# Un-supervised Learning, Dimensionality reduction and their influence on Neural Network

Umesh Jadhav
*OMSCS*
*Georgia Institute of Technology*
Georgia, USA
ujadhav6@gatech.edu

***Abstract*—Unsupervised learning is a machine learning technique used when target variables are not available, unlike supervised learning. It works by uncovering hidden patterns in the data and grouping similar data points into clusters. The goal is to minimize the distance within each cluster and maximize the distance between clusters. Dimensionality reduction techniques transforms the higher-dimensional data to lower dimensional space without loosing the information during the conversion. In this study, K-Means, and Expectation Maximization algorithms are implemented over Company Bankruptcy dataset and Diabetes Classification dataset. Further, Principal Component Analysis(PCA), Independent Component Analysis(ICA), and Random Projection(RP) are used to transform the aforementioned datasets into lower dimensions. Then, the two clustering algorithms are re-implemented over the lower dimensional spaces generated by the three dimensionality reduction algorithms to analyze their effect. Finally, two Neural networks are developed by re-instating the target variables where the first model leverages the projected data from dimensionality reduction algorithms for predicting the class label whereas, the second model uses the the intra-cluster distances and density as training set. The performance of these models is then compared with the previous study[1] for type 1 and type 2 errors. It is found that the K-Means performed well on continuous feature space while EM favors categorical feature space. Similarly, PCA due to its linear projection outperforms ICA and RP for continuous variables whereas, the same cannot be considered optimal when implemented over discrete variables. Also, neural network with ICA showcase better metrics than PCA and RP whereas, neural network with EM cluster density shows greater type 1 and type 2 error.***

## I. Introduction

Un-supervised Learning is a sub-domain of Machine Learning where the models have no prior knowledge of target variables. This technique groups the dataset represented by data points in form of clusters where, the data points that show identical patterns belong to the same cluster. The main idea here is to make the clusters as compact as possible by reducing the intra-cluster distance between the data points and the cluster center. On the other hand, to ensure consistency, the distance between various clusters in the space needs to be maximum known as inter-cluster distance. In this study, K-Means clustering and Expectation Maximization algorithms are implemented over Company Bankruptcy dataset and Diabetes classification dataset. To choose optimal number of clusters, Sum of Squared Distance(SSD) is leveraged in case of K-Means clustering whereas, for Expectation Maximization, Bayesian Information Criterion(BIC) is main metrics used for cluster evaluation.

Dimensionality Reduction techniques transform the higher dimensional dataset into lower dimensional dataset by projecting it onto a space where a higher variance is anticipated. The main idea here is that higher variance captures more data points resulting in compressing greater information. As a result, in this study, three dimensionality reduction algorithms that are Principal Component Analysis(PCA), Independent Component Analysis(ICA) and Randomized Projection(RP) are implemented over the two datasets to study their effect. For performance evaluation, Scree plots, variance plots for PCA, Kurtosis plot for ICA, and Reconstruction error plot for RP are used. Further, the effect of dimensionality reduction techniques over clustering algorithms is analyzed and compared with the results of clustering for the original datasets.

Finally, two Neural network models are developed based upon the feature space generated by the clustering algorithms and dimensionality reduction algorithms by completely replacing the original datasets' feature space. However, the target variable is not replaced, making the final step as the classification problem. Here, accuracy and type 1 and type 2 errors are compared with the previous study[2].

## II. Datasets

Company Bankruptcy dataset[3] has 73 continuous independent features and 6300 observations with binary label target variable. Here, the large feature space makes it favorable for PCA and RP since ICA would suffer for determination of independent features. However, the threshold choice of RP would introduce noise into the feature space since the algorithm converges to lowest reconstruction error for the same number of features. Whereas, EM would generate lower clusters as that of the K-Means technique as the data points are overlapped making it difficult for K-Means to cluster them separately. Similarly, the inherent feature of K-Means of bias towards Spherical clusters would still suggest comparatively greater clusters than EM for dimension reduction due to the continuous feature space.

As for the neural network, for dimension reduction algorithms, neural network with PCA would converge quickly as that of the other two algorithms with better accuracy and type 1 and type 2 errors since ICA would generate greater

dimensions than PCA and RP. Additionally, neural network with EM would provide better accuracy than with K-Means due to the lower number of cluster space and since the density suggest the likelihood of data points belonging to the cluster, it is easier for Neural network to identify the target label based upon it.

Whereas, Diabetes Classification dataset[4] contains 13 independent features with mixture of discrete and continuous nature with a binary target variable and 1,00,000 records. Here, the linearly separable data point make it favorable for both, K-Means and EM. This is due to the fact that K-Means tends to create clusters with clear division of the data points, creating higher inter-cluster distance. On the other hand, although EM provides the likelihood of the data points belonging to the clusters, the speration between them will follow a similar pattern as that of K-Means.

However, since the original feature space is tremendously small as compared to that of the number of samples, RP would suffer and suggest higher number of resultant dimensions as that of the original feature space since RP follows Johnson Lindenstrauss lemma. Whereas, since the features of this dataset follow discrete categorical and numerical values, PCA would extract the information in minimum number of components. Similarly, ICA will generate number of components closer to that of PCA sine the feature space is low and majority of data points are categorical.

