Unsupervised Learning and Dimensionality Reduction

Varun Dani (vdani3)

1. Datasets:

1.1. Landsat Satellite (Statlog)

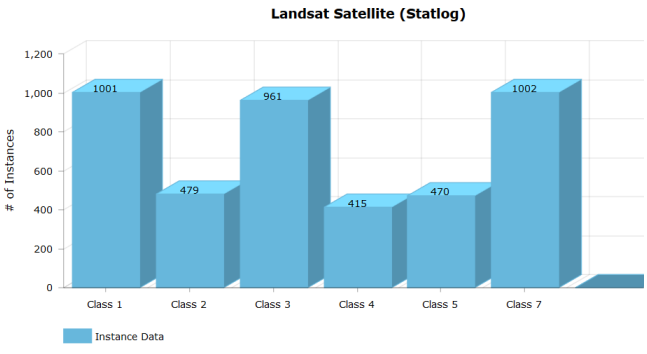
This dataset is image pixel data generated from Landsat Multi-Spectral Scanner and consists of the multi-spectral values of pixels in a satellite image(NASA). *Classification Problem*: The aim is to predict this classification associated with the central pixel in the 3x3 matrix, given the multi-spectral values. This will predict given image pixel is associated with which type of land. [from red soil (1), cotton crop (2), grey soil (3), damp grey soil (4), soil with vegetation stubble (5), or very damp grey soil (7)] Dataset contains 36 attributes, 6 decision classes, 4435 training instances and 2000 test instances.

1.2. Balance Scale Weight & Distance

This dataset is generated to model psychological experimental results of an attempt that characterize and explain developmental differences in children's thinking, specifically in their understanding of balance scale problems. *Classification Problem*: Given attributes right-balance, right-weight, left-balance, left-weight we should identify whether the result is Balanced(B), Right-Aligned(R), or Left-Aligned(L). The Dataset contains 4 attributes, 3 decision classes, 625 training instances and 289 test instances.

1.3. Why Datasets are Interesting?

Following graphs shows how both datasets are distributed among decision classes and generate interesting insights from distribution.



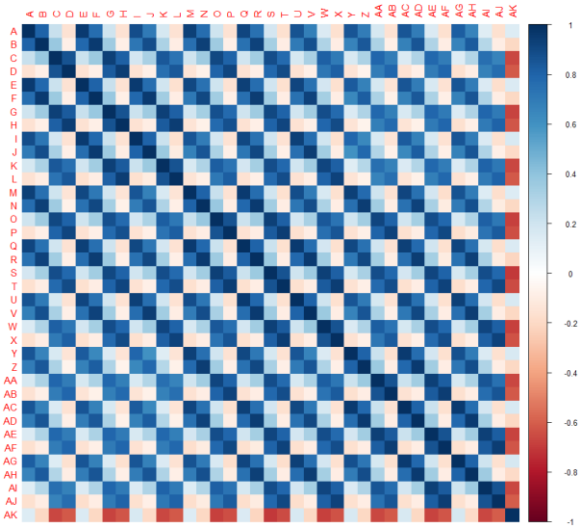


Figure 1: Distribution of Landsat and attribute correlation (Statlog) (design tool AmCharts Ref: README: Section-M-1)

If we observe above Landsat testing Data, it is distributed almost uniformly across decision classes. Large number of instances (4435) enable algorithms to train model more precisely. Basic intuition for choosing this dataset that it will give almost near values for training and testing accuracy because of uniform distribution of data. Supervised learning algorithm will perform better for this dataset for training and testing set. We can analyse from above that, each attribute related strongly with fourth attribute in grid because it is data of 3x3 grid spectral values and neighbourhood values are mostly tends to same classification of image other than far values.

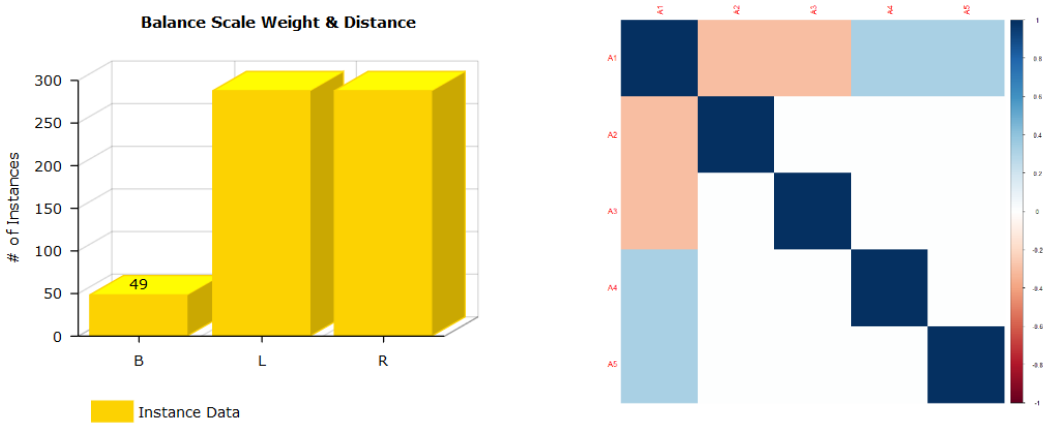


Figure 2: Distribution of Balance Scale Data and attribute correlation (design tool AmCharts)

In Contrast, if we observe Balance Scale data, it has 4 attributes and only 625 instances. This number is very low in comparison to Landsat dataset. Because of the small number of instances, accuracy will be very low, and we will not be able to reach accuracy more than some threshold limit for this dataset (Blumer Bound). Also, major observation to note here is very less number of instances for Class B(Balanced) will lead to significant degradation in TP and FP for that class in testing data. Observation from the correlation matrix, A4 and A5 are prominent features for the dataset. Also, A2 and A3 are negatively related to class and will degrade the performance of overall accuracy.

2. Clustering Algorithms: Following are the analysis of K-means and Expectation Maximization(EM) on both dataset followed by the comparison of both algorithms.

