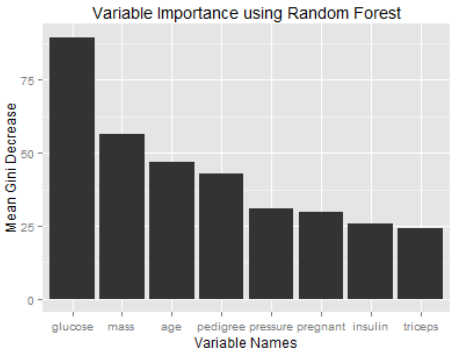
	Min	1st Qu.	Median	Mean	3rd Qu.	Max	# non-Gaussian components
Wine Qu.	0.2779	2.3570	2.9530	11.2400	15.6900	42.090	2

Pima Indian Diabetes Plots

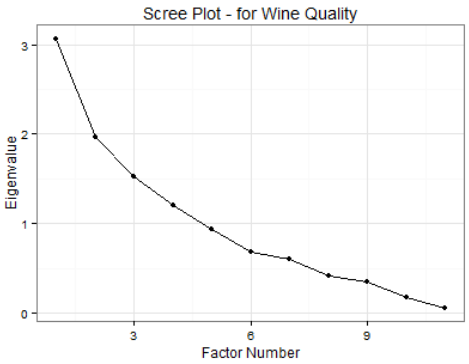


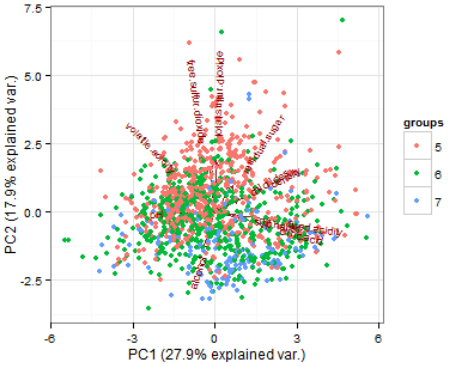


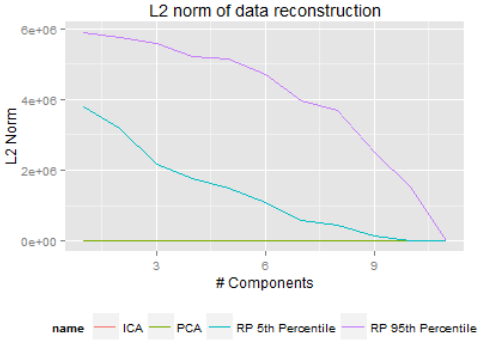


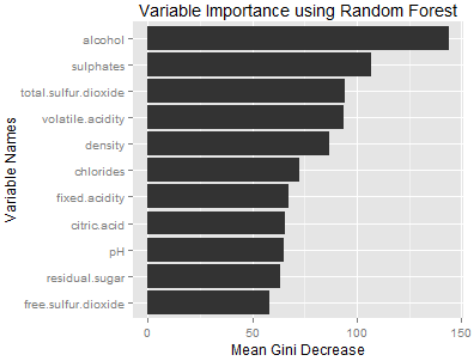


Wine Quality Plots









Top left to bottom right: PCA Scree plot, Biplot, L2 norm of data reconstruction, and Mean Gini Difference using Random Forests

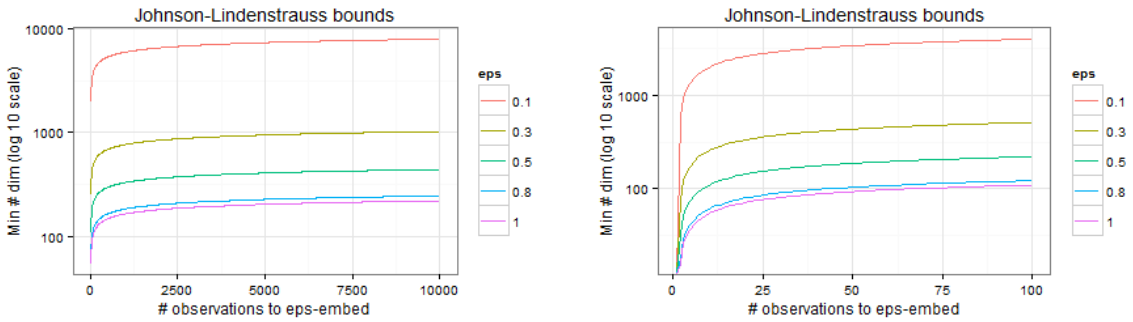


Figure 2 - Even with 100 features and epsilon = 1, we still need over 100 features to meet the Johnson-Lindenstrauss bounds, thereby not reducing the dimension.