Finally, the implementation of Clustering algorithms over the dimension reduced features could match the ground truth label since the original low dimensional space would further be reduced to lower dimensions. As a result, K-Means and EM, both are expected to follow similar metrics in case of PCA and ICA. However, in this case, both the clustering algorithms would suffer when working upon the resultant feature space of Sparse RP.

## III. Methodology

The entire experimentation is done using Python programming language with scikit learn as the core framework for implementing all the algorithms. For visualization purposes, matplotlib library is leveraged with an additional usage of pandas and numpy. The experimentation methodology for both the datasets follow similar structure and is discussed sequentially. For clustering as well as for dimension reduction algorithms, single random seed of 29 is defined.

### A. Clustering

1) K-Means - First, the optimal setting is chosen based upon comparison between hyper-parameters of $KMeans()$ from scikit learn such as $algorithm$, $init$, and $n\_init$ for cluster sizes between 2 and 20 for both the datasets. Further, the optimal number of clusters is estimated using elbow plot over cluster Inertia(Sum of Squared distance).
2) Expectation Maximization(EM) - Here, $GaussianMixture()$ from scikit learn module is used to implement the algorithm by first comparing the hyper-parameters performance for optimal clusters such as $algorithm$, $covariance\_type$, and $init\_params$. The algorithm for both the datasets is tested for clusters between 2 and 20 with 1000 iterations per component with analysis of BIC score.

### B. Dimensionality Reduction

1) Principal Component Analysis - PCA is implemented using $PCA()$ from scikit learn module, tested for components between 2 and 15 for Company Bankruptcy dataset and 2 to 10 for diabetes classification dataset. Additionally, various values of $svd\_solver$ with $power\_iteration\_normalize = "auto"$ are compared for defining the optimal settings. For performance evaluation, scree plot, and cumulative variance plot are leveraged.
2) Independent Component Analysis - ICA is implemented using $FastICA()$ from scikit-learn module and analyzed for components between 2 and 30 for Company Bankruptcy dataset, and 2 to 16 for the second. Here, $algorithm$, $fun$, and $whiten\_solver$ are compared and evaluated using average kurtosis plot.
3) Randomized Projection - Here, $GaussianRandomProjection()$ and $SparseRandomProjection()$ from scikit-learn are compared with $eps$ values between 0.1 and 0.9 for both the datasets for components between 1 and the number of features of the dataset. The performance is evaluated using Reconstruction error plot of reconstructing original feature space from the reduced dataset.

### C. Clustering over Dimensionality Reduction

1) K-Means with PCA, ICA and RP - For PCA, 6 components and 3 components with auto SVD solver for Company Bankruptcy dataset and Diabetes Classification dataset is defined here respectively based upon the output from previous section.
   For ICA, 10 components with "svd" svd solver and "logcosh" function is preferred for Diabetes Classification dataset whereas, 14 components with "eigh" svd solver and "cube" function is chosen for Company Bankruptcy dataset.
   For RP, in case of Company Bankruptcy dataset, Gaussian model with 38 components and 0.9 eps is defined whereas, Gaussian model with 8 components and 0.9 eps is chosen for Diabetes classification dataset. These hyper-parameter settings are obtained by considering 50% threshold of reconstruction error.
   $KMeans()$ is evaluated for clusters between 2 and 20 over the reduced dataset generated by the three dimensionality reduction algorithms. Finally, The results are then compared with that of the results from Clustering sub-section.
2) Expectation Maximization with PCA, ICA and RP - Similarly, on the other hand, the hyper-parameter setting for PCA, ICA and the RP model as seen for KMeans are evaluated for clusters between 2 and 20, generated by $GaussianMixture()$.
   These results are then compared with the results of EM from Clustering sub-section.

### D. Neural Network with Dimensionality Reduction

1) Neural Network with PCA - Based upon the best set of parameters, for Company Bankruptcy dataset, $PCA()$ with 6 components and "auto" SVD solver is defined. As for the neural network, two layered architecture is followed where the first layer defined by RELU activation function with 48 neurons. The second layer performing as the output layer is defined by Sigmoid activation function with a single neuron for binary classification. Finally, $binary\_crossentropy$ is defined as the loss function for the model.
2) Neural Network with ICA - A similar neural network structure is followed here as that of PCA. Here, $FastICA()$ with 14 components, "eigh" as $svd\_solver$ and $algorithm = "parallel"$ is defined as from the kurtosis plot.
3) Neural Network with RP - For RP, Gaussian Random Projection is preferred with 35 components and eps of 0.9 for Company Bankruptcy dataset with a similar neural network structure as that of PCA and ICA.

### E. Neural Network with Clustering

1) Neural Network with K-Means - A two layered neural network as defined in previous sub-section is implemented here. Since this is a classification task, the output layer has single neuron and the input space is comprised of intra-cluster distance from centroid of the original dataset however, the original target variable is the prediction label for our model. For Company Bankruptcy dataset, the first layer is built upon 48 neurons and 7 clusters.
2) Neural Network with Expectation Maximization - A similar neural network structure is followed here as that of K-Means. However, the input space here is the density of original dataset with respect to the clusters, obtained by $predict\_proba()$ of $GaussianRandomProjection()$ with 9 components for Company Bankruptcy dataset. The original target variable is reinstated here for the classification task in both the datasets.

