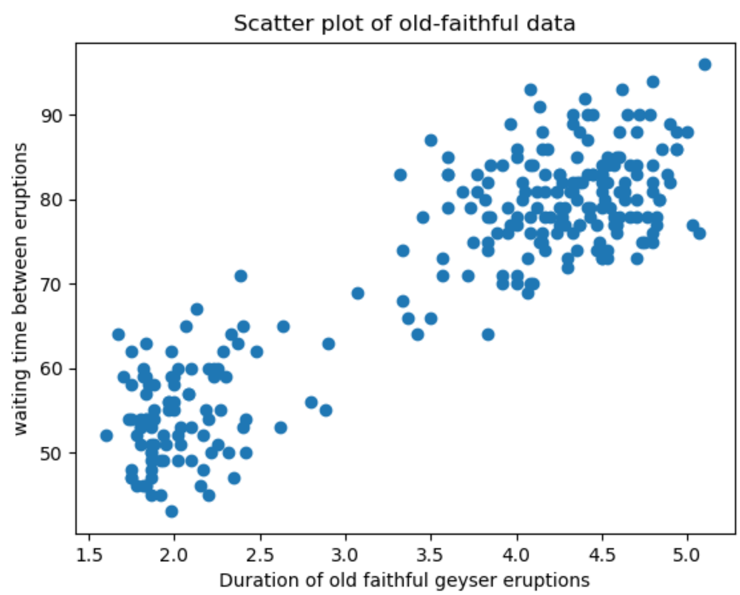
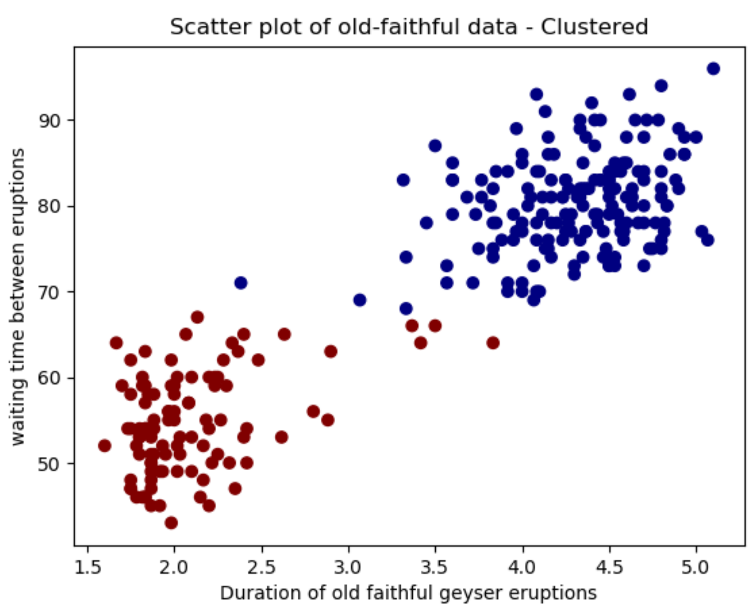
EE 511
SIMULATION METHODS FOR STOCHASTIC SYSTEMS
PROJECT – 3
ABINAYA MANIMARAN
SPRING 2018
03/23/2018

1. K-MEANS CLUSTERING ON "OLD FAITHFUL" DATASET

For the given Old Faithful Dataset,

- Feature 1: Duration of old faithful geyser eruptions
- Feature 2: Waiting time between eruptions
- Hence, Dimension of data = 2
- Scatter plot between both the features is shown below
- The data points were clustered using K-Means clustering, in which the centroids were initialized using k-means++ algorithm
- Clustered scatter plot is also shown below (different colors)





2. GAUSSIAN MIXTURE MODEL USING EXPECTATION MAXIMIZATION:

For this experiment, 4 datasets were used:

a) Spherical Covariance Matrix:
   - Generated using make_blobs function with (-5,0) and (0,1.5) as centers
   - Note that the centers are far apart

b) Elliptical Covariance Matrix:
   - Generated using make_blobs function with (-5,0) and (0,1.5) as centers
   - Transformed using ((0.4,0.2),(-0.4,1.2)) as transformation functions

c) Poorly Separated sub-populations:
   - Generated using make_blobs function with (-5,0) and (-4,0) as centers
   - Note that the centers are very close

d) Old-Faithful data given

GMM Clustering with Expectation Maximization was implemented using the following algorithm:
(Source: https://www.youtube.com/watch?v=ZBLyXgjBx3Q&feature=youtu.be)

- Initialize Means μ_j, Covariance Matrix Σ_j and Mixing coefficients Π_j
- Expectation Step:

$$\gamma_k(x) = \frac{\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x \mid \mu_j, \Sigma_j)}$$

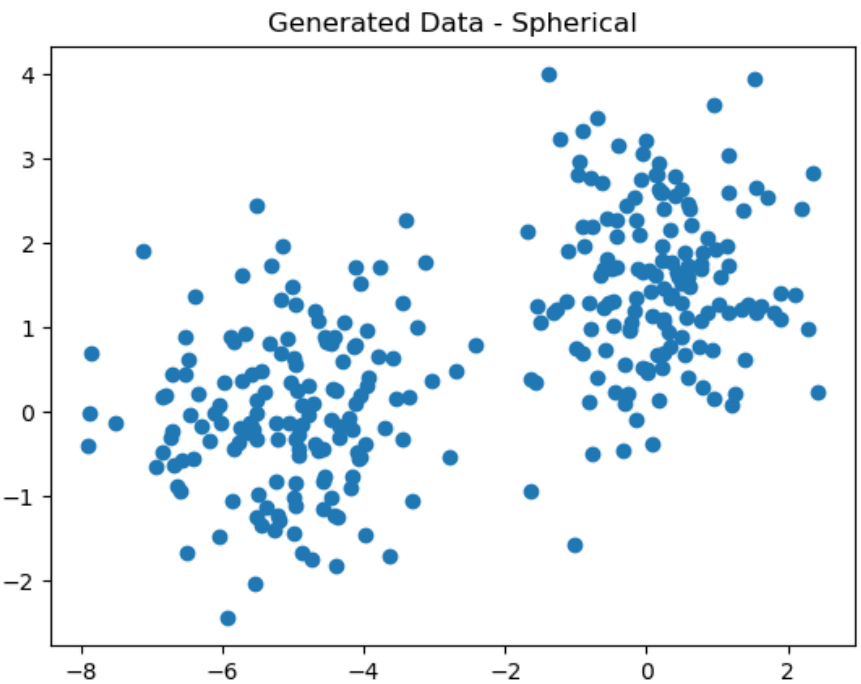
- Maximization Step: Update all the parameters