3 Unsupervised Learning Algorithms

Two unsupervised learning algorithms were considered. K-means and Expectation maximization (EM) algorithm. To determine the optimal clusters for these two algorithms, a clustergram was used for K-means, whilst EM was solved analytically by maximizing the Baysian information criterion for each combination of type of mixture and number of clusters. For information on the various model labels and their references please refer to the R mclust documentation.

3.1 Pima Indian Diabetes

For the clustergram we choose the cluster number where the points are most spread out. For example, in the situation where there is no transformation, when we arrive at six clusters, we are actually essentially left with five clusters as indicated in the plot, hence we would choose the optimal number of clusters to be five. In a similar manner, the ideal number of clusters for each set of points can be chosen.

Data Set	Ideal # of Clusters (Kmeans)	Accuracy against labels (Kmeans)	Ideal # of Clusters (EM)	Accuracy against labels (EM)
No Transformations	5	0.7121	6	**0.6746**
PCA	3	0.6746	4	*0.6603*
ICA	2	0.6731	7	0.6719
RP	3	*0.6510*	4	0.6628
RF	3	**0.7461**	8	0.6694

Since the no information rate of the Pima Indian data set is 65%, the using the binomial test, the resulting p-values range from 0.91-0.95, suggesting that the model accuracies produced by the best clusters are no better than using the no information rate provided by the data.

3.2 Wine Quality

The same approach can be table for the wine quality data set. The results are summarized below:

Data Set	Ideal # of Clusters (Kmeans)	Accuracy against labels (Kmeans)	Ideal # of Clusters (EM)	Accuracy against labels (EM)
No Transformations	5	*0.4918*	7	0.5498
PCA	3	**0.5747**	8	**0.5510**
ICA	2	0.4352	10	*0.5004*
RP	6	0.4978	4	0.5024
RF	4	0.5050	13	0.5052

The no information rate for the Wine Quality data set is 51%, the p-values for the binomial test range from 0.6-0.8, again suggesting that the model accuracies produced by the best clusters are no better than using the information rate provided by the data.

3.3 Cluster Performance

To compare the two algorithm's cluster performance against different feature reduction techniques, the number of clusters for each algorithm was fixed at three. The colors of each plot refer to the particular labels of each of the data set.

For the Pima Indian data set looking at the diagonals on the plot below, in general the clusters for each algorithm do a fair job of splitting the two labels. In the Kmean plots all transformations appear to be loosely related to each other, whilst in the EM plots, PCA and ICA almost match each other perfectly, and there appears to be some similarity in the clusters derived from RF, with PCA and ICA.

Figure 3 - Comparison of Cluster performance for Pima Diabetes data set, on the left is Kmeans and right EM

With the wine quality data set, none of the transformations appear to conform to the labels, this aligns with the numbers in the tables above. On the cluster comparison standpoint, PCA and ICA almost match each other perfectly within the EM plots, whilst there does not appear to be much relationship between any of the other transformations within EM or Kmean clusters.