## IV. Results

Please note here that, all the time related metrics may vary depending upon the system configurations.

### A. Clustering

#### Company Bankruptcy Dataset

1) K-Means
First from figure 7, elbow plot of K-means clustering with elkan algorithm and random initialized centroids for Sum of Squared distance defined by inertia, showcase an elbow at cluster 7 for normalized SSD score of approximately 0.28.
2) Expectation Maximization(EM)
For EM algorithm in figure 8, covariance type as "full" and "random" initialized centroids agree upon the number of clusters of 8 for lowest BIC score with a value of 6550000 and 6750000 respectively.

#### Diabetes Classification Dataset

1) K-Means
Secondly, for Diabetes Classification dataset, K-means clustering with "elkan" algorithm, with "random" init parameters agree upon 4 clusters through elbow point of normalized sum of squared distance score of 0.4 from figure 9.
2) Expectation Maximization(EM)
In case of EM, elbow point is evident at cluster 4 with 0.6 BIC score for "spherical" algorithm with "kmeans" init parameters as seen in figure 10. This metric and results are generated for both, EM trained with target label as well as EM trained without target label.

### B. Dimension Reduction

#### Company Bankruptcy Dataset

1) PCA
From figure 1, with a cumulative variance threshold of 90%, the optimal number of components is estimated to be 6. The same is asserted by individual variance value and variance ratio.

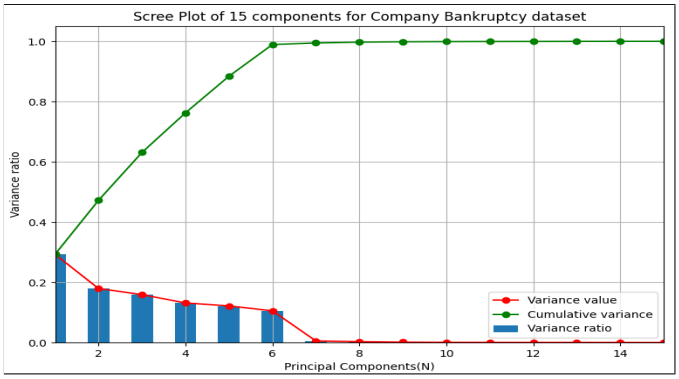


Fig. 1: PCA Scree plot for Company Bankruptcy dataset.

2) ICA
On the other hand, the average absolute kurtosis value per component is illustrated in figure 2. Since the component with highest average kurtosis value is considered to be the optimal, here the component is evident to be 14 where the average absolute kurtosis value is seen to be 40. Additionally, these results are obtained by defining "eigh" as whiten solver with cubical function by comparing with other values for the hyper-parameters.
3) RP
To determine the optimal components for Sparse RP and Gaussian RP are first compared for various values of eps. However, as evident from figure 3, all the eps values follows an overlapping trajectory, declining to the lowest reconstruction error of 0 at 73 components. As a result, Gaussian RP is chosen for further experimentation with 35 features for a thresold of 50% reconstruction error. However, the dimensions generated have minor correlation between them, indicating the algorithm's poor performance.

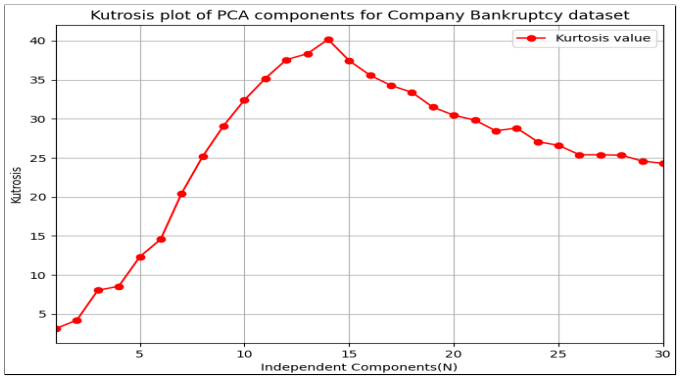


Fig. 2: ICA Kurtosis plot for Company Bankruptcy dataset.

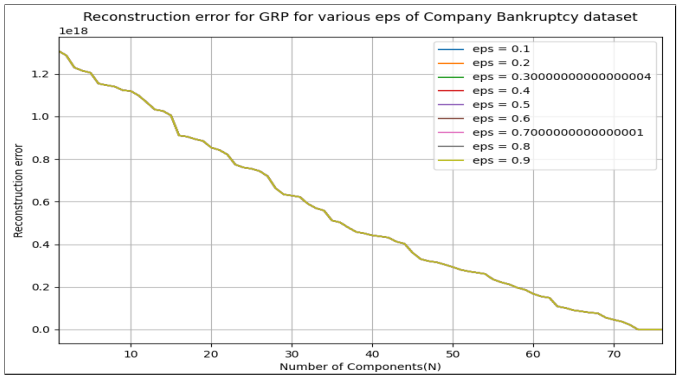


Fig. 3: Gaussian RP Reconstruction Error for Company Bankruptcy dataset.

