August - December 2018
Odd Semester
CS669: Pattern Recognition
Programming Assignment 4
Preethi Srinivasan S18001
Ganesan S18005

1. Introduction

When the number of features or dimensions in the data increases, it makes the training extremely slow, the Principal Component Analysis (PCA) technique is used to reduce the dimension of the data while keeping the information of the original features as close as possible. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

Often, the desired goal is to reduce the dimensions of a *d*-dimensional dataset by projecting it onto a (*l*)-dimensional subspace (where $l<d$) in order to increase the computational efficiency while retaining most of the information. An important question is "what is the size of *l* that represents the data 'well'?"
Later, we will compute eigenvectors (the principal components) of a dataset and collect them in a matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the "length" or "magnitude" of the corresponding eigenvector.
In this assignment, we performed Principal Component Analysis (PCA) on 32D Bag of Visual Words (BoVW) representation for scene images. In each class we had 50 images for training and 50 images for testing. In assignment 2, we were unable to build GMM for mixtures >= 2, on this BoVW features due to non-convergence of EM algorithm. This was due to curse of dimensionality problem.
Here we reduce the dimension of the BoVW features using PCA. The GMM was built on this reduced dimensional feature. Log Likelihood graphs for l = 1 is shown for all mixture components. Beyond that, Confusion matrix for different number of principal components and GMM mixtures are plotted.

Summary of the PCA approach performed in this assignment:

1. Take all the training examples including all classes
2. Compute the Covariance Matrix
3. Perform Eigen analysis to obtain ‘d’ eigenvalues and the corresponding eigenvectors
4. Choose ‘l’ eigenvectors corresponding to ‘l’ leading eigenvalues (significant eigenvalues)
5. Project each of the $y_n = x_n$ - mu (mean subtracted data for mathematical ease) on to the ‘l’ eigenvalues to get ‘l’ principal components.
6. Perform GMM on the reduced data and compute the Performance matrix.

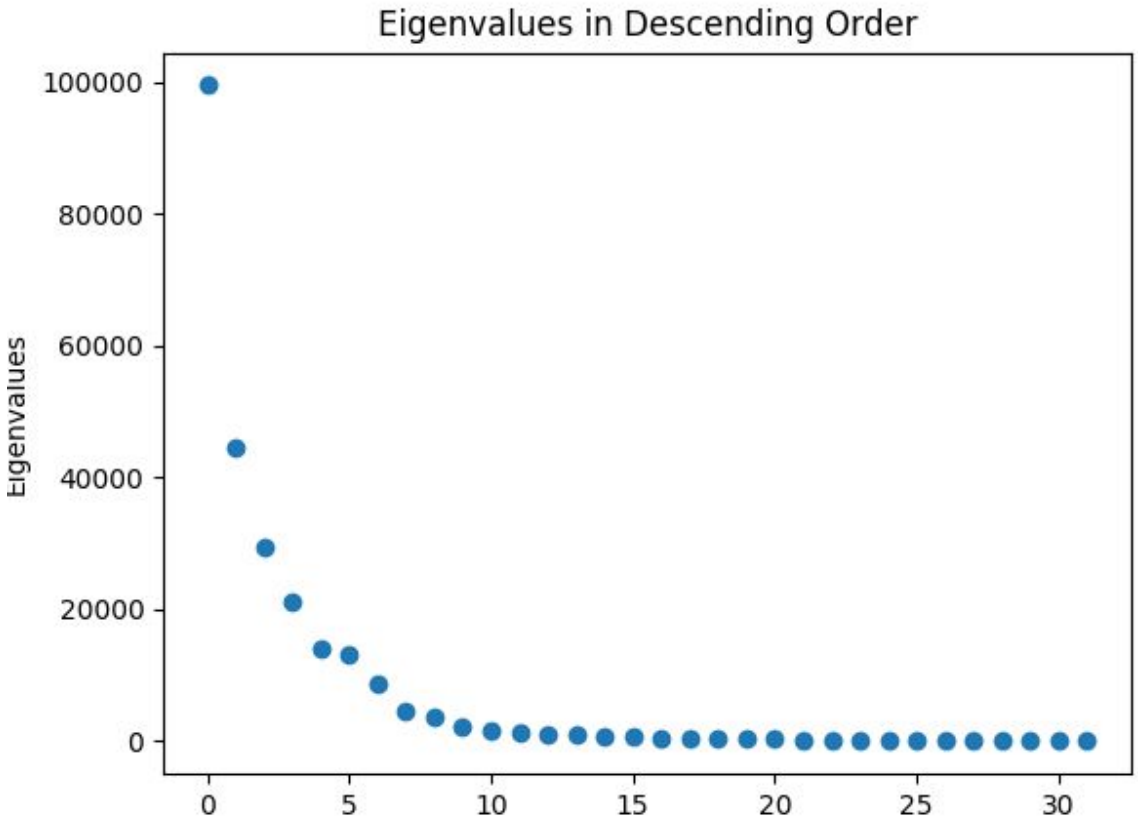

Figure 1.1: Plot of Eigenvalues in Descending Order

The plot above clearly shows that most of the variance (39.66% of the variance to be precise) can be explained by the first principal component alone. The second principal component still bears some information (17.78%). Similarly as the later eigenvalues carry lesser information. Together, the first 10 eigenvalues carry 96.26% information. Rest of them can be safely dropped.