2.1. K-Means

K-Means is a simpler version of the clustering algorithms. it randomly selects cluster centers. It assigns each point to one center based on distance matrix and re-evaluates centers to minimize error. K-Means iteratively repeat this process of cluster assignment and center calculations till convergence. For both datasets, best cluster centers are chosen from 25 random generated center sets with Euclidian distance as a distance measure for pixels and weights.

Optimal selection of k:

Elbow method is used to calculate percentage of variance explained as a function of the number of clusters for optimal k selection. The second method is silhouette plot. This plot indicates for which value of k, silhouette value is higher for objects and that value indicates optimal k. The silhouette value is a measure of how similar an object is to its own cluster.

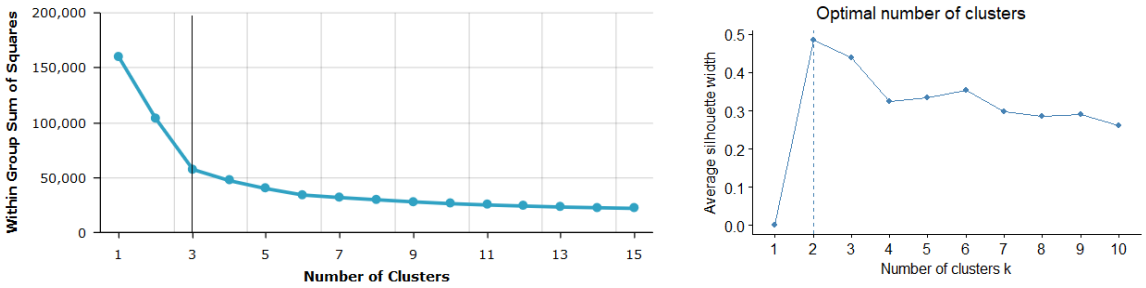


Figure 3: Elbow Plot and silhouette Plot (landsat)

Dataset 1 (Landsat):

Figure 3 shows Elbow plot for Landsat with the sudden decrease after value k=3. and silhouette plot shows value k=2 as optimal. Why optimal less then labels? because 6 classes of land can be further classified into 3 or 2 categories according to type of land.

cluster analysis:

From silhouette plot, I have picked top three values of k and performed cluster analysis. The reason for highest values are 2,3 and 6 is: clusters aligned with original data and 6 types of lands can be further classified into 3 and 2 categories based on similarities. If we observe confusion matrix in figure 5, for k=3 and k=6 most of the points are well defined into clusters except cluster 3 (k=3) and cluster 2(k=3). Figure 5 shows feature contribution for all combination of clusters. Also, because of random point selection, clusters are vertical instead of horizontal (opposite of point projection in the figure), this resulted in low accuracy because a substantial chunk of one cluster is identified as another.

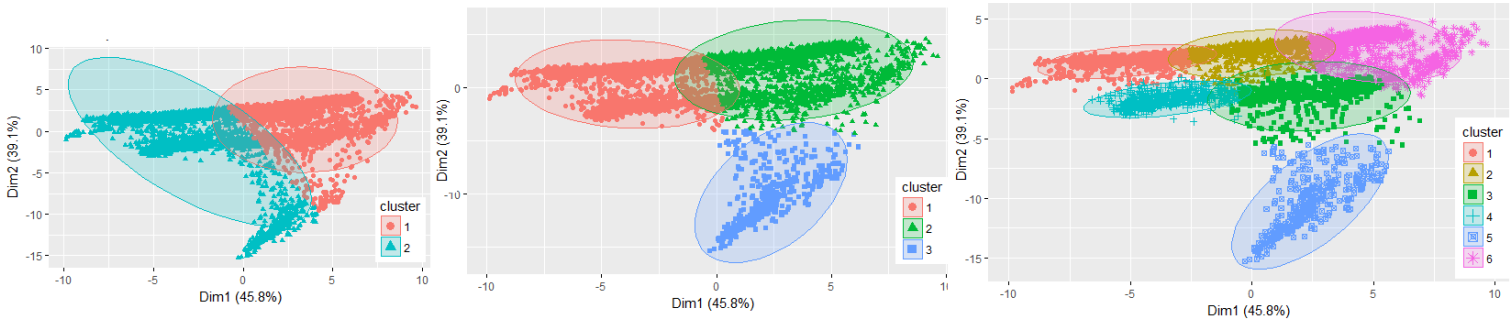


Figure 4: Clusters for k=2, k=3, k=6 (landsat)

	1	2	3	4	5	6
1	708	14	954	204	39	85
2	363	52	7	211	429	953
3	1	413	0	0	2	0

	1	2	3	4	5	6
1	20	0	874	68	0	13
2	13	8	75	317	39	319
3	405	85	0	8	98	2
4	625	1	11	1	24	0
5	0	382	0	0	0	0
6	9	3	1	21	309	704

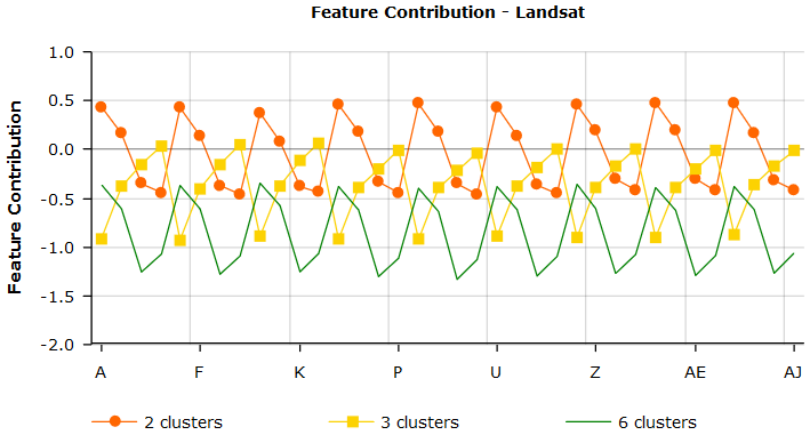


Figure 5: Confusion Matrix and Feature contribution (landsat)

Dataset 2 (Balance Scale):

Figure 6 shows Elbow plot for Balance Scale with the no clear value of k this indicates the critical separation of data according to classes. Also, silhouette plot shows value k=8 as optimal where total classes are 3 only. Why? It might be because this data can be well defined in higher dimensional space of 8 dimensions.