### Diabetes Classification dataset

1) PCA

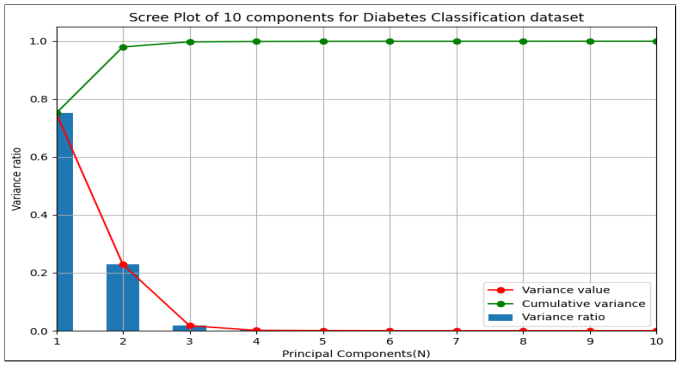


Fig. 4: PCA Scree plot for Diabetes Classification dataset.

As evident from scree plot of figure 4, the optimal components are 3 with cumulative variance threshold value of 90%, which is adduced by the variance ratio. This model is chosen after its comparison with "arpack" SVD solver which showcased similar results.

2) ICA

Here as seen in figure 5, the highest average kurtosis value of 150 is achieved for 10 components for model with "logcosh" function and "eigh" whiten solver. Other models with function "exp" and "cube" and "svd" whiten solver displayed high kurtosis at 10 components however their absolute average value did not match this model.

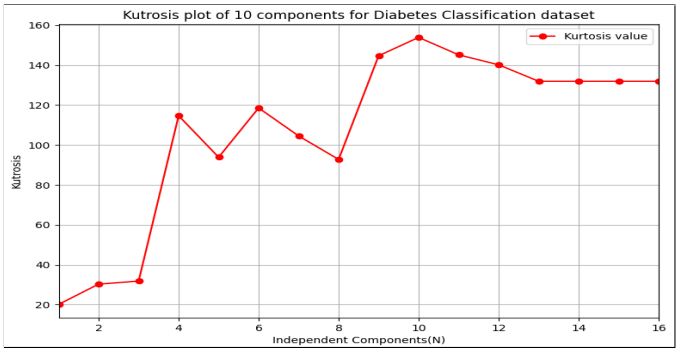


Fig. 5: ICA Kurtosis plot for Diabetes Classification dataset.

3) Randomized Projection

In case of RP, Gaussian RP as well as Sparse RP illustrates similar trend for all eps values as seen in figure 6. The reconstruction error attains the value of 0 at 13 components hence, for further tasks, Sparse RP with 0.9 eps and 8 features is chosen.

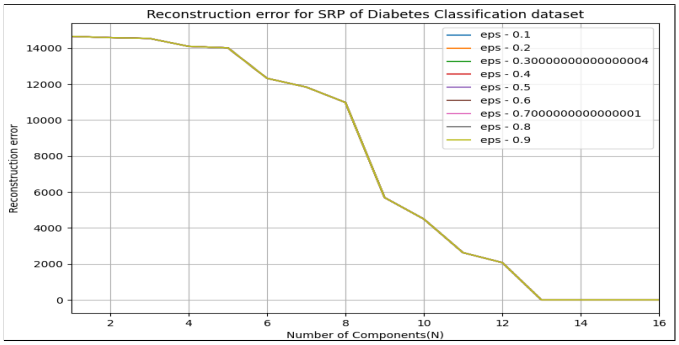


Fig. 6: Sparse RP Reconstruction Error for Diabetes Classification dataset.

## C. Clustering with Dimension Reduction

### Company Bankruptcy Dataset

1) K-Means with PCA, ICA and RP

Figure 7 illustrates elbow plot of K-Means for PCA, ICA and RP reduced dataset along with the Original dataset's feature space. First from PCA elbow plot, an elbow point is seen at cluster 7 with a normalized SSD score of 0.29. ICA follows the trend however, the elbow point is evident for 6 clusters with the SSD score of 0.3. Whereas, Gaussian RP appears to be struggling with uncovering the hidden pattern since it indicates an elbow point at 9 clusters with normalized SSD score of 0.3. Finally, K-Means on the original feature space follows a tight relationship with the trend line of PCA.

2) Expectation Maximization with PCA, ICA and RP

Figure 8 illustrates BIC score of EM for PCA, ICA and RP reduced dataset compared with the original dataset's

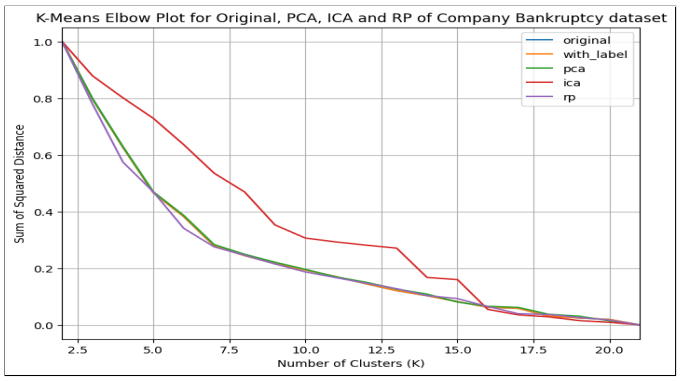


Fig. 7: K-Means Elbow plot of PCA, ICA and RP for Company Bankruptcy dataset.