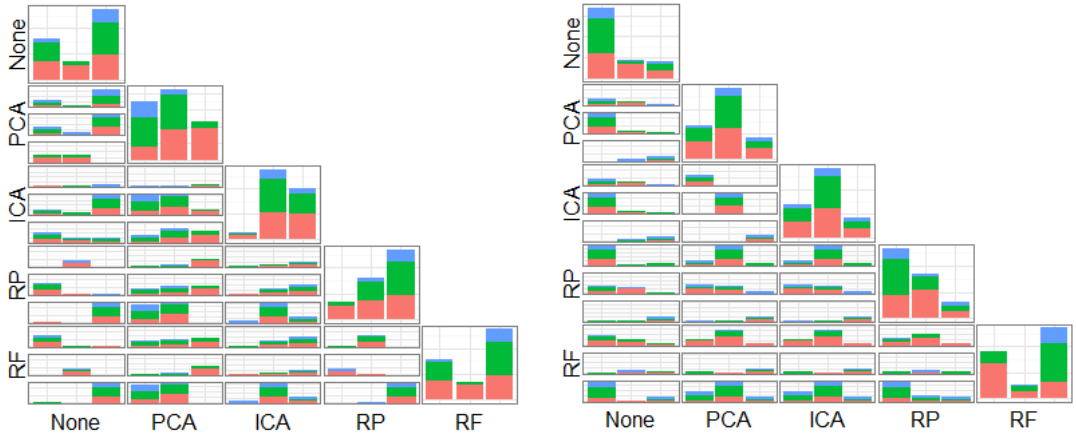


Figure 4 - Comparison of Cluster performance for Wine Quality data set, on the left is Kmeans and right EM