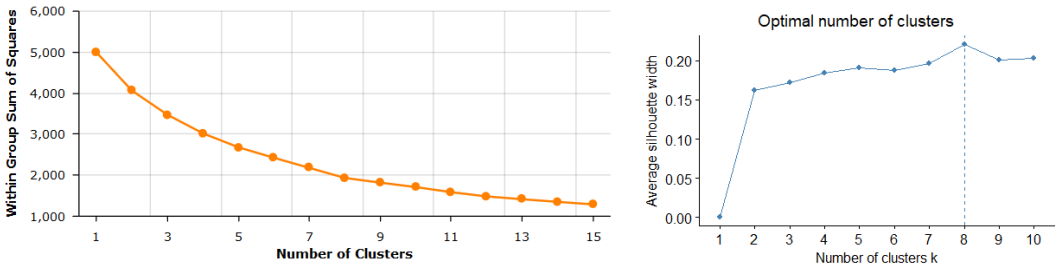


Figure 6: Elbow Plot and silhouette Plot (Balance scale)

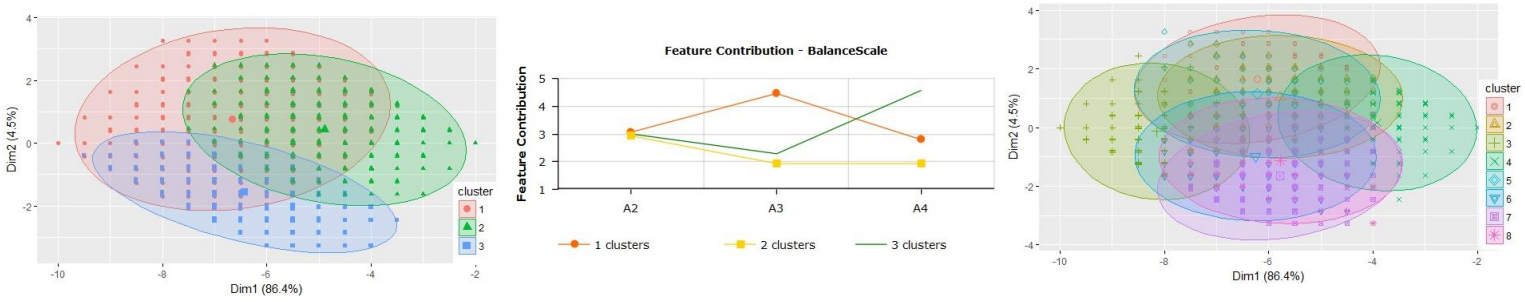


Figure 7: Clusters k=3, k=8, Feature Contribution (Balance scale)

cluster analysis:

From silhouette plot, I have picked top two values of k (k=3, k=8) and performed cluster analysis. it is clear that clusters do not line up with original class and accuracy to class correctly degraded to 22.67%(very low). Why it's not aligning? If we see from data distribution from Figure 7, there is lack of instances to successfully classify class "B" and this leads to wrong assignments of the cluster. Also, for k=8, k-Means tries to define classes in higher dimensions but also failing on that perspective. From the Feature contribution plot, we can see that only A4 feature contributes towards classification.

2.2. Expectation Maximization

EM finds k probability distribution of data such that log-likelihood of data, given distribution, is maximized. EM has 2 steps: 1. performing an expectation (E) and 2. maximization (M) step. In detail, estimating the log-likelihood for the current estimates for the parameters and maximizing the likelihood found on above step. I have used model-based clustering technique and library for performing EM.

Optimal selection of k:

I have used Bayesian information criterion (BIC) plot for finding optimal k. If the value of BIC is lowest that value of k is preferred for model selection. This plot gives an estimation of information loss by calculating likelihood function. This measurement takes more time to find out optimal, but we can see from after results that it is more accurate than K-Means.

Dataset 1 (Landsat):

Figure 8 shows a plot of BIC values w.r.t number of components for two different models. First is equal volume ellipsoidal model (VEV) and other is varying volume ellipsoidal model. From the analysis, observed drastically increased after k=3 and highest on k=5. I have analyzed both cases (highest and lowest) in the subsequent figure.

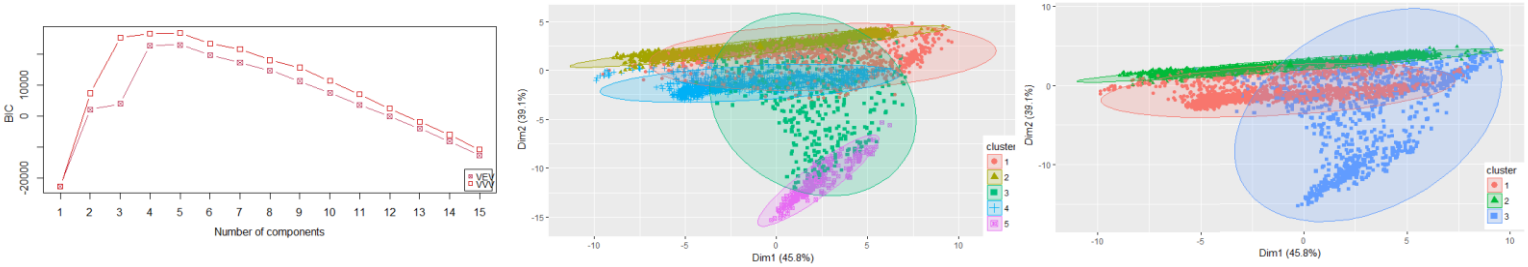


Figure 8: BIC plot and clusters (Landsat)

Dataset 2 (Balance Scale):

Figure 9 shows a plot of BIC values w.r.t number of components for two different models. First is equal volume ellipsoidal model (VEV) and other is equal volume ellipsoidal model(EII). Figure 9 shows analysis of both highest and lowest values of k (k=3, k=6), Why BIC value? Good clustering accuracy in high dimensions might lead BIC values to go lower but it does not mean aligned result with actual classes.