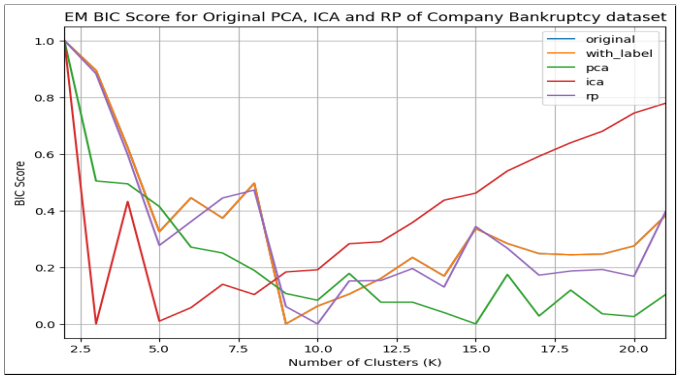


Fig. 8: EM BIC score with PCA,ICA and RP for Company Bankruptcy dataset.

feature space. First for EM with PCA reduced features, lowest normalized BIC score of 0.0 is evident at 15 clusters although the line is filled with various spikes and troughs throughout. Secondly, ICA indicate similar BIC score of 0.0 for 2 clusters specifically for 3 and 5 followed by a significant linear increase in the metric. Since ICA illustrate number of clusters close to that of the actual ground truth, 3 clusters are considered here. Finally, Gaussian RP showcase similar number of clusters as that of PCA, indicating 10 as the ideal number. However, the trajectory before and after the lowest BIC score is different as that of PCA.

### Diabetes Classification Dataset

1) K-Means with PCA, ICA and RP
Figure 9 illustrates elbow plot of K-Means for PCA, ICA and RP reduced dataset along with the Original dataset's feature space. Here, all the dimension reduction algorithms appears to agree upon the optimal number of clusters as 6 since the elbow point of the normalized sum of squared distance is visible at cluster 6 though the original dataset, PCA and RP appears to converge at 0 SSD score quicker as that of ICA. Specifically, K-Means over PCA reduced dataset illustrates a sudden change in the momentum of the trend with the normalized SSD score of 0.2. However, Sparse RP showcase a relaxed decrease in the metric

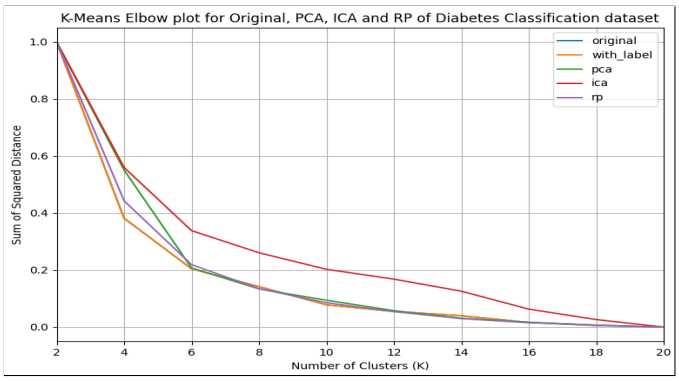


Fig. 9: K-Means Elbow plot with PCA, ICA and RP for Diabetes Classification dataset.

bending at two points though similar SSD score followed by it. Although facing a delay in convergence, ICA also follows the same trend proportional to RP with a tilt in its graph evident at two points however, in this case the normalized SSD appears to be 0.35

2) Expectation Maximization with PCA, ICA and RP
Figure 10 illustrates BIC score of EM for PCA, ICA and RP reduced dataset along with the Original dataset's feature space. Here, instead of choosing lowest BIC score, the existence of elbow point is considered. First, EM over PCA reduced dataset indicate 6 as the optimal clusters with normalized BIC score of 0.4. On the other hand, in case of ICA, the elbow point is evident at cluster 8 indicating as the optimal clusters with the normalized BIC score of 0.35. Whereas, Sparse RP indicate the optimal clusters as 4 as evident from the elbow point for normalized SSD score of 0.62. When compared with the BIC score of EM with original dataset, Sparse RP matches the BIC score of it for optimal clusters as 4.

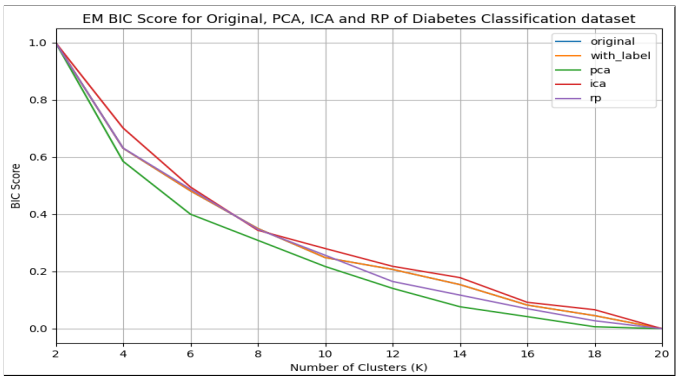


Fig. 10: EM BIC score with PCA, ICA and RP for Diabetes Classification dataset.