Number of principal components = 1

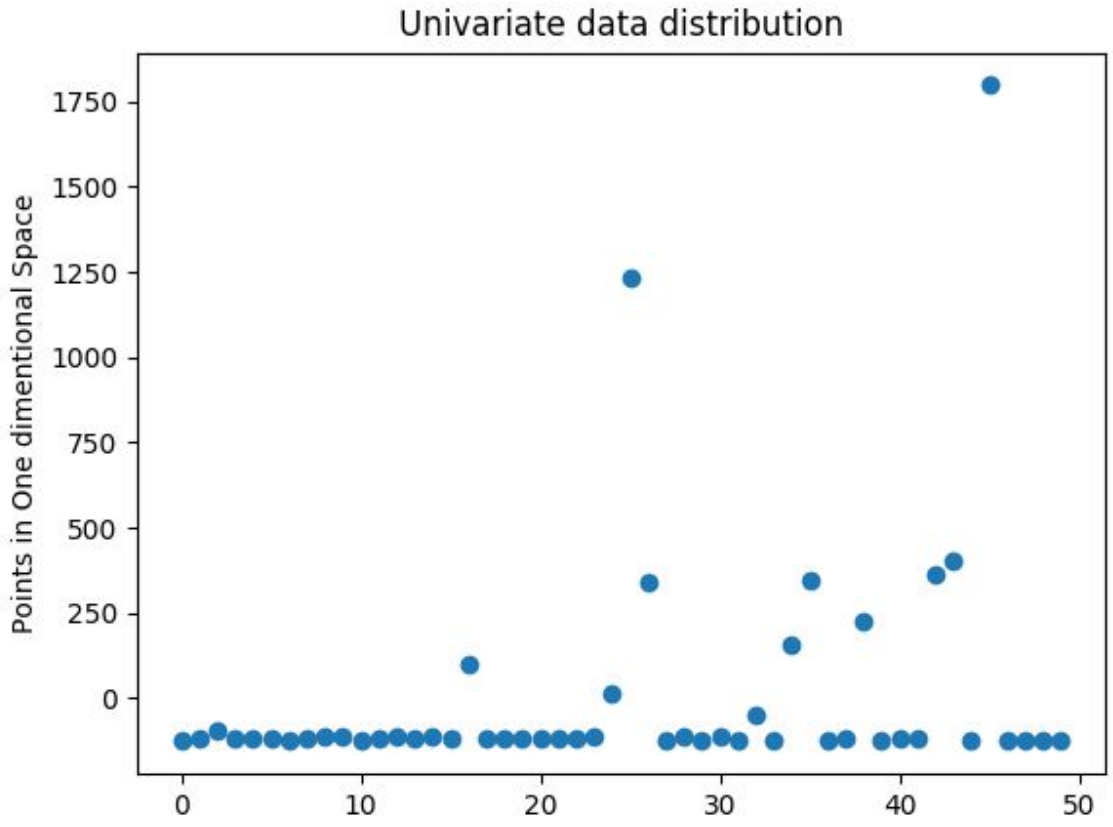

Figure 1.2 : Univariate data(50 features) distribution of Class "Coast" after PCA with l = 1.

As it can be seen from the above figure, the data is distributed in a very non uniformly fashion. Some of the clusters contain only 1 point. For example, in the above case, point 1760 is the outlier which is falling singly in to a cluster. Covariance matrix of such a cluster returns 'nan' because finding covariance of a single point is not defined. So, we hard coded to return 0 when ever Covariance of matrix is resulting in 'nan'. This issue occurs when the number of clusters is increased (as in k = 4 and above).