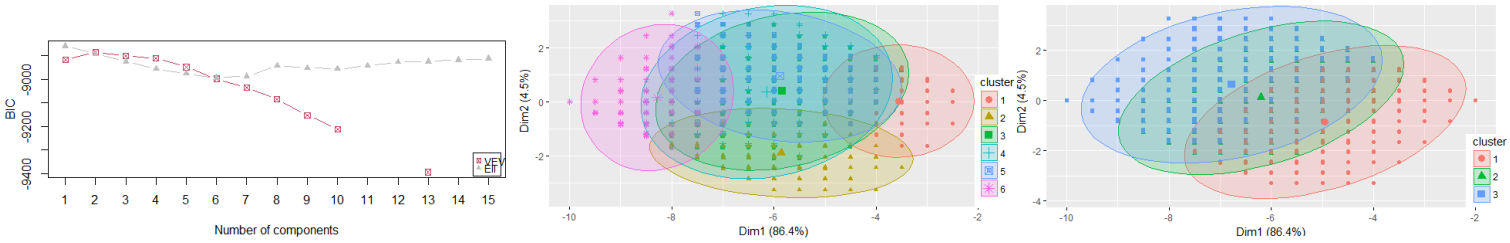


Figure 9: BIC plot and clusters (Balance Scale)

Cluster analysis:

For Dataset 1, Figure 9 shows cluster distribution for k=5 and k=3. The significant observation is, clusters are horizontals and totally aligned to points distribution (Exactly opposite to K-means). This enables EM higher overall accuracy over K-means i.e. 62.49%. Also, it accurately identifies clustering for class 1,2,4,5,7 for Landsat dataset leaving class 3 with accuracy 28.82%. For Dataset 2, the results for EM is better in comparison of K-Means algorithm but again it has a low overall accuracy of 26.03%. For most labels from "R" and "L", EM able to classify correctly. However, for label "B" algorithm only able to classify 18.75% correctly, that leads to lower overall accuracy for Balanced Scale Dataset.

2.3. Comparison and Performance Analysis

For Landsat dataset, EM performs better than K-Means. Time taken by EM is more but in terms of accurate identification of clustering, EM is the best algorithm. Why? performance of EM on Landsat is dependent on nature of distribution of dataset over classes. For balanced scale dataset, both algorithms performed poorly to classify class "B" and leads to a low overall accuracy of clustering. Time taken by EM is comparatively higher than K-Means, this is because of iteratively calculating log-likelihood for current parameters and maximization of that.

3. Dimensionality Reduction Algorithms:

The main idea behind Dimensionality Reduction is to pre-process data and identify/transform important features from a large set of features. Feature selection selects important features from a given set of features. Feature Extraction transform features to lower dimensions.

3.1. PCA (Principal Component Analysis)

PCA achieves dimension reduction by creating new, artificial variables called principal components. Each principal component is a linear combination of the observed variables. It is a well-established technique for reducing the dimensionality of data while keeping as much variation as possible (without minimum information loss). I have used a library that calculates singular value decomposition of the data matrix and return list of principal components. The measurement of the amount of data retained by PCA is calculated by variance and eigen value.

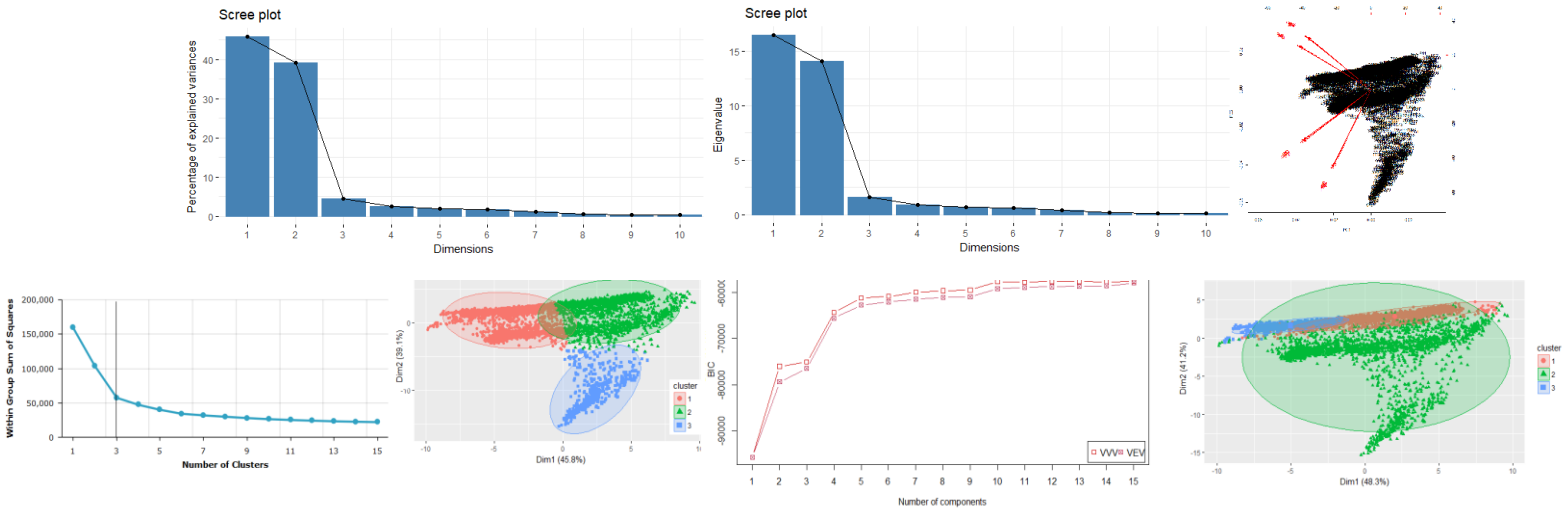


Figure 10: Scree Plot, K-Means and EM Clusters (landsat)