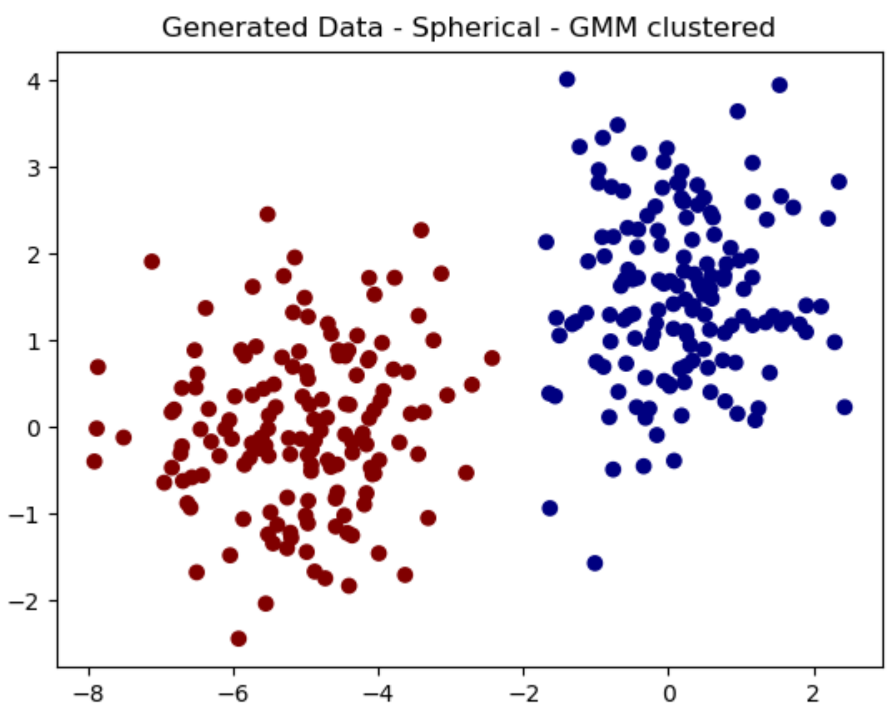
$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)} \qquad \boldsymbol{\pi}_j = \frac{1}{\mathbf{N}} \sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)$$

$$\Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(x_n)}$$

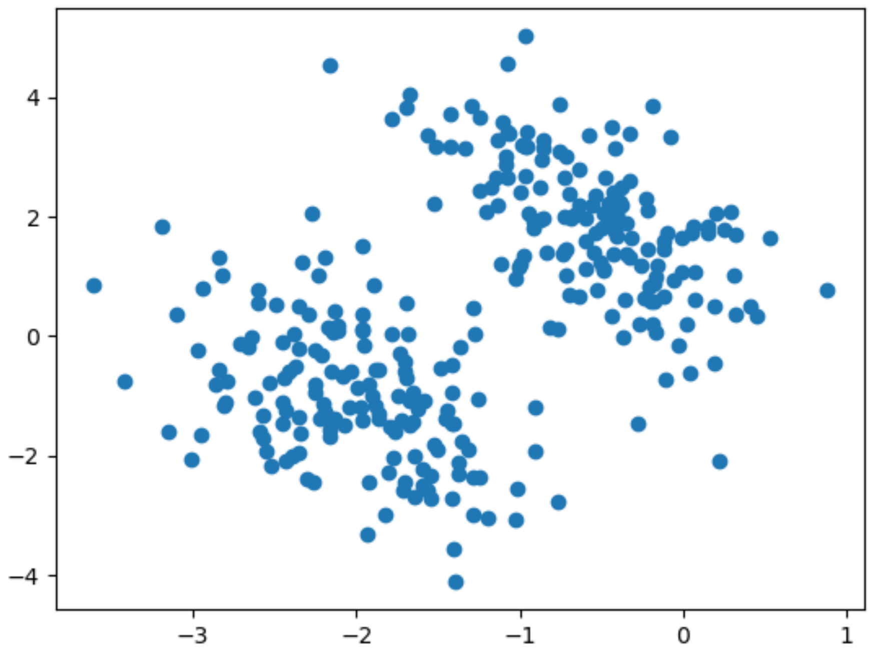
- Calculate Log Likelihood for convergence check

$$\ln p(\mathbf{X} \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathrm{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right\}$$