Number of principal components = 1, Number of GMM mixtures = 1

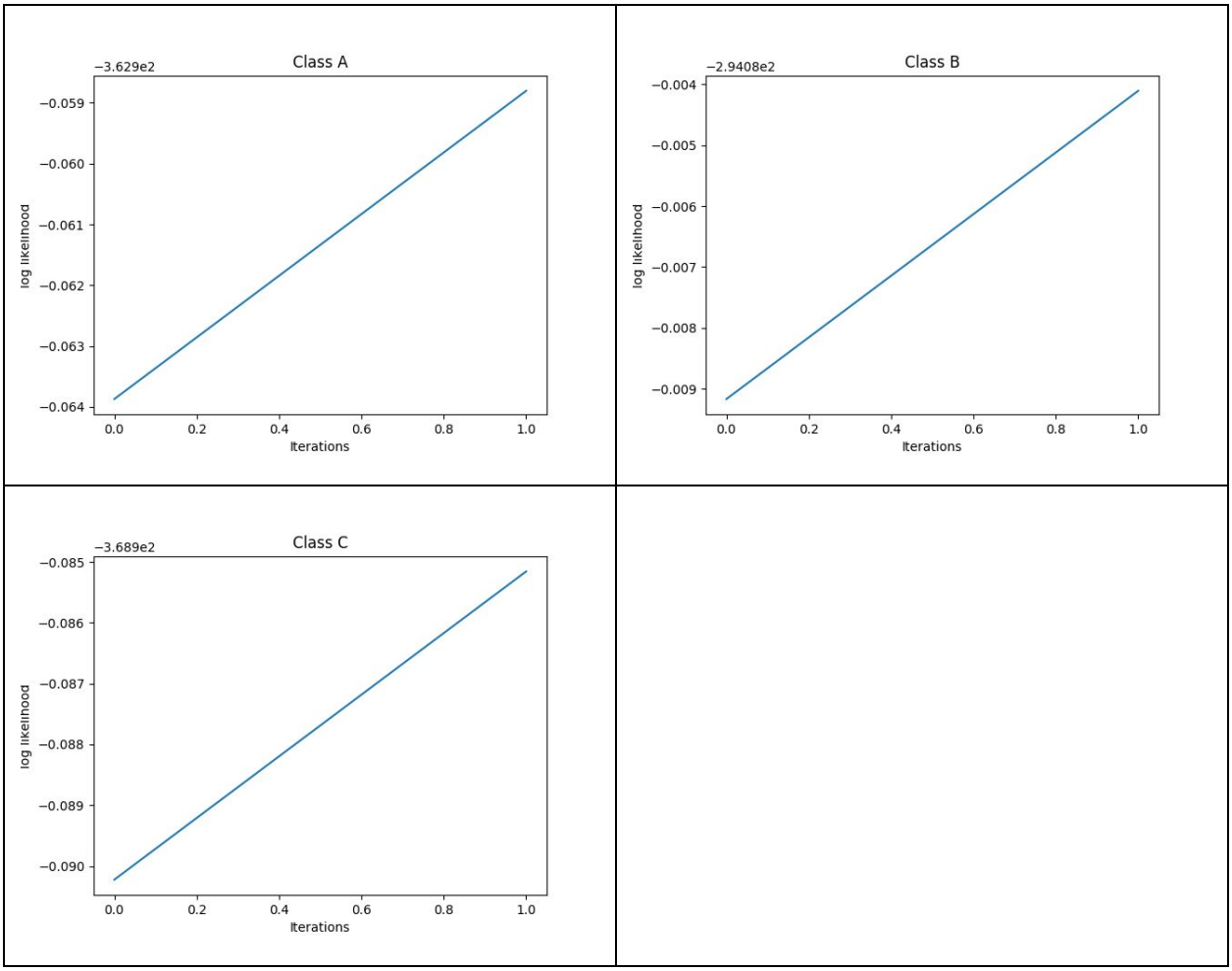

Figure 1.3 : Log-likelihood graph of GMM convergence with l=1, k=1

Table 1.1 Confusion Matrix for the classifier after PCA reduction with l = 1, k = 1

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*5*	*42*	*3*
	Industrial Area	*1*	*47*	*2*
	Pagoda	*9*	*23*	*18*

Table 1.2 the performance matrix for the classifier with l =1, k =1

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Coast	*33.33*	10.0	15.38
Industrial Area	41.96	94.0	58.02
Pagoda	78.26	36.0	49.32
Mean Value	51.19	46.67	40.91