Dataset 1 (Landsat):

Figure 10 shows when we apply clustering algorithms on PCA. It is clear from the first plot that, we can obtain about 80% of variance by retaining first 2 PCs. One of the important observation for cluster distribution that for K-means value of k and cluster separation is almost same but for EM it is changed from the original cluster. Why? It is possible that distribution of first 2 PCs contribute more in K-means and comparatively less in EM. There is no significant increase in overall accuracy after application of PCA. I think it is because of 3x3 pixel distribution of spectral value in Landsat data. In conclusion, data re-construction does not give much advantage in Landsat dataset in terms of accuracy.

Dataset 2 (Balance Scale):

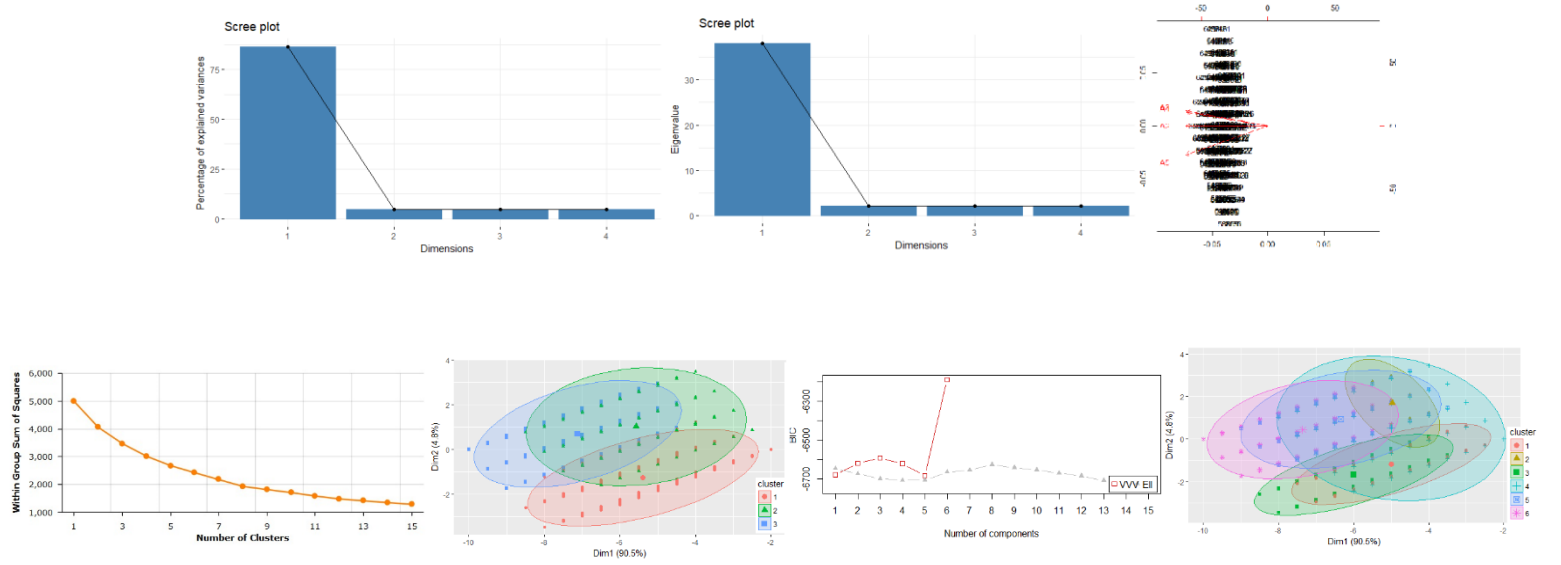


Figure 11: Scree Plot, K-Means and EM Clusters (Balance Scale)

Figure 11 shows when we apply clustering algorithms on PCA for Balance scale dataset. It is clear from the first plot that, we can obtain about 85% of variance by retaining the first PC. This indicates data distribution is highly dependent on less number of features than actual. The graph shows Scree plot for Eigen value and percentage of explained variances. Clusters observed very sparse and overlapping (more dimensions) from original ones. Why? Mostly because of balance scale classes correlated to features A4 and A5. PCA tried to transform features into another space and ends up with slightly lower overall accuracy for both K-means (20.79%) and EM (25.87%).

3.2. ICA (Independent Component Analysis)

The method originated from signal processing research, where signal sources are mixed to a new set of signals. It separates a multivariate signal into additive subcomponents by assuming that signals are non-Gaussian and independent of each other. The used library is "fastICA" attempts to 'un-mix' the data by estimating an un-mixing matrix.

Dataset 1 (Landsat):