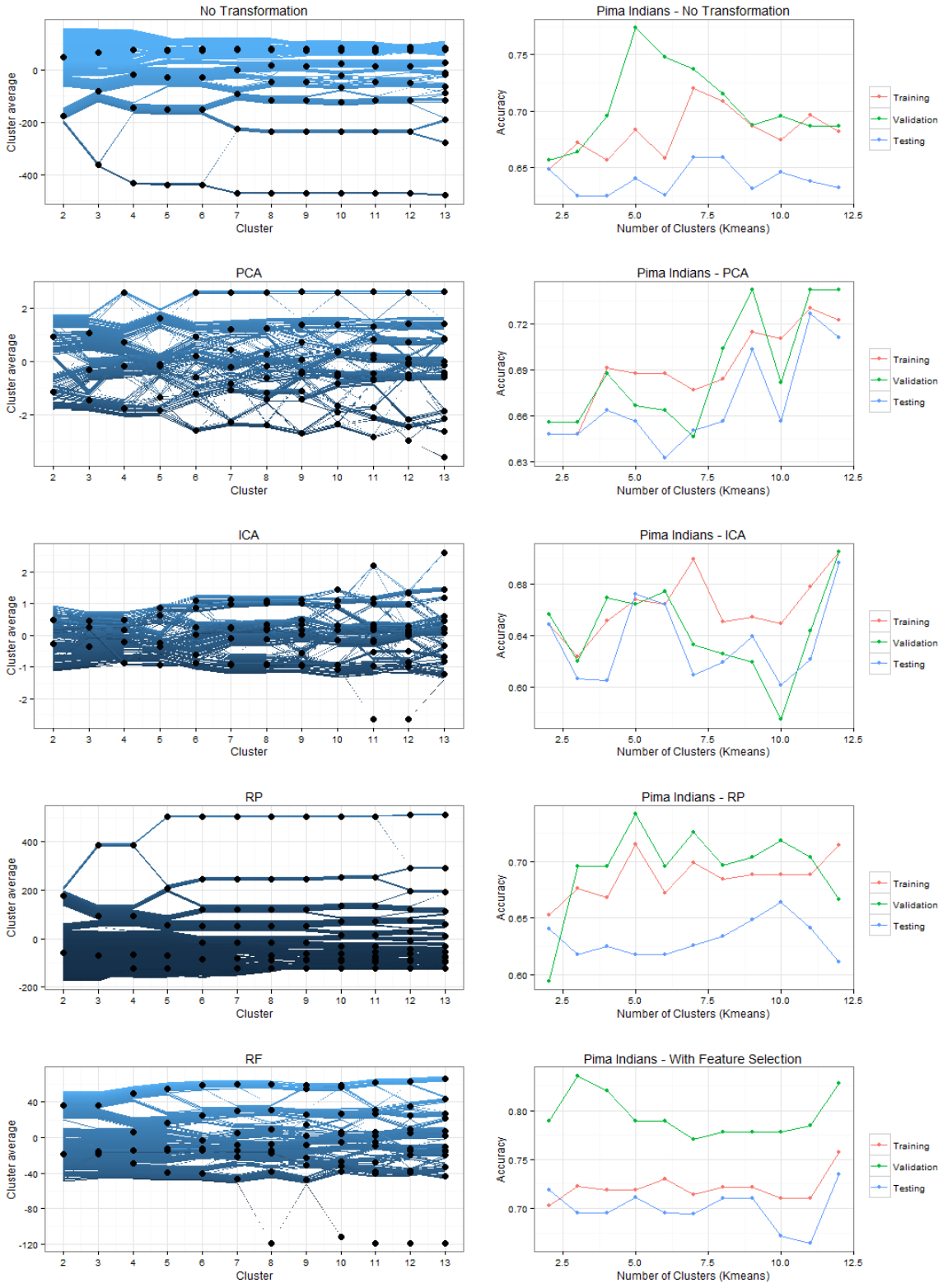


Figure 5 - Pima Indians - Kmeans - Optimal Clusters

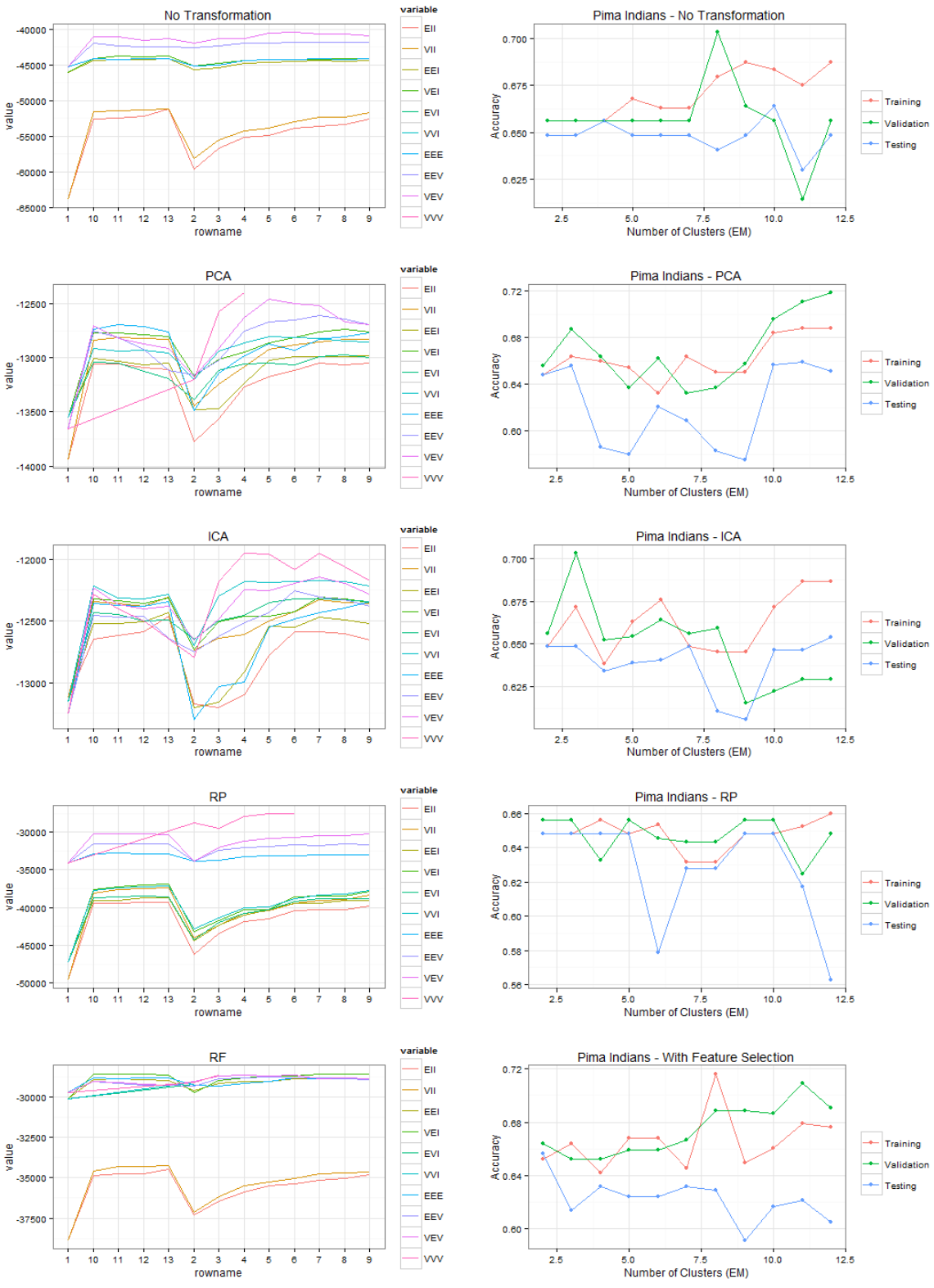


Figure 6 - Pima Indians - EM Optimal Clusters

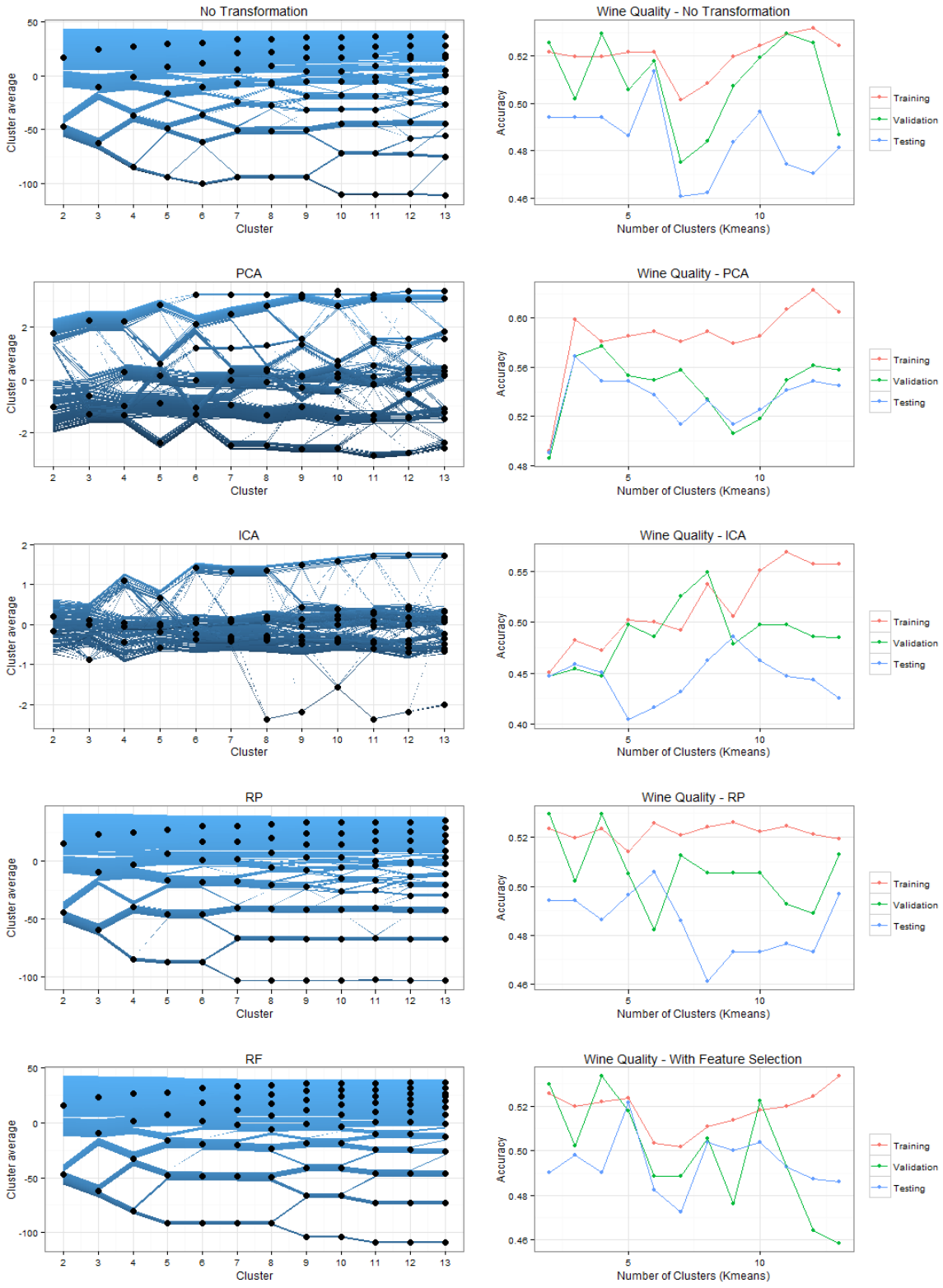


Figure 7 - Wine Quality - Kmeans - Optimal Clusters

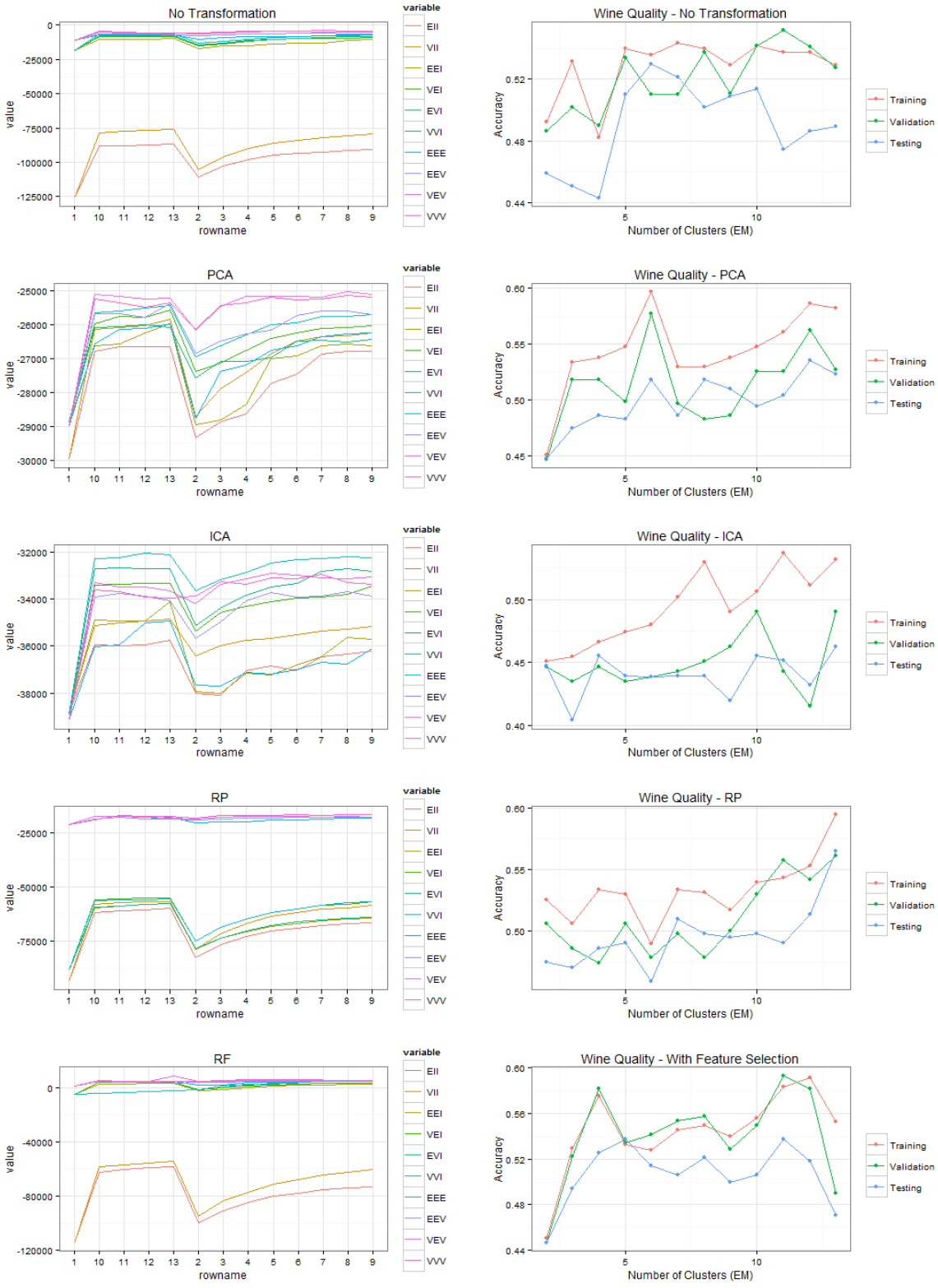


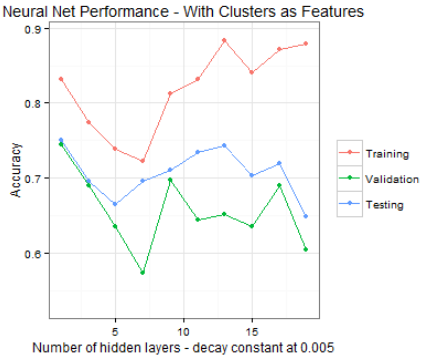
Figure 8 - Wine Quality - EM - Optimal Clusters

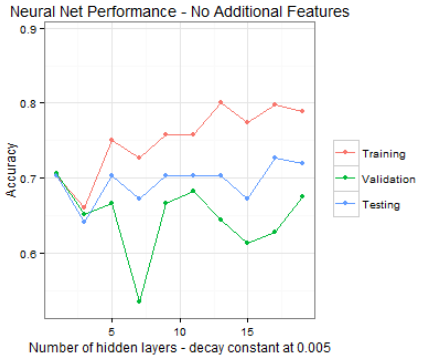
4 Neural Networks