Class Accuracy: 46.66

Number of principal components = 1, Number of GMM mixtures = 2

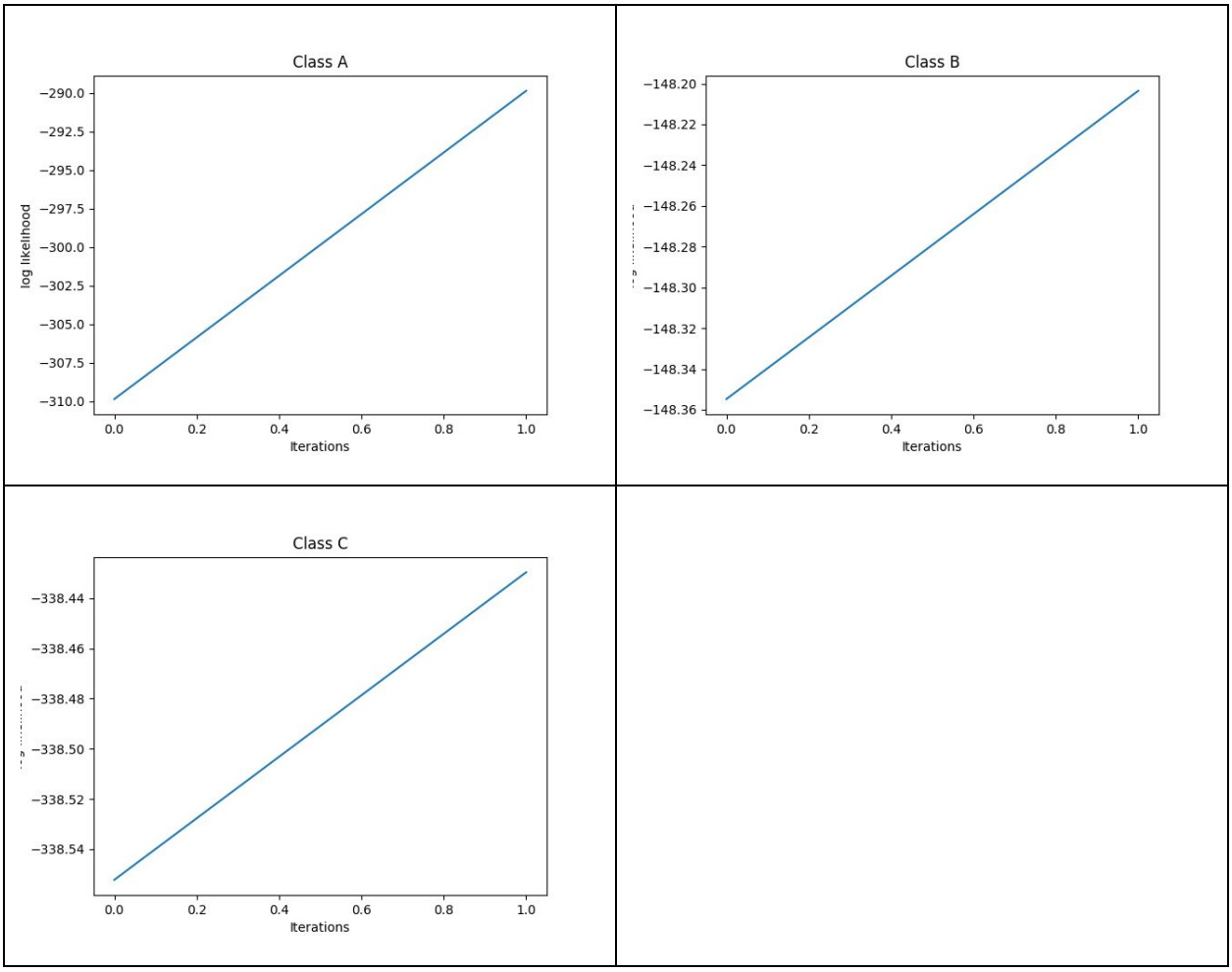

Figure 1.4 : Log-likelihood graph of GMM convergence with l=1, k=2

Table 1.3 Confusion Matrix for the classifier after PCA reduction with l = 1, k = 2

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*8*	*33*	*9*
	Industrial Area	*3*	*44*	*3*
	Pagoda	*21*	*8*	*21*

Table 1.4 the performance matrix for the classifier with l =1, k =2

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*25.0*	16.0	19.51
Industrial Area	51.79	88.0	65.19
Pagoda	63.64	42.0	50.6
Mean Value	46.8	48.67	45.1