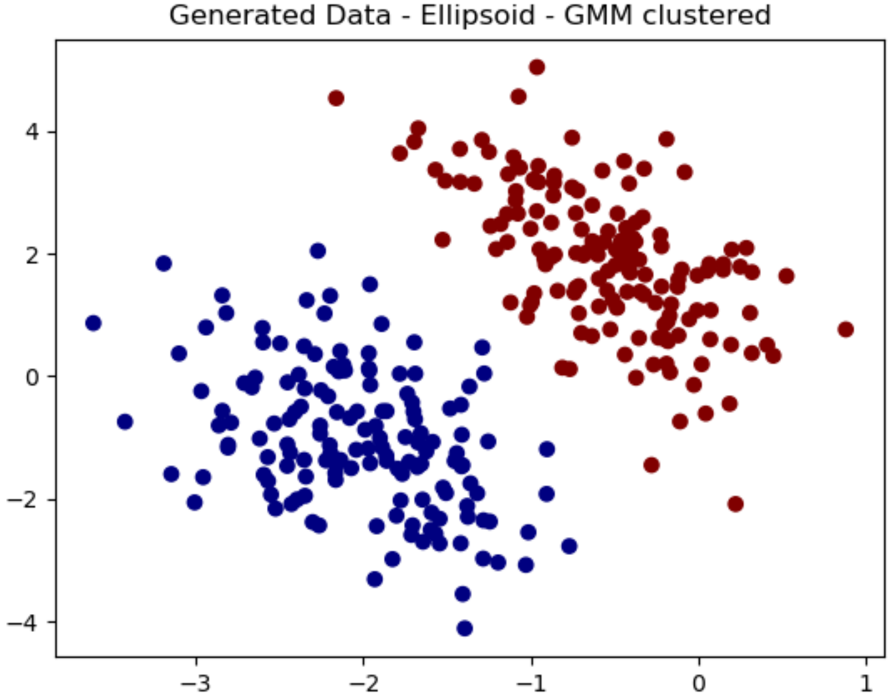

Generated Data - Spherical
4
3
2
1
0
−1
−2
−8
−6
−4
−2
0
2

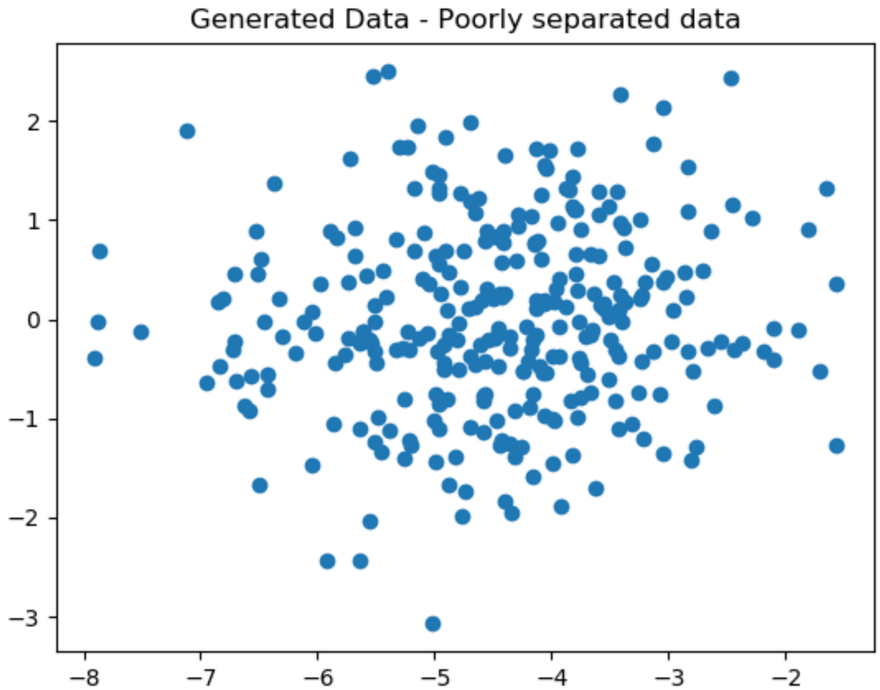

Generated Data - Spherical - GMM clustered
4
3
2
1
0
−1
−2
−8
−6
−4
−2
0
2

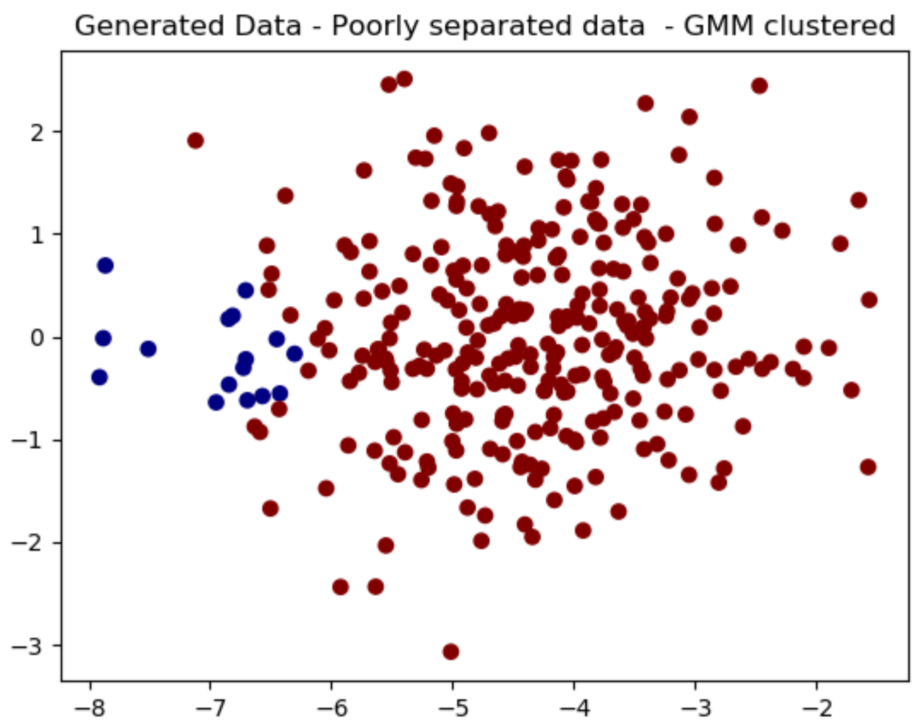
Generated Data - Ellipsoid

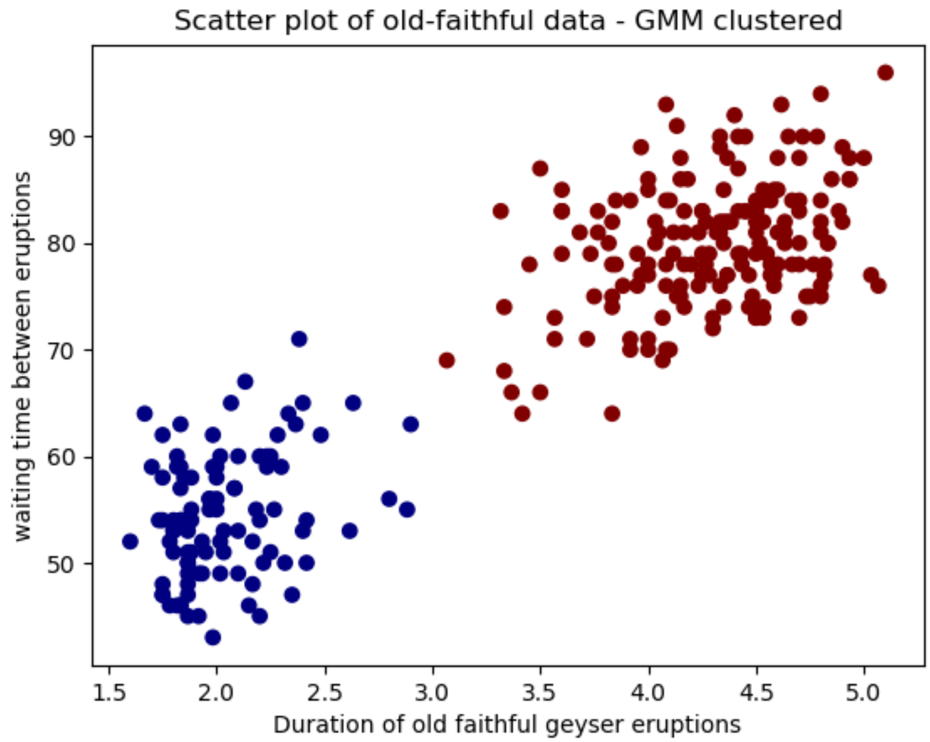


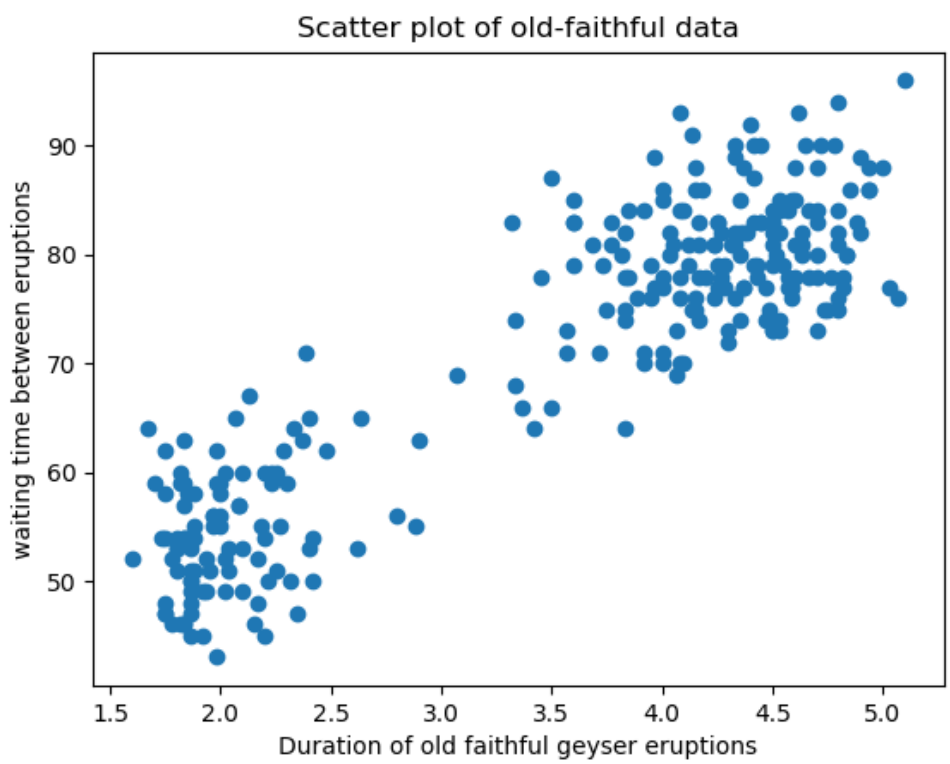
Generated Data - Ellipsoid - GMM clustered

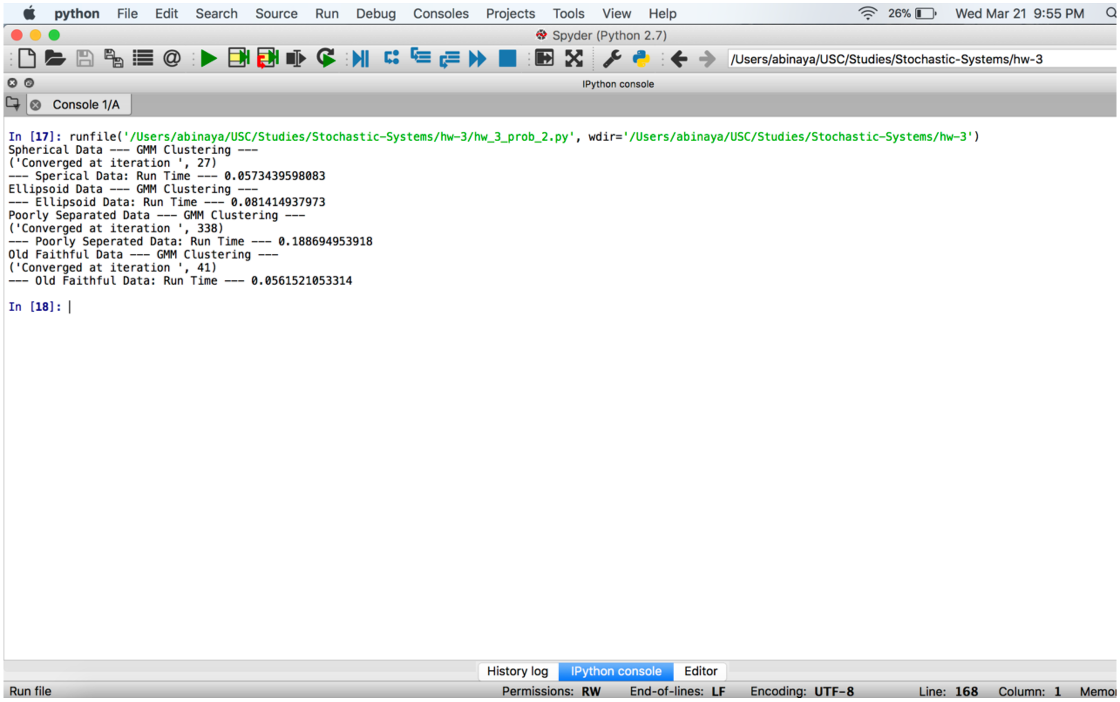



Generated Data - Poorly separated data
2
1
0
−1
−2
−3
−8
−7
−6
−5
−4
−3
−2


Generated Data - Poorly separated data - GMM clustered
2
1
0
−1
−2
−3
−8
−7
−6
−5
−4
−3
−2


Scatter plot of old-faithful data - GMM clustered
waiting time between eruptions
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Duration of old faithful geyser eruptions


Scatter plot of old-faithful data
waiting time between eruptions
90
80
70
60
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Duration of old faithful geyser eruptions