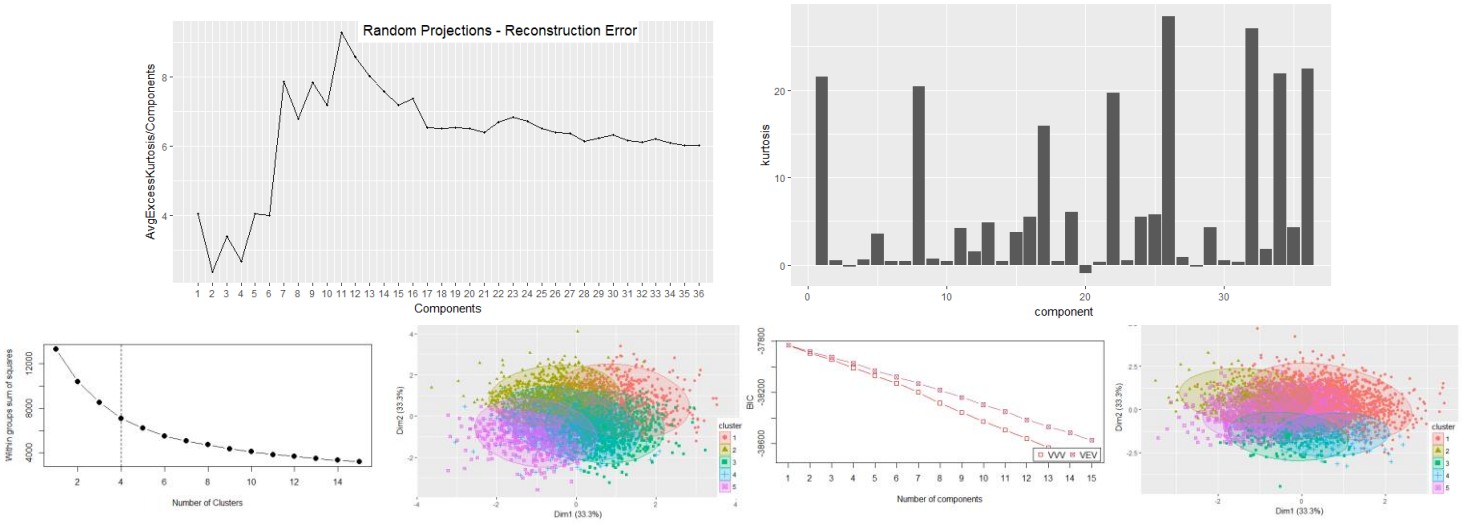


Figure 12: Reconstruction Error plot, kurtosis plot, K-Means, EM Clusters *(Landsat)*

Figure 12 shows average kurtosis over a 20 runs w.r.t to the components for Landsat dataset. The maximum value for average kurtosis is 9 as in plot. This is used to extract parameters as shown in the subsequent figure. Components 26, 32 and 34 has higher kurtosis values than other components. These parameters passed to K-Means and EM with k=5. Result cluster is observed with no clear separation between clusters opposite from original cluster. Why? this might be because of neighborhood pixels would have independent spectral pixel values than other pixels and can be seen in lower dimensions. K-Means and EM had performed well after performing ICA and overall accuracy increased for both K-Means (63.04%) and EM (65.49%).

Dataset 2 (Balance Scale):

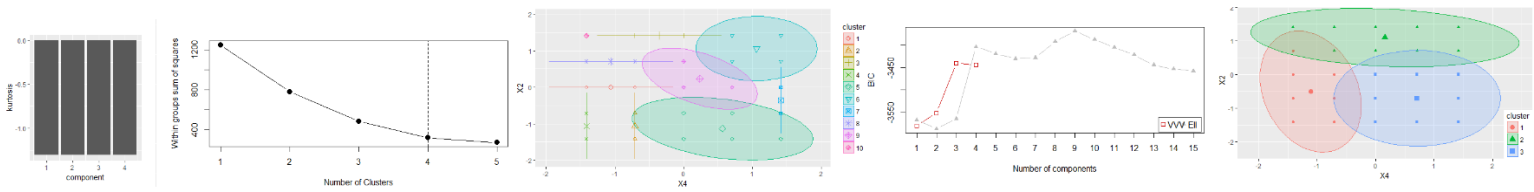


Figure 13: kurtosis plot, K-Means, EM Clusters *(Balance Scale)*

Figure 13 shows average kurtosis over a 20 runs w.r.t to the components for Balance scale dataset. I came here with the significantly ominous observation that kurtosis of all components is same and it is 0. This is only possible when the distribution is Normal distribution or Mesokurtic distribution. These parameters passed to K-Means and EM with k=10 and k=3. It has given poor result and relatively different clusters than original clusters. It seems that there is the only random generation of cluster and assignment of points. It is clear that ICA does not generate meaningful results and also poorly performed in terms of overall accuracy for this dataset.

3.3. RP (Random Projection)

The Random Projection is dimensionality reduction method suitable for the distance-based method (Points that lies in Euclidian Space). The dimensions and distribution of random projections matrices are controlled so as to preserve the pairwise distances between any two samples of the dataset.

Dataset 1 (Landsat):

Figure 14 shows reconstruction error found by adding one component over time for multiple runs of Random Projection on Landsat Dataset. K-means optimal value is 4. However, to compare this clustering with optimal clustering, it is taken as 5 for both EM and K-Means. K-means cluster is observed with change in skew but with same horizontal separation. Also, EM cluster is skewed from original one but if we observe overall accuracy, it is less than optimal distribution (61.75%).