Class Accuracy: 48.66%

Number of principal components = 1, Number of GMM mixtures = 4

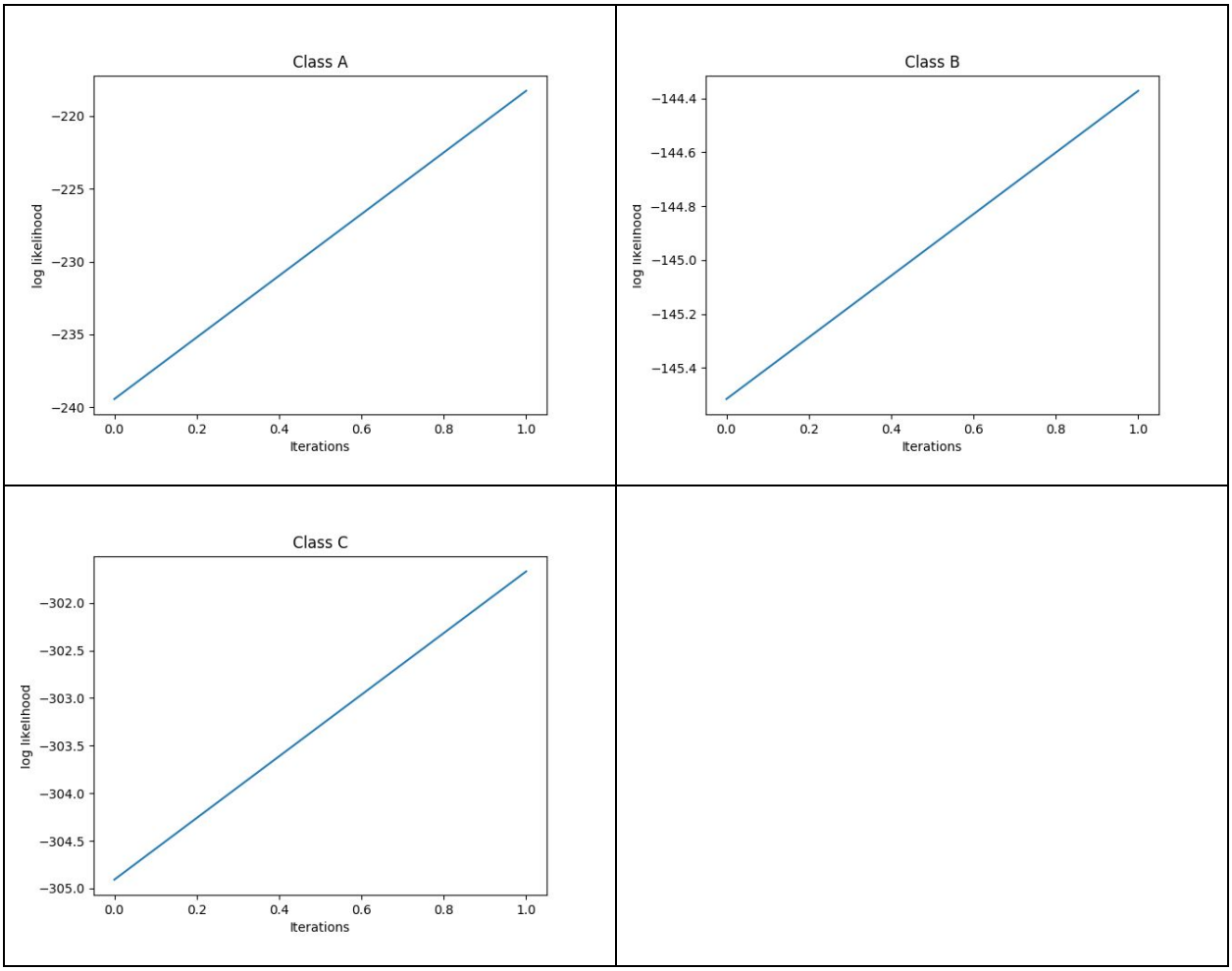

Figure 1.5 : Log-likelihood graph of GMM convergence with l=1, k=4

Table 1.5 Confusion Matrix for the classifier after PCA reduction with l = 1, k = 4

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*11*	*27*	*12*
	Industrial Area	*18*	*27*	*5*
	Pagoda	*10*	*7*	*33*

Table 1.6 the performance matrix for the classifier with l =1, k =4

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*28.21*	22.0	24.72
Industrial Area	44.26	54.0	48.65
Pagoda	66.0	66.0	66.0
Mean Value	46.16	47.33	46.46

Class Accuracy: 47.33%

Number of principal components = 1, Number of GMM mixtures = 8

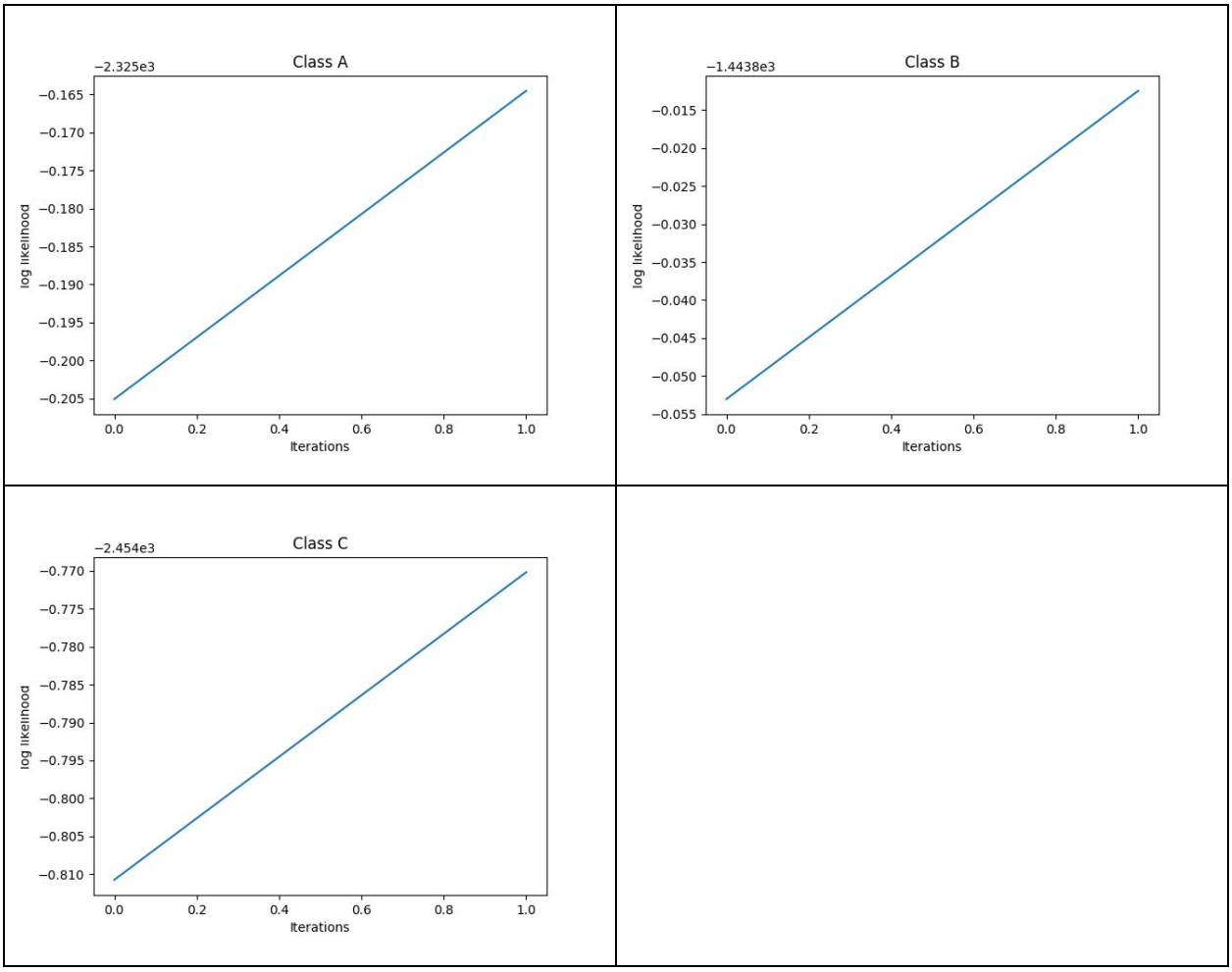

Figure 1.6 : Log-likelihood graph of GMM convergence with l=1, k=8

Table 1.7 Confusion Matrix for the classifier after PCA reduction with l = 1, k = 8