### *D. Neural network with Dimension Reduction for Company Bankruptcy dataset*

1) Learning Curve

TABLE I: Type I and Type II score of Neural network with Dimension Reduction algorithms.

| Error | PCA | ICA | RP |
|---|---|---|---|
| **Type I** | 49 | 55 | 53 |
| **Type II** | 69 | 3 | 30 |

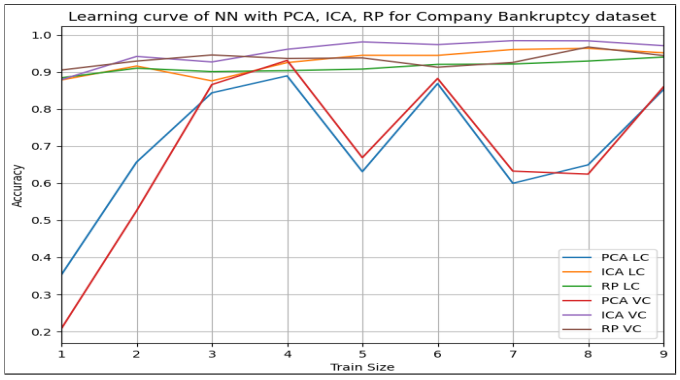


Fig. 11: Learning Curve of NN with PCA, ICA and RP for Company Bankruptcy dataset.

Figure 11 illustrates learning curve for neural network for PCA, ICA and RP reduced dataset with original target variable. First for PCA, the learning curve displays a linear growth in accuracy upto 40% of the training data. A dip in the accuracy can be seen at training size 50% before rising at the 60% however, another decline of 60% can be seen at training size of 70%, rising gradually thereafter to terminate at 87% accuracy.

In case of ICA, the learning curve in figure 11 crossed 98% accuracy for train size 70% and 80%. Both the train and validation sets follows similar trajectory indicating good learning rate.

On the other hand, for RP, a similar pattern as that of PCA is visible from figure 11 for Learning curve. The model reaches 90% accuracy at the 40% training size. However, none of the further incremental sizes breaks the record though both the trends follow close trajectory with respect to each other indicating a good learning rate.

2) Validation Curve

As for the validation curve in figure 12 for PCA, a linear rise in accuracy for first 20 iterations can be seen before reaching 90% accuracy for the next 20 iterations. Though the momentum subsides for the further iterations, the gradual increase in the accuracy reaches at the peak of 93% for the last iteration.

The validation curve of ICA from figure 12 indicates better convergence of train and validation sets, rising quickly to 95% at $10^{th}$ iteration. Further iterations for both the sets maintains constant accuracy of 97%.

Whereas, the validation curve from right sub-figure of figure 24 begins with 65% accuracy while crossing the mark of 95% quickly with a couple of spikes evident at $20^{th}$ and $40^{th}$ iteration. The final accuracy of the model remains at 96%.

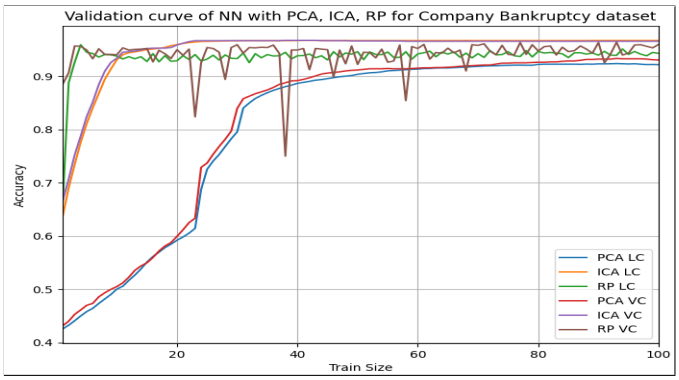


Fig. 12: Validation Curve of NN with PCA, ICA and RP for Company Bankruptcy dataset.

Finally, for table 1, the type 1 and type 2 errors show a significant difference with that from study[2]. Specifically, ICA and RP reduced dataset outperforms the baseline neural network . Although PCA showcase lower type 1 error of 49, it loses for the type 2 error with the baseline model since the former has 69 samples while the latter has 37.

### *E. Neural Network with Clustering for Company Bankruptcy dataset*

TABLE II: Type I and Type II score of Neural network with Clustering algorithms.

| Error | K-Means | Expectation Maximization |
|---|---|---|
| **Type I** | 55 | 55 |
| **Type II** | 71 | 0 |

1) Neural Network with K-Means

From the learning curve from figure 13, a pair of troughs and plateau is seen where train set and validation set, both following the similar trajectory. The trend begins with 85% accuracy for the 10% training size while decreasing up to 60% during 30% size. However, the accuracy reaches 93% and 96% for train and validation set respectively for 60% size before dipping quickly at 80% train size.

In case of validation curve from figure 14, the model rises rampantly from 10% to 80% accuracy before 20 iterations. Although the strength of growth lowers, the model maintains an accuracy of 97% after 60 iterations.

Also, neural network with K-Means indicate type 1 error of 55, outperforming the baseline model[2] by large margin however, trade-off for type 2 error is evident by two times.

2) Neural Network with Exp. Max.