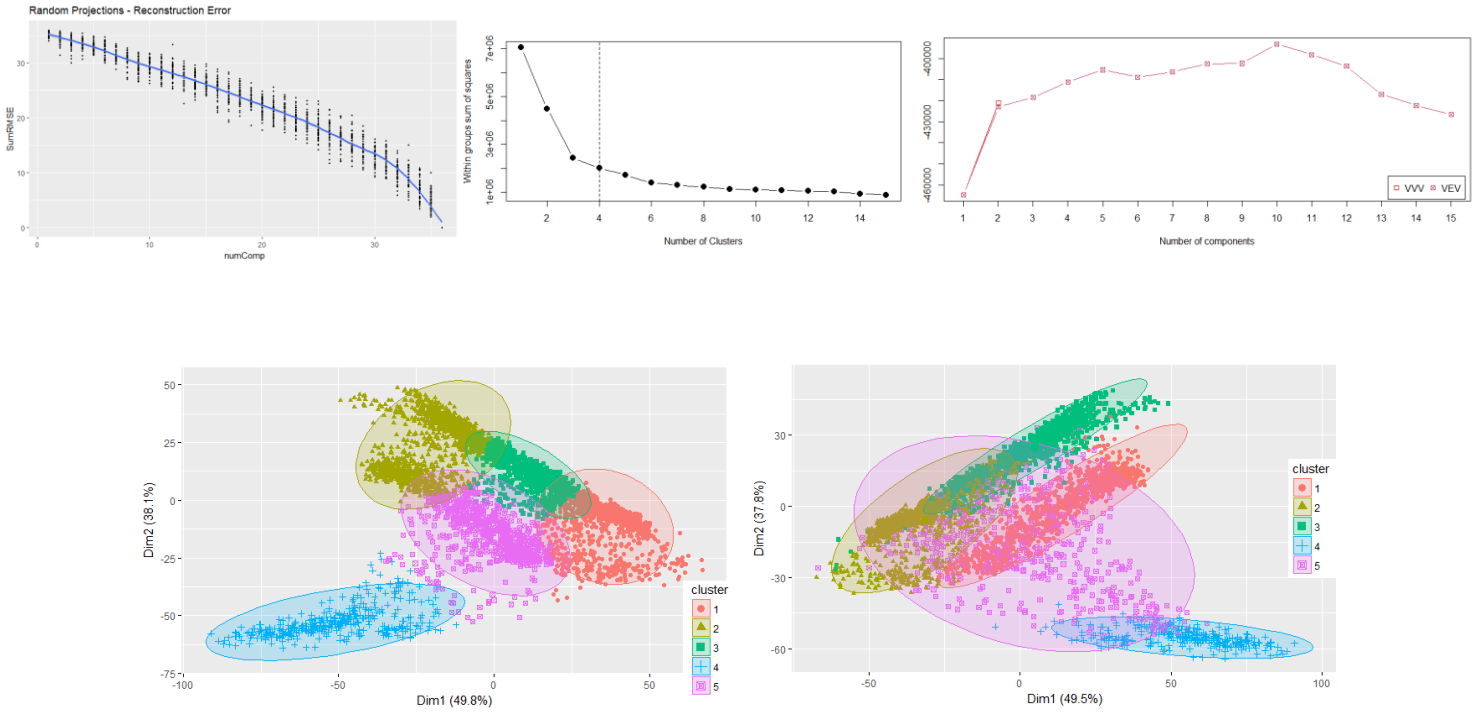


Figure 14: K-Means, EM Clusters over runs of RP *(Landsat)*

Dataset 2 (Balance Scale):

Figure 15 shows reconstruction error found by adding one component over time for multiple runs of Random Projection on Balance Scale dataset. For K-Means optimal k=4 but clusters are constructed across data and not beneficial to increase overall accuracy. For EM optimal value calculated from EEI values and clusters are as shown. One of the disadvantages of RD in balance Scale dataset is, running it multiple times and averaging the value will give lower accuracy. Data points in clusters are very near and can be in multidimensional space. EM performs better than K-Means in this case, but it is not better than above algorithms.

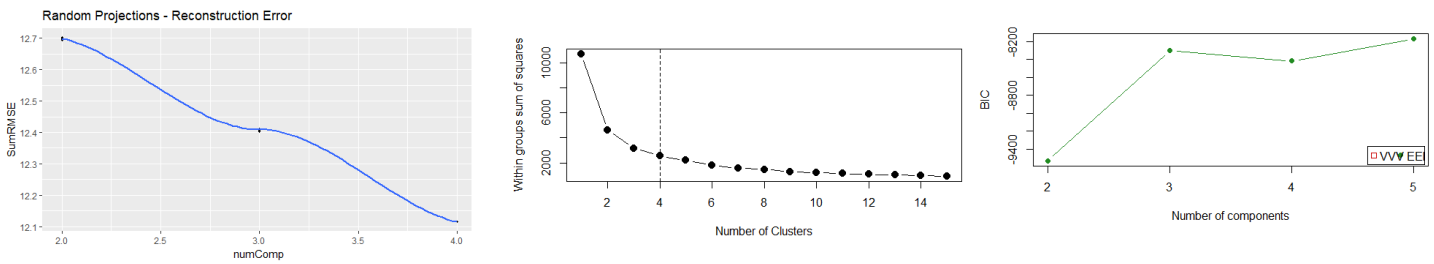


Figure 15: K-Means, EM over runs of RP *(Balance Scale)*

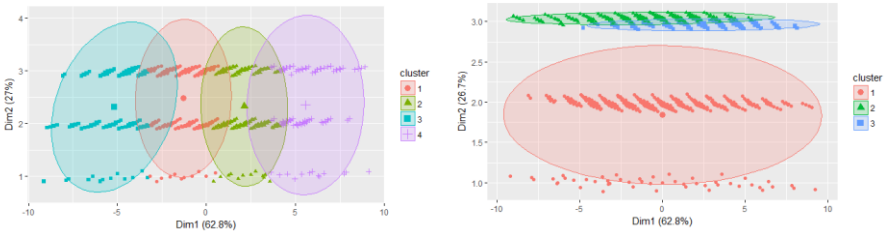


Figure 16: K-Means, EM Clusters *(Balance Scale)*

3.4. IG (Information Gain)

The information gain of an attribute indicates the significance of that feature with respect to the classification. It depends on information value before and after knowing the attribute value. (less entropy, high IG).

Dataset 1 (Landsat):

Figure 17 shows information Gain for all features. From the plot, we can conclude that, features: T, R, Q, U, P has high information-gain, and these passed to the clustering algorithm. K-Means optimal value k=4. Clusters and data points distribution look similar to previous one in two dimensions. However, results are changed, and K-Means poorly performed for identification of class 5(28.65%) and class 7(16.12%). For EM, from VVV model as shown in the figure, the optimal value I observed for both k=3 and k=4 values. For comparison of EM to K-Means k=4 clusters are shown. EM gets near similar overall accuracy (64.91%) to ICA.