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*16*	*21*	*13*
	Industrial Area	*16*	*27*	*7*
	Pagoda	*7*	*7*	*36*

Table 1.8 the performance matrix for the classifier with l =1, k =8

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*41.03*	32.0	35.96
Industrial Area	49.09	54.0	51.43
Pagoda	64.29	72.0	67.92
Mean Value	51.47	52.67	51.77

Class Accuracy: 52.66%

Number of principal components = 2, Number of GMM mixtures = 1

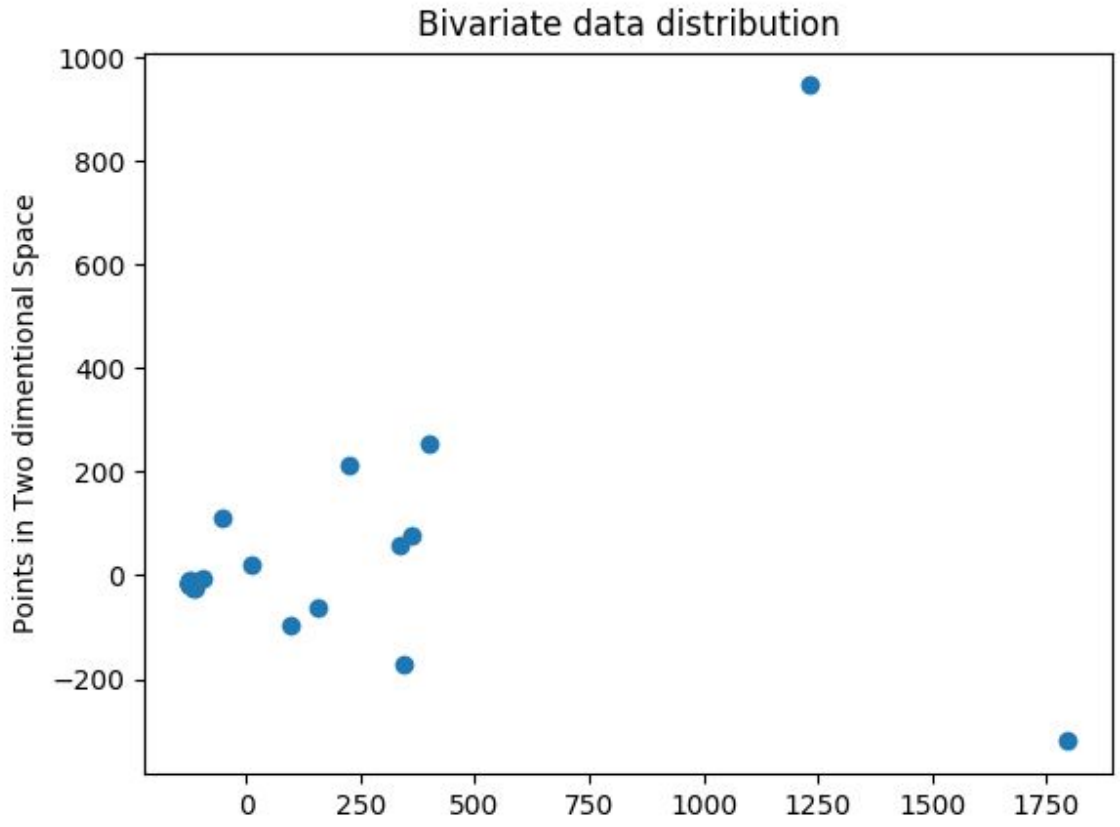

Figure 1.7 : Multivariate data(50 features) distribution of a Class "Coast" after PCA with l = 2.

Above figure shows the data distribution of 2d points after reducing the data with l = 2. From the scatter plot, we can see that points are scattered sparsely. So when multiple mixtures are considered, it is possible that some clusters contain only 1 point leading to undefined covariance matrices.

Table 1.9 Confusion Matrix for the classifier after PCA reduction with l = 2, k = 1

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*11*	*38*	*1*
	Industrial Area	*3*	*47*	*0*
	Pagoda	*18*	*18*	*14*

Table 1.10 the performance matrix for the classifier with l =2, k =1

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*34.38*	22.0	26.83
Industrial Area	45.63	94.0	61.44
Pagoda	93.33	28.0	43.08
Mean Value	57.78	48.0	43.78

Class Accuracy: 48.0%

Number of principal components = 2, Number of GMM mixtures = 2

Table 1.11 Confusion Matrix for the classifier after PCA reduction with l = 2, k = 2

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*6*	*38*	*6*
	Industrial Area	*2*	*48*	*0*
	Pagoda	*10*	*25*	*15*

Table 1.12 the performance matrix for the classifier with l =2, k =2

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*33.33*	12.0	17.65
Industrial Area	43.24	96.0	59.63
Pagoda	71.43	30.0	42.25
Mean Value	49.34	46.0	39.84

Class Accuracy: 46.0%

Number of principal components = 2, Number of GMM mixtures = 4

Table 1.13 Confusion Matrix for the classifier after PCA reduction with l = 2, k = 4