The learning curve as seen in figure 13 starts with an excellent learning rate of 95% and 97% for training and validation sets respectively. The trend maintains the momentum until 30% of train size after which a declining accuracy is evident in both the sets. However, the trend recovers from the fall quickly following 60% train size,

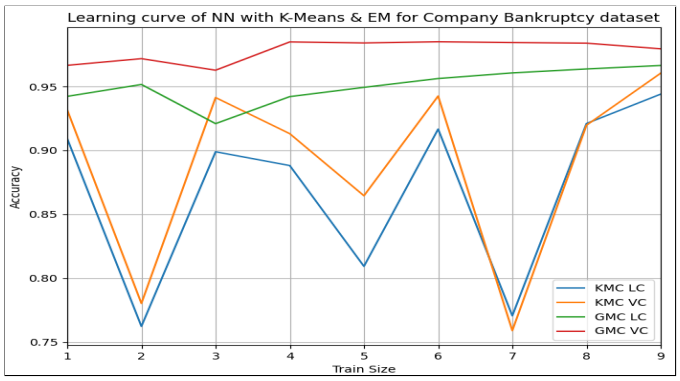


Fig. 13: Learning Curve of NN with K-Means and EM for Company Bankruptcy dataset.

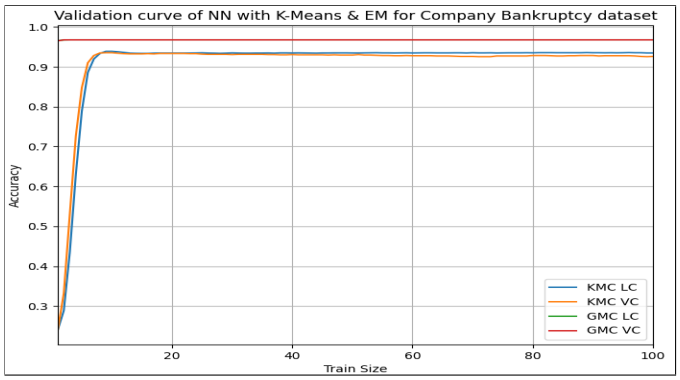


Fig. 14: Validation Curve of NN with K-Means and EM for Company Bankruptcy dataset.

reaching a maximum of 98% validation accuracy in the process.

From figure 14 of Validation curve, a stepped rise in accuracy can be seen beginning from 10% where the accuracy reaches its epitome by $20^{th}$ iteration of 96%. Further iterations maintain the accuracy of 96% till the end of experiment.

Finally, in terms of Type 1 and Type 2 errors, EM matches type 1 score of K-Means of 55. However, an ideal performance in terms of Type 2 error is shown by EM by classifying 0 False Negative samples, outperforming the baseline model[2] in the process.

## V. Discussion

### A. Company Bankruptcy dataset

Beginning with the clustering algorithms, K-Means indicated lower clusters as compared to EM evident from their respective elbow plots. Here for K-Means, randomly assigned initial centroids worked well as that of "K-Means++". This is due to the continuous feature space of the dataset where various data points overlap in the input space. For EM, "full" covariance indicated lower BIC score as that of "tied", "spherical" and "diagonal" because each component computes its own general covariance matrix and since the dataset is not standardized, single variance is not guaranteed, hence spherical and diagonal covariance failed here. Here, the null hypothesis that EM would generate lower clusters as that of K-Means is rejected. Also, the elbow plot of K-Means when fit with label and without label show no difference since the specification of target label is ignored internally for both the clustering algorithms.

Secondly, with 90% variance ratio, PCA reduced the entire feature space to 7 components, outperforming ICA and RP. This is due to the fact that the linear projection of PCA suited well for the continuous feature space dataset since all the data points resides in close proximity with each other. Since ICA assumes the independent features existence, it tries to uncover the independent feature set however, the continuous feature space makes it difficult for ICA to reduce the dataset into similar components produced by PCA. Also, RP following Johnson-Lindenstrauss Lemma, suggests more than 100 components for data samples greater than 5000 which results in the RP indicating original number of features. The null hypothesis that ICA would suffer here is accepted however, RP failed to out-perform ICA though PCA illustrated optimal performance.

Further, applying K-Means on PCA reduced dataset illustrate no effect of dimension reduction as the resultant clusters here is also 7 seen in elbow plot of figure 7. Whereas, ICA reduced dataset for K-Means indicate 9 optimal clusters, exceeding the number obtained over original dataset. However, the results for Gaussian RP indicate similar results as that of K-Means with PCA reduced dataset evident from elbow plot of figure 7. The identical performance of K-Means with PCA and K-Means with RP is because the dimensions when reduced to half of the number of original dataset, makes RP to work similar to that of PCA however, the choice of the components by threshold makes RP to suffer from noise.

On the other hand, the BIC score for EM with PCA does not align with the results of K-Means with PCA reduced dataset and suggest higher clusters. Similarly, EM with Gaussian RP reduced dataset resides in close proximity as that with PCA. However, EM with ICA reduced dataset illustrates lowest number of clusters which is close to that of ground truth labels. The reason behind EM with ICA outperforming EM with PCA, and RP is the ICA reduced dataset has highly separated features than that of PCA followed by RP. Also, K-Means in general is biased towards spherical distribution of data whereas EM is flexible enough with support for Diagonal distribution also. Additionally, the probabilistic estimation in terms of log-likelihood of EM makes it superior to K-Means here in case of ICA reduced dataset since the feature set is highly un-correlated. Here, the null hypothesis that K-Means with ICA would perform poor is accepted however, EM with ICA showcase lower clusters.