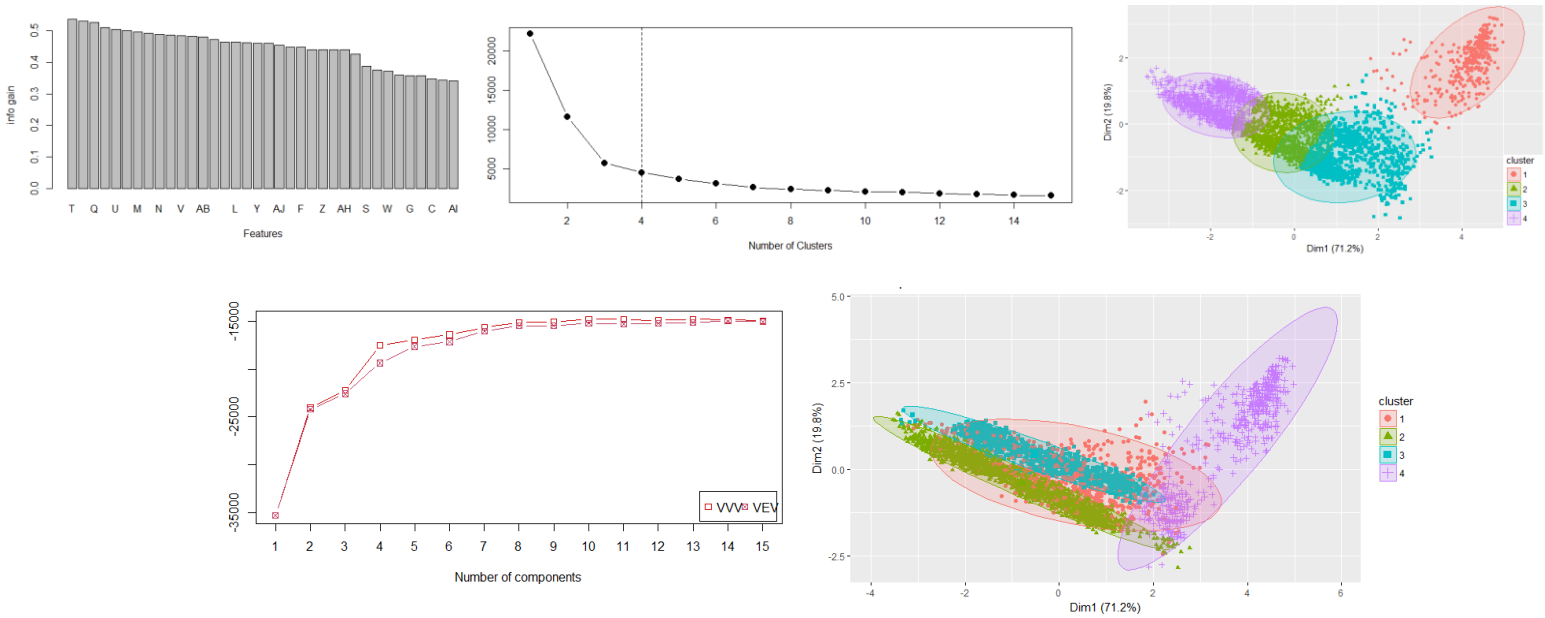


Figure 17: IG Plot, K-Means, EM Clusters *(Landsat)*

Dataset 2 (Balance Scale):

Figure 18 shows information Gain for all features for Balance scale. One important observation that IG is opposite from data distribution graph in the previous figure. A2 features have more information

gain. Executing K-Means for optimal k=4 and EM for k=3 fives similar clusters as EM of the dataset. Why? while observing correlation plot of balance scale it is clear that A2 is negatively correlated with classification. Using A2 as information gain attribute will result in reverse assigned clusters. This led to the overall accuracy of 24.44% for the dataset.

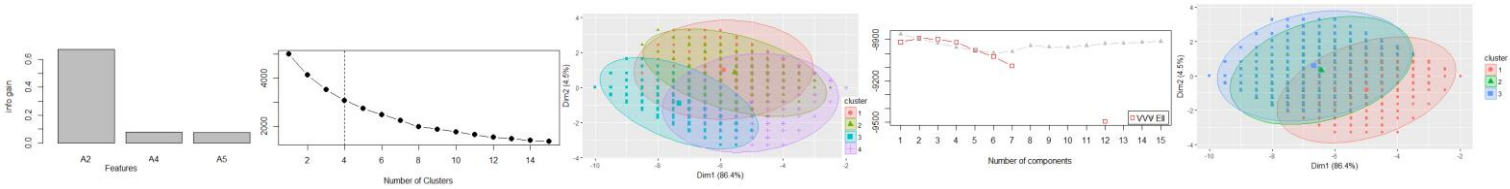


Figure 18: IG Plot, K-Means, EM Clusters *(balance Scale)*

4. Neural Network Analysis on Landsat:

I have selected Landsat dataset for observing the effect of Neural Network (Multilayer Perceptron) after dimensionality reduction and clustering on the dataset.

Figure 19 shows training and testing errors with No clustering algorithm applied, with K-Means and with EM on PCA, ICA, RP, IG. RP gives so far best accuracy result (89.1%) that is closer to a simple Neural net (90.15%). Overfitting is observed in PCA where it fails to give accurate results on the test dataset. Also, one key observation for IG is high variance giving the poor performance of training and test datasets.

	No Cluster		K-Means **k=3**		EM **k=3**	
	Train %	Test %	Train %	Test %	Train %	Test %
PCA	83.29	70.05	81.59	71.9	80.06	68.89
ICA	80.27	83.95	80.04	84.3	81.17	83.8
RP	87.52	89.1	89.36	86.77	79.54	76.89
IG	62.89	55.79	61.56	47.72	63.5	60.85

Figure 19: Comparison Table (Landsat)

One significant Observation for execution Time is, the time taken to train model is very less than Simple Neural net without dimensionality reduction algorithm. If we compare the maximum time taken by any combination, dimensionality reduction algorithms (58.92s) are best than without it (309.54s).

The key learning from above experiments proves that Dimensionality reduction and clustering algorithm decreasing the processing time of algorithms with lowest information loss and without losing accuracy. For Landsat, accuracy not improved from simple Neural Network experiment, but it provides the higher perspective of dimensionality reduction with a huge number of attributes with high accuracy.