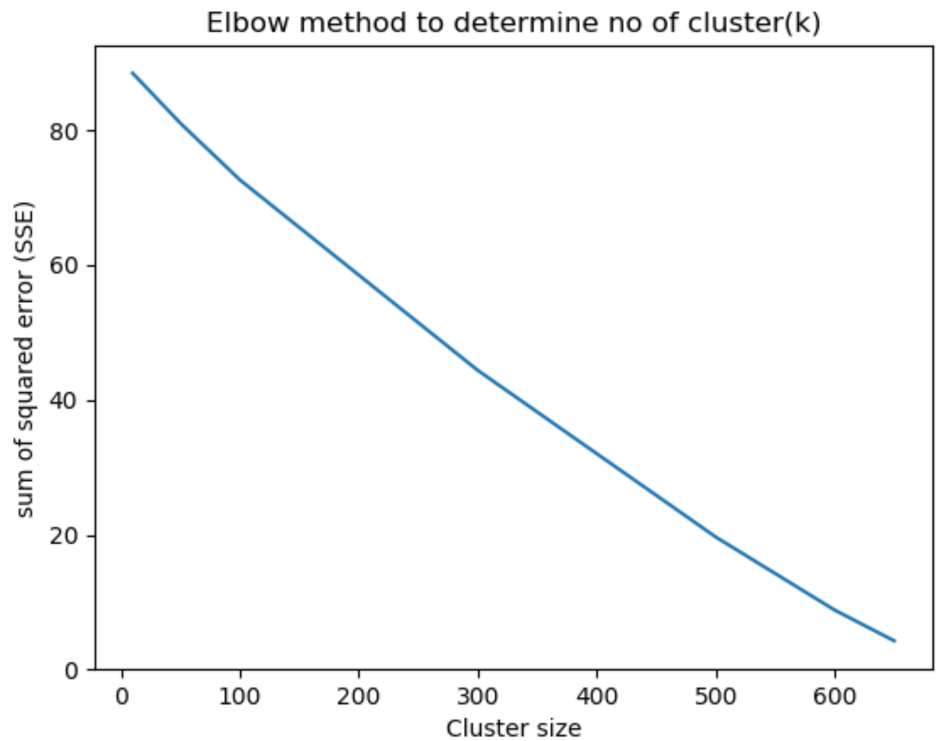


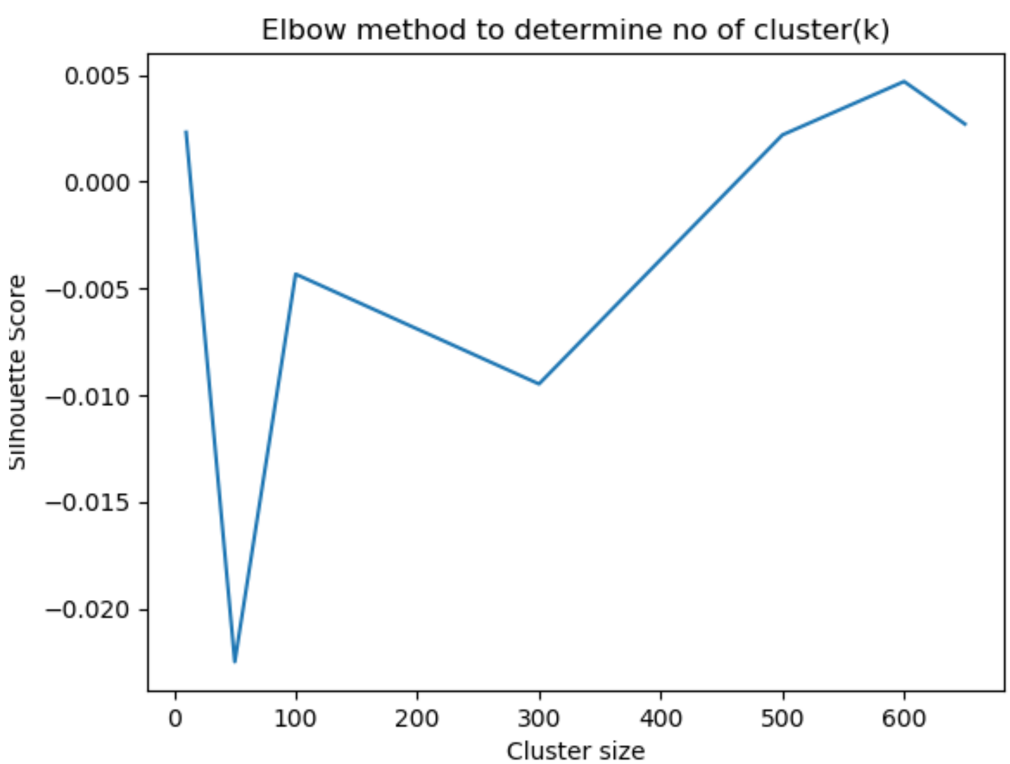
Key observations from the above experiment were:

- Spherical data run time < Ellipsoid data run time < Poorly Separated data
- Spherical data converges quickly, with clear separation
- As the means are well separated in first two cases, the clustering also is 100% efficient
- In the third case, since the means are close together and the data is poorly separated, most of the data points fall into one cluster. Poor clustering occurs
- Old Faithful data set also gets clustered better than K-Means model. In K-means model, 2 data points are misclassified. But using GMM, all the data points are correctly classified.

3. CLUSTERING OF TEXT USING K-MEANS

- Given data 11000 dimensions. They were clustered using different k – values where, k = number of clusters
- I ran for different cluster size: (Maximum cluster size can be 700 since 700 data points are only available
  - Cluster Size = [10,50,100,300,500,600,650]
- For each cluster size, two different scores were calculated
  - Silhouette Score
  - Sum of Squared Distance Score
- Cluster size, was chosen based on minimum score value
- The plots for both the score values for each cluster size is shown below
- Chose cluster size k = 100
- The document ids for each cluster for chosen cluster size is printed below


Elbow method to determine no of cluster(k)
sum of squared error (SSE)
Cluster size
0
20
40
60
80
0
100
200
300
400
500
600


Elbow method to determine no of cluster(k)
Silhouette Score
Cluster size
0.005
0.000
−0.005
−0.010
−0.015
−0.020
0
100
200
300
400
500
600

DOCUMENT IDs FOR CLUSTER SIZE = 100:

Cluster : 100
Document Id's belonging to Cluster: 0
Index([u'1992_18'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 1
Index([u'1987_33', u'1987_44', u'1987_68', u'1987_79', u'1988_19', u'1988_75',
       u'1989_53', u'1989_60', u'1989_82', u'1990_86', u'1990_112', u'1991_5',
       u'1991_40', u'1992_4', u'1992_19', u'1992_22', u'1992_45', u'1992_104',
       u'1992_106'],
      dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 2
Index([u'1987_27', u'1989_40', u'1989_84', u'1990_56', u'1990_61', u'1990_62',
       u'1990_63', u'1990_64', u'1990_65', u'1990_67', u'1990_68', u'1990_90',
       u'1991_64', u'1991_65', u'1991_69', u'1991_70', u'1992_33', u'1992_37',
       u'1992_38', u'1992_40', u'1992_41'],
      dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 3
Index([u'1987_4', u'1988_20', u'1988_23', u'1988_79', u'1990_143', u'1991_143',
       u'1992_9', u'1992_27', u'1992_39'],
      dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 4
Index([u'1988_26', u'1989_27'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 5
Index([u'1987_35'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 6
Index([u'1987_20', u'1987_26', u'1987_29', u'1987_71', u'1987_74', u'1987_75',
       u'1988_22', u'1988_46', u'1988_47', u'1988_55', u'1988_81', u'1989_7',
       u'1989_16', u'1989_17', u'1989_18', u'1990_3', u'1990_6', u'1990_70',
       u'1991_6', u'1991_7', u'1991_10', u'1991_93', u'1991_94', u'1991_100',
       u'1992_47', u'1992_51', u'1992_113', u'1992_124'],
      dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 7
Index([u'1987_7'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 8
Index([u'1988_38', u'1988_39', u'1989_49', u'1989_51', u'1990_76', u'1991_59',
       u'1991_60', u'1991_61'],
      dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 9
Index([u'1987_3', u'1987_14', u'1987_18', u'1987_24', u'1987_25', u'1987_30',

```
       u'1987_42', u'1987_52', u'1987_56', u'1987_62', u'1987_70', u'1987_77',
       u'1987_81', u'1987_85', u'1987_90', u'1988_53', u'1988_59', u'1988_65',
       u'1988_74', u'1989_9', u'1989_39', u'1989_44', u'1989_100', u'1990_9',
       u'1990_13', u'1990_14', u'1990_15', u'1990_24', u'1990_71', u'1990_87',
       u'1990_88', u'1990_118', u'1990_120', u'1991_13', u'1991_15',
       u'1991_27', u'1991_35', u'1991_37', u'1991_89', u'1991_90', u'1992_16',
       u'1992_50', u'1992_69', u'1992_70', u'1992_71', u'1992_83', u'1992_99'],
      dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  10
Index([u'1989_19'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  11
Index([u'1987_58', u'1988_10', u'1988_32', u'1988_57', u'1988_61', u'1989_55',
       u'1989_61', u'1989_69', u'1989_83', u'1989_101', u'1990_93', u'1990_94',
       u'1990_95', u'1990_98', u'1990_100', u'1990_101', u'1990_102',
       u'1990_110', u'1990_128', u'1990_129', u'1991_102', u'1991_105',
       u'1991_108', u'1991_114', u'1991_115', u'1991_120', u'1991_128',
       u'1991_130', u'1992_10', u'1992_60', u'1992_67', u'1992_73', u'1992_75',
       u'1992_77', u'1992_79', u'1992_98'],
      dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  12
Index([u'1987_55', u'1988_66'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  13
Index([u'1991_17', u'1992_86', u'1992_88'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  14
Index([u'1987_38', u'1987_53', u'1987_66', u'1988_17', u'1988_60',
       u'1990_123'],
      dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  15
Index([u'1989_52'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  16
Index([u'1987_60'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  17
Index([u'1989_47', u'1991_39'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  18
Index([u'1989_21', u'1990_132'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  19
Index([u'1988_49'], dtype='object', name=u'doc_id')
----------------------------------------
```

```
Document Id's belonging to Cluster:  20
Index([u'1988_6'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  21
Index([u'1992_80'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  22
Index([u'1987_76'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  23
Index([u'1987_45', u'1988_77', u'1988_80', u'1989_92', u'1990_54', u'1991_98',
      u'1992_102'],
     dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  24
Index([u'1988_27', u'1988_58', u'1988_94', u'1989_23', u'1989_25', u'1989_26',
      u'1990_27', u'1990_29', u'1990_30', u'1990_31', u'1990_32', u'1990_34',
      u'1990_35', u'1990_36', u'1991_21', u'1991_22', u'1991_25', u'1991_30',
      u'1992_81', u'1992_84', u'1992_87', u'1992_90'],
     dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  25
Index([u'1990_119', u'1991_122', u'1992_21'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  26
Index([u'1987_84'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  27
Index([u'1988_15', u'1991_4', u'1991_9', u'1992_122'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  28
Index([u'1987_39', u'1988_48', u'1989_36', u'1989_91', u'1990_5', u'1991_43',
      u'1991_47', u'1991_51', u'1991_73', u'1991_74', u'1991_75', u'1992_118',
      u'1992_125'],
     dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  29
Index([u'1990_25', u'1991_113'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  30
Index([u'1987_9', u'1987_10', u'1989_5', u'1989_14', u'1990_12', u'1990_38',
      u'1991_2', u'1991_66', u'1992_120'],
     dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  31
Index([u'1992_76'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  32
```