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*5*	*30*	*15*
	Industrial Area	*9*	*36*	*5*
	Pagoda	*8*	*7*	*35*

Table 1.14 the performance matrix for the classifier with l =2, k =4

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*22.73*	10.0	13.89
Industrial Area	49.32	72.0	58.54
Pagoda	63.64	70.0	66.67
Mean Value	45.23	50.67	46.36

Class Accuracy: 50.66%

Number of principal components = 2, Number of GMM mixtures = 8

Table 1.15 Confusion Matrix for the classifier after PCA reduction with l = 2, k = 8

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*29*	*11*	*12*
	Industrial Area	*26*	*24*	*4*
	Pagoda	*15*	*7*	*28*

Table 1.16 the performance matrix for the classifier with l =2, k =8

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*41.43*	58.0	48.33
Industrial Area	52.63	40.0	45.45
Pagoda	66.67	56.0	60.87
Mean Value	53.58	51.33	51.55

Class Accuracy: 51.33%

Number of principal components = 5, Number of GMM mixtures = 1

Table 1.17 Confusion Matrix for the classifier after PCA reduction with l = 5, k = 1

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*12*	*34*	*4*
	Industrial Area	*4*	*45*	*1*
	Pagoda	*19*	*11*	*20*

Table 1.18 the performance matrix for the classifier with l =5, k = 1

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*34.29*	24.0	28.24
Industrial Area	50.0	90.0	64.29
Pagoda	80.0	40.0	53.33
Mean Value	54.76	51.33	48.62

Class Accuracy: 51.33%

Number of principal components = 5, Number of GMM mixtures = 2

Table 1.19 Confusion Matrix for the classifier after PCA reduction with l = 5, k = 2

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*8*	*34*	*8*
	Industrial Area	*3*	*45*	*2*
	Pagoda	*12*	*12*	*26*

Table 1.20 the performance matrix for the classifier with l =5, k = 2

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*34.78*	16.0	21.92
Industrial Area	49.45	90.0	63.83
Pagoda	72.22	52.0	60.47
Mean Value	52.15	52.67	48.74

Class Accuracy: 52.66%

Number of principal components = 5, Number of GMM mixtures = 4

Table 1.21 Confusion Matrix for the classifier after PCA reduction with l = 5, k = 4

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*7*	*31*	*12*
	Industrial Area	*9*	*39*	*2*
	Pagoda	*14*	*7*	*29*

Table 1.22 the performance matrix for the classifier with l =5, k = 4

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*23.33*	14.0	17.5
Industrial Area	50.65	78.0	61.42
Pagoda	67.44	58.0	62.37
Mean Value	47.14	50.0	47.09

Class Accuracy: 50.0%

Number of principal components = 5, Number of GMM mixtures = 8

Table 1.21 Confusion Matrix for the classifier after PCA reduction with l = 5, k = 8

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*20*	*16*	*14*
	Industrial Area	*16*	*30*	*4*
	Pagoda	*13*	*10*	*27*

Table 1.22 the performance matrix for the classifier with l = 5, k = 8

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*40.82*	*40.0*	*40.4*
Industrial Area	*53.57*	*60.0*	*56.6*
Pagoda	*60.0*	*64.0*	*56.84*
Mean Value	*51.46*	*51.33*	*51.28*

Class Accuracy: 51.33%

Number of principal components = 10, Number of GMM mixtures = 1

Table 1.23 Confusion Matrix for the classifier after PCA reduction with l = 10, k = 1

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*11`*	*34*	*5*
	Industrial Area	*4*	*44*	*2*
	Pagoda	*15*	*11*	*24*

Table 1.24 the performance matrix for the classifier with l =10, k = 1

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*36.67*	22.0	27.5
Industrial Area	49.44	88.0	63.31
Pagoda	77.42	48.0	59.26
Mean Value	54.51	52.67	50.02

Class Accuracy: 52.66%

Number of principal components = 10, Number of GMM mixtures = 2

Table 1.25 Confusion Matrix for the classifier after PCA reduction with l = 10, k = 2

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*4*	*34*	*12*
	Industrial Area	*3*	*44*	*3*
	Pagoda	*10*	*11*	*29*

Table 1.26 the performance matrix for the classifier with l =10, k = 2

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*23.53*	8.0	11.94
Industrial Area	49.44	88.0	63.31
Pagoda	65.91	58.0	61.7
Mean Value	46.29	51.33	45.65

Class Accuracy: 51.33%

Number of principal components = 10, Number of GMM mixtures = 4

Table 1.27 Confusion Matrix for the classifier after PCA reduction with l = 10, k = 4

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*8*	*31*	*11*
	Industrial Area	*6*	*37*	*7*
	Pagoda	*14*	*8*	*28*

Table 1.28 The performance matrix for the classifier with l =10, k = 4

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*28.57*	16.0	20.51
Industrial Area	48.68	74.0	58.73
Pagoda	60.87	56.0	58.33
Mean Value	46.04	48.67	45.86

Class Accuracy: 48.66%

Number of principal components = 12, Number of GMM mixtures = 1

Table 1.29 Confusion Matrix for the classifier after PCA reduction with l = 12, k = 1

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*8*	*33*	*9*
	Industrial Area	*4*	*42*	*4*
	Pagoda	*9*	*11*	*30*