R's caret library was used to fit neural networks for each of the data set. The data set was split into training (70%) validation (15%) and testing (15%) portions. Hyperparameters were optimized through exhaustive grid search, with each model training used 10 fold cross validation, with the best set of parameters chosen based on validation set performance. Each data set was trained twice, once with no changes, and the other time by adding the results of the cluster as features; adding an additional five features to the data set.

4.1 Pima Indian

The optimal set of parameters was when decay was set at a level of 0.05, and the training, validation and testing performance is graphed below. With the range of the **Accuracy** axis the same in both models, it is clear that the model with additional features performs stronger. Unsurprisingly, adding additional features meant longer training times. for the Pima Indian data set the training time increased by roughly 60%.

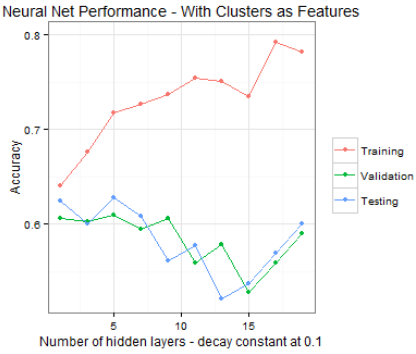


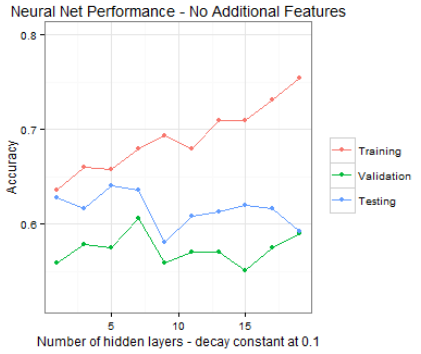


Time to train neural network with additional features	3 min 7 seconds
Time to train neural network with no additional features	1 minute 52 seconds

4.2 Wine Quality

The neural network model for Wine quality data set was optimized and chosen in the same way as the Pima data set. Comparing the accuracy curves in the two graphs below shows similar performance compared with the Pima Indian dataset which was consistently higher. For the wine quality data set, adding the additional features increased training time by roughly 30%.





Time to train neural network with additional features	4 min 54 seconds
Time to train neural network with no additional features	2 minute 59 seconds