Index([u'1989_87', u'1990_16', u'1991_36', u'1991_41', u'1991_71', u'1992_23'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 33
Index([u'1989_88', u'1990_33', u'1990_108', u'1991_52', u'1991_57', u'1992_1',
      u'1992_49'],
     dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 34
Index([u'1989_33', u'1990_40', u'1991_54'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 35
Index([u'1992_34'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 36
Index([u'1989_10', u'1989_11'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 37
Index([u'1987_83', u'1991_32'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 38
Index([u'1987_69', u'1987_73', u'1990_46', u'1992_46', u'1992_52'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 39
Index([u'1987_37', u'1988_1', u'1988_4', u'1988_21', u'1989_42', u'1989_54',
      u'1989_57', u'1989_65', u'1989_67', u'1989_74', u'1989_79', u'1990_89',
      u'1990_91', u'1990_103', u'1990_115', u'1990_117', u'1990_121',
      u'1990_122', u'1990_125', u'1991_58', u'1991_109', u'1991_117',
      u'1991_123', u'1991_144', u'1992_11', u'1992_20', u'1992_26',
      u'1992_31', u'1992_42', u'1992_56', u'1992_57', u'1992_59', u'1992_61',
      u'1992_63', u'1992_65', u'1992_78'],
     dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 40
Index([u'1987_15', u'1987_47', u'1987_78', u'1988_24', u'1988_25', u'1989_22',
      u'1989_31', u'1989_98', u'1990_21', u'1991_19', u'1991_20', u'1992_85'],
     dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 41
Index([u'1992_6'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 42
Index([u'1989_75', u'1989_76', u'1989_77', u'1990_73', u'1990_111',
      u'1990_124', u'1991_45', u'1991_79', u'1991_81', u'1991_82', u'1991_84',
      u'1991_86', u'1991_87', u'1991_103', u'1991_104', u'1991_118',
      u'1991_129', u'1991_133', u'1991_134', u'1992_2', u'1992_29',
      u'1992_72', u'1992_74'],
     dtype='object', name=u'doc_id')

```
----------------------------------------
Document Id's belonging to Cluster:  43
Index([u'1992_89'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  44
Index([u'1987_12'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  45
Index([u'1991_62', u'1992_55'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  46
Index([u'1987_6', u'1987_23', u'1987_36', u'1987_65', u'1987_72', u'1987_89',
      u'1988_2', u'1988_3', u'1988_7', u'1988_12', u'1988_14', u'1988_16',
      u'1988_29', u'1988_30', u'1988_33', u'1988_40', u'1988_42', u'1988_63',
      u'1988_64', u'1988_70', u'1988_76', u'1988_85', u'1988_90', u'1989_30',
      u'1989_43', u'1989_59', u'1989_62', u'1989_66', u'1989_68', u'1989_73',
      u'1989_80', u'1990_23', u'1990_37', u'1990_39', u'1990_42', u'1990_44',
      u'1990_60', u'1990_72', u'1990_74', u'1990_75', u'1990_77', u'1990_78',
      u'1990_81', u'1990_82', u'1990_83', u'1990_85', u'1990_92', u'1990_96',
      u'1990_99', u'1990_104', u'1990_105', u'1990_107', u'1990_116',
      u'1990_126', u'1990_131', u'1990_139', u'1990_140', u'1990_142',
      u'1991_34', u'1991_38', u'1991_46', u'1991_55', u'1991_78', u'1991_80',
      u'1991_83', u'1991_85', u'1991_96', u'1991_99', u'1991_110',
      u'1991_121', u'1991_125', u'1991_127', u'1991_131', u'1991_132',
      u'1991_135', u'1991_136', u'1991_137', u'1991_138', u'1992_5',
      u'1992_7', u'1992_12', u'1992_14', u'1992_24', u'1992_25', u'1992_53',
      u'1992_62', u'1992_91', u'1992_93', u'1992_107', u'1992_117'],
     dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  47
Index([u'1987_5', u'1989_86', u'1990_130', u'1991_116', u'1992_3', u'1992_8'], dtype='object',
name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  48
Index([u'1987_57', u'1988_41', u'1990_59'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  49
Index([u'1991_23', u'1991_26'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  50
Index([u'1991_50'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  51
Index([u'1988_36', u'1990_58', u'1992_35'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  52
Index([u'1987_2', u'1987_8', u'1988_5', u'1988_87', u'1989_58', u'1990_109',
      u'1990_138', u'1991_107', u'1991_140', u'1992_36', u'1992_103'],
```

dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 53
Index([u'1992_30'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 54
Index([u'1990_28'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 55
Index([u'1992_114', u'1992_127'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 56
Index([u'1988_54', u'1990_43'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 57
Index([u'1987_64', u'1989_95', u'1991_97'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 58
Index([u'1989_34', u'1991_56'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 59
Index([u'1988_13', u'1989_64', u'1990_26'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 60
Index([u'1992_109'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 61
Index([u'1988_44', u'1992_121'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 62
Index([u'1990_79'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 63
Index([u'1991_88', u'1992_96'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 64
Index([u'1992_119'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 65
Index([u'1987_13'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 66
Index([u'1988_82', u'1988_84', u'1990_53'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 67
Index([u'1989_45'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 68

```
Index([u'1987_88'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  69
Index([u'1991_67'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  70
Index([u'1987_46', u'1987_63', u'1988_18', u'1988_89', u'1989_70', u'1990_97',
       u'1991_42', u'1991_68', u'1992_15', u'1992_17'],
      dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  71
Index([u'1992_44'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  72
Index([u'1989_13'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  73
Index([u'1991_29'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  74
Index([u'1991_112'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  75
Index([u'1987_41'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  76
Index([u'1989_50'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  77
Index([u'1989_12', u'1990_1', u'1991_12'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  78
Index([u'1989_29'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  79
Index([u'1987_50'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  80
Index([u'1987_1', u'1987_16', u'1987_17', u'1987_22', u'1987_28', u'1987_31',
       u'1987_32', u'1987_34', u'1987_43', u'1987_48',
       ...
       u'1992_68', u'1992_82', u'1992_92', u'1992_105', u'1992_108',
       u'1992_110', u'1992_111', u'1992_112', u'1992_116', u'1992_123'],
      dtype='object', name=u'doc_id', length=124)
----------------------------------------
Document Id's belonging to Cluster:  81
Index([u'1989_32'], dtype='object', name=u'doc_id')
----------------------------------------
```

Document Id's belonging to Cluster: 82
Index([u'1987_87'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 83
Index([u'1988_45', u'1989_3', u'1990_52', u'1992_100'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 84
Index([u'1992_115'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 85
Index([u'1991_119'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 86
Index([u'1988_9', u'1989_63', u'1989_90'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 87
Index([u'1989_24'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 88
Index([u'1987_21', u'1987_40', u'1987_54', u'1987_59', u'1988_78', u'1988_83',
    u'1988_86', u'1988_88', u'1989_93', u'1989_94', u'1989_96', u'1989_97',
    u'1990_135', u'1990_136', u'1990_137', u'1990_141', u'1991_91',
    u'1991_92', u'1991_95', u'1992_94', u'1992_95', u'1992_97',
    u'1992_101'],
   dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 89
Index([u'1991_63'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 90
Index([u'1987_11'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 91
Index([u'1987_51'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 92
Index([u'1988_43', u'1989_2', u'1989_4', u'1992_126'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 93
Index([u'1987_19', u'1988_68', u'1989_56'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 94
Index([u'1990_22', u'1991_18'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 95
Index([u'1991_31'], dtype='object', name=u'doc_id')

Document Id's belonging to Cluster: 96

```
Index([u'1991_28'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  97
Index([u'1987_67'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  98
Index([u'1988_28'], dtype='object', name=u'doc_id')
----------------------------------------
Document Id's belonging to Cluster:  99
Index([u'1990_20'], dtype='object', name=u'doc_id')
----------------------------------------
```

--- CODES FROM FOLLOWING PAGE ---

CODES:

PROBLEM 1:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans


X = np.loadtxt("old-faithful.txt")

plt.figure()
plt.scatter(X[:,1], X[:,2])
plt.title("Scatter plot of old-faithful data")
plt.xlabel("Duration of old faithful geyser eruptions")
plt.ylabel("waiting time between eruptions")

kmeans_model = KMeans(n_clusters=2)
kmeans_model.fit(X[:,1:])

plt.figure()
plt.scatter(X[:,1], X[:,2], c=kmeans_model.labels_, cmap="jet")
plt.title("Scatter plot of old-faithful data - Clustered")
plt.xlabel("Duration of old faithful geyser eruptions")
plt.ylabel("waiting time between eruptions")
```

PROBLEM 2:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
from sklearn.datasets.samples_generator import make_blobs
from numpy.core.umath_tests import inner1d
from sklearn import preprocessing
import time

class GMMClustering:

  def __init__(self, n_clusters, max_iterations, initializeMethod):
    self.name = "GMMClustering"
    self.n_clusters = n_clusters
```

```
        #self.n_features = n_features
        #self.n_data = n_data
        self.max_iterations = max_iterations
        self.initializeMethod = initializeMethod
        self.labels_ = []

    def fit(self, X):

        self.n_data, self.n_features  = X.shape

        if self.initializeMethod == "Random":
            self.mean_vectors = np.random.uniform(0,1, size=[self.n_clusters,self.n_features])
            self.covariance_matrices = np.random.uniform(0,1,
size=[self.n_features,self.n_features,self.n_clusters])
            self.alpha_vectors = np.array([0.5,0.5])

        self.log_likelihood_values = []
        for iteration in range(0,self.max_iterations):
            ### Expectation Step
            for i in range(0, self.n_clusters):
                temp_likelihood = inner1d(((X -
self.mean_vectors[i]).dot(np.linalg.inv(self.covariance_matrices[i]))), ((X -
self.mean_vectors[i])))
                likelihood = np.exp(-temp_likelihood/2.0)
                likelihood = likelihood / ((2* math.pi *
abs(np.linalg.det(self.covariance_matrices[i])))**0.5)
                e_step = self.alpha_vectors[i] * likelihood
                if i==0:
                    expectation = e_step
                else:
                    expectation = np.vstack([expectation,e_step])
            expectation = expectation /  np.sum(expectation, axis=0)

            ### Maximization step (Updation)
            for i in range(0, self.n_clusters):
                ### update mean vector
                mean_temp = np.multiply(X.T, expectation[i]).T
                self.mean_vectors[i,:] = np.sum(mean_temp, axis=0) / np.sum(expectation[i])
                ### update covariace matrix
                covariance_temp1 = np.einsum('ij,kj->jik',(X - self.mean_vectors[i]).T,(X -
self.mean_vectors[i]).T)
                covariance_temp2 = np.multiply(covariance_temp1.T, expectation[i]).T
                self.covariance_matrices[i,:,:] = np.sum(covariance_temp2, axis=0) /
np.sum(expectation[i])
```

```
            ### update alpha vector
            self.alpha_vectors[i] = np.sum(expectation[i]) /  self.n_data

        ### evaluate log likelihood
        for i in range(0, self.n_clusters):
            temp_likelihood2 = inner1d(((X -
self.mean_vectors[i]).dot(np.linalg.inv(self.covariance_matrices[i]))), ((X -
self.mean_vectors[i])))
            likelihood2 = np.exp(-temp_likelihood2/2.0)
            likelihood2 = likelihood2 / ((2* math.pi *
abs(np.linalg.det(self.covariance_matrices[i])))**0.5)
            e_step2 = self.alpha_vectors[i] * likelihood2
            if i==0:
                expectation2 = e_step2
            else:
                expectation2 = np.vstack([expectation2,e_step2])
        log_likelihood = np.sum(expectation2, axis=0)
        log_likelihood = np.sum(np.log(log_likelihood))
        self.log_likelihood_values.append(log_likelihood)
        if iteration > 0:
            if self.log_likelihood_values[iteration] == self.log_likelihood_values[iteration-1]:
                print("Converged at iteration ", iteration)
                break
        ### assign cluster values
        self.labels_ = np.argmax(expectation2,axis=0)



n_features=2
n_clusters = 2
n_data = 300

### Generate data Spherical structure
centers = [[-5, 0], [0, 1.5]]
X1, y1 = make_blobs(n_samples=n_data, centers=centers, random_state=40)
X1_scaled = preprocessing.scale(X1)

plt.figure()
plt.scatter(X1[:,0],X1[:,1])
plt.title('Generated Data - Spherical')

print "Spherical Data --- GMM Clustering --- "
start_time = time.time()
gmm_model1 = GMMClustering(n_clusters, 100, "Random")
```

```
gmm_model1.fit(X1_scaled)
plt.figure()
plt.scatter(X1[:,0], X1[:,1], c=gmm_model1.labels_, cmap='jet')
plt.title('Generated Data - Spherical - GMM clustered')
print "--- Sperical Data: Run Time ---", (time.time() - start_time)

### Generate data Ellipsoid structure
centers = [[-5, 0], [0, 1.5]]
X2, y2 = make_blobs(n_samples=n_data, centers=centers, random_state=40)
transformation = [[0.4, 0.2], [-0.4, 1.2]]
X2 = np.dot(X2, transformation)
X2_scaled = preprocessing.scale(X2)

plt.figure()
plt.scatter(X2[:,0],X2[:,1])
plt.title('Generated Data - Ellipsoid')

print "Ellipsoid Data --- GMM Clustering --- "
start_time = time.time()
gmm_model2 = GMMClustering(n_clusters, 100, "Random")
gmm_model2.fit(X2_scaled)
plt.figure()
plt.scatter(X2[:,0], X2[:,1], c=gmm_model2.labels_, cmap='jet')
plt.title('Generated Data - Ellipsoid - GMM clustered')
print "--- Ellipsoid Data: Run Time ---", (time.time() - start_time)

### Generate poorly seperated subpopulations
centers = [[-5, 0], [-4, 0]]
X3, y3 = make_blobs(n_samples=n_data, centers=centers, random_state=40)
X3_scaled = preprocessing.scale(X3)

plt.figure()
plt.scatter(X3[:,0],X3[:,1])
plt.title('Generated Data - Poorly separated data')

print "Poorly Separated Data --- GMM Clustering --- "
start_time = time.time()
gmm_model3 = GMMClustering(n_clusters, 1000, "Random")