Table 1.30 the performance matrix for the classifier with l =12, k = 1

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*38.1*	16.0	22.54
Industrial Area	48.84	84.0	61.76
Pagoda	69.77	60.0	64.52
Mean Value	52.23	53.33	49.61

Class Accuracy: 53.33%

Number of principal components = 12, Number of GMM mixtures = 2

Table 1.31 Confusion Matrix for the classifier after PCA reduction with l = 12, k = 2

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*6*	*33*	*11*
	Industrial Area	*6*	*41*	*3*
	Pagoda	*11*	*9*	*30*

Table 1.32 the performance matrix for the classifier with l =12, k = 2

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*26.09*	12.0	16.44
Industrial Area	49.4	82.0	61.65
Pagoda	68.18	60.0	63.83
Mean Value	47.89	51.33	47.31

Class Accuracy: 51.33%

Number of principal components = 12, Number of GMM mixtures = 4

Table 1.33 Confusion Matrix for the classifier after PCA reduction with l = 12, k = 4

	Class assigned by the Classifier			
Actual Values		*coast*	*Industrial Area*	*Pagoda*
	Coast	*10*	*29*	*11*
	Industrial Area	*6*	*38*	*6*
	Pagoda	*14*	*6*	*30*

Table 1.34 the performance matrix for the classifier with l =12, k = 4

	Precision (%)	***Recall Rate (%)***	***F Score (%)***
Coast	*33.33*	20.0	25.0
Industrial Area	52.05	76.0	61.79
Pagoda	63.83	60.0	61.86
Mean Value	49.74	52.0	49.55

Class Accuracy: 52.0%

Conclusion:

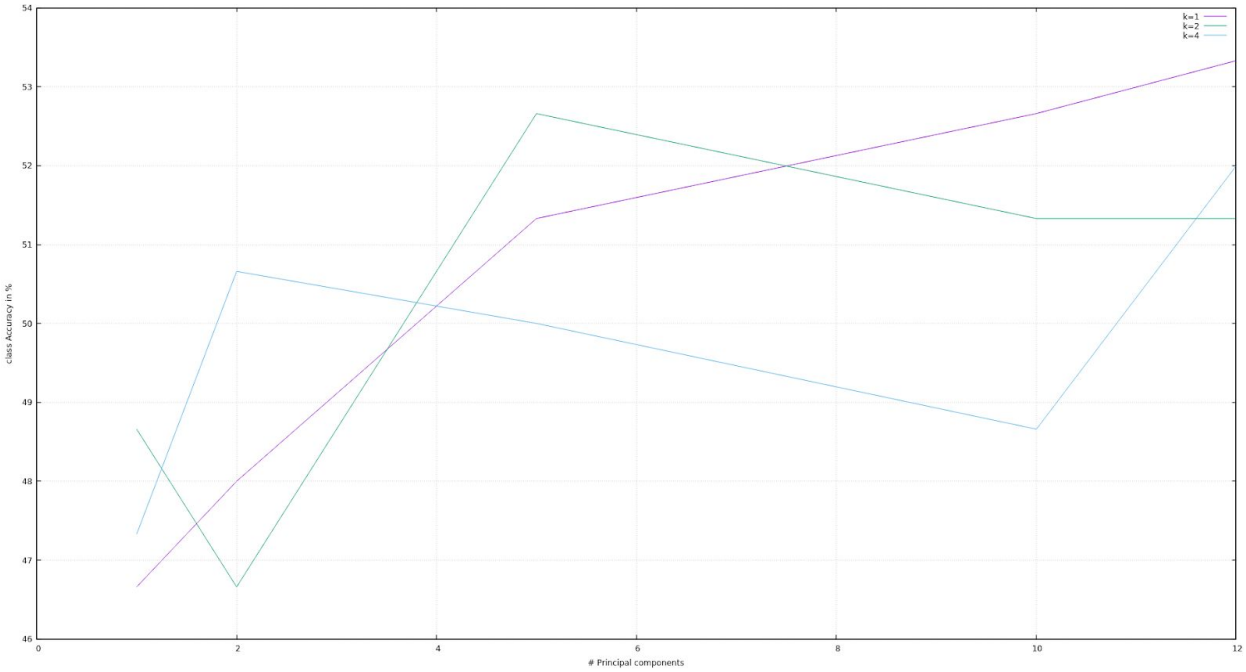

Figure 1.8: Comparison of class accuracy with varying value of l and k

We have considered maximum 12 principal components. Beyond 12 components, accuracy does not change. We have got maximum accuracy of 53.33% with l=12 and k=1. We observed that accuracy is increased with increase in the number of principal components. When we considered only first principal component (l = 1), the accuracy was about 46%. It slowly increased with increasing the 'l' value. This is because, the first few principal components

(precisely in our case around first 12 components) corresponds to the directions in which the overall information of the data set is maximum.

In assignment 2, when we ran BOVW of 32d data on GMM with number of mixtures = 1, accuracy was 44% and GMM did not converge beyond that. After reducing the dimension of these 32D features enabled us to overcome the curse of dimensionality problem for GMM and we successfully built GMM on these features using the PCA technique.

If the number of GMM mixtures are further increased, 50 data points in a class, will become too little data for the number of mixtures.