Comparing neural networks with PCA, ICA and RP reduced datasets shows a better learning rate for ICA whereas, an identical learning curve evident in case of PCA and RP as the lower dimensional space makes if less feasible for RP to work upon, making it similar to PCA. However, all the

three techniques outperform the baseline model because the existence of higher features act as noise to the models which is removed by these dimension reduction algorithms. The performance is evident from type 1 and type 2 error of table 1. Here, the null hypothesis of Neural network with PCA will converge better with high accuracy is rejected.

Finally, the performance of neural network with K-Means and EM indicate the strength of log-likelihood where the density of data point belonging to a cluster is used as input space. Also, the data distribution is not spherical here which is favored by K-Means and hence the support of EM to various distributions provide better feature space. Here, the null hypothesis of EM outperforming K-Means is accepted.

### *B. Diabetes Classification dataset*

First, K-Means and EM illustrate similar optimal clusters since the dataset is linearly separable, making the data points highly distinguishable between the clusters. This is evident from the performance of EM with the hyper-parameters following K-Means model with Spherical convergence with k-means function. As a result, both the models indicate similar optimal number of clusters resulting in acceptance of null hypothesis. Also, the elbow plot of K-Means when fit with label and without label show no difference since the specification of target label is ignored internally for both the clustering algorithms.

Further, following 95% variance, PCA indicate 3 principal components however, the kurtosis plot indicate 10 components. The lower components indicated by PCA cannot be considered optimal here because PCA works by identifying the variance between the data but, since this dataset contains categorical and binary features majorly, PCA cannot grasp the information from variance during projection and favor one of the value from the categorical variable. On the other hand, Sparese RP suffers from the problem of low dimensional features of the original dataset. Hence, the reconstruction error converges to 0 for original number of features. Hence, the resultant features also shows minor degree of correlation. However, ICA does not appear to be suffering from lower dimensional problem as well as nature of the dataset's features. Hence, it spontaneously captures information within 10 independent and non-correlated features. Here, the null hypothesis that RP would work poorly on this dataset is accepted. This result of RP is highly likely since Johnson Lindenstrauss lemma suggest 25 features for 0.9 eps.

Since PCA did not perform well over categorical variables, its effect is evident from elbow plot of K-Means with PCA and EM with PCA where both the clustering algorithms indicate similar clusters once again. Similary, ICA and Sparse RP follows the trend with 6 optimal clusters. However, the Sum of Squared Distance(SSD) for ICA at 6 clusters is much less than that of PCA and RP standing at 0.2 and 0.38 for ICA, PCA and RP respectively. These metrics showcase the ideal number of components generated by ICA which K-Means and EM are able to cluster the data with lower SSD. However, the cluster labels do not fit well with the ground truth labels that are binary. As a result, the null hypothesis that K-Means K-Means with ICA would outperform remaining combinations is rejected given the similar performance in all the cases.

## VI. Conclusion

K-Means and Expectation Maximization(EM) algorithms are implemented over Company Bankruptcy dataset and Diabetes Classification dataset. Both the datasets have binary prediction label however, their feature spaces vary over continuous and categorical features. First the effect of continuous feature space is evident where K-Means clustered the overlapping data points into spherical structures. However, EM uses probabilistic placement of data point in the clusters and given the nature of the dataset, the data point might be available in more than one cluster. This difference is clearly evident when compared with second dataset which has categorical and hence the linearly separable data points make it highly suitable for EM. Additionally, the introduction of target label while fitting the clustering algorithms do not have any influence on the clusters since it is ignored by the algorithms. However, other metrics such as Silhouette score, Rand Index(RI) can be used to identify the number of clusters.

The effect of dataset is also evident from dimensionality reduction algorithms between both the datasets. PCA in case of Company Bankruptcy dataset, suggest lower dimensions than that of ICA and RP whereas, PCA in case of Diabetes Classification dataset cannot be much trusted due to its linear projection. ICA in this case discard the biased compress of information towards single feature value and generates dimensions with non-correlated features. RP on the other hand, suffers from the curse of dimensions as it is suitable for large feature space and none of the two dataset follows this notion. Hence, it generates output identical to PCA or output correlated reduced dimensions.

However, dimensionality reduction and clustering technique performs tremendously well by replacing the original dataset either by the intra-cluster distance in case of K-Means, Density functions of feature space in case of EM and the reduced dimensions of PCA, ICA and RP. The results of these two technique showcase better type 1 and type 2 scores when compared with the baseline model trained over the original datasets. Also, the reduced dataset and the usgae of cluster space to replace the entire original feature space reduces the wall clock time for training the neural network in both the caes.

## References


[1] Umesh Jadhav, Supervised Learning, https://www.overleaf.com/read/vbgxwtzrmyxc#3b06b5.

[2] Umesh Jadhav, Comparative study of Randomized Optimization Algorithms for N-Queens problem, Knapsack Problem, and Neural Network weights optimization, https://www.overleaf.com/read/rtwmvgxzkwdp#4a63fd.

[3] Fedesoriano, Company Bankruptcy Prediction, https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction.

[4] Priyam Chowksi, Comprehensive Diabetes Clinical Dataset(100k rows), https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset, July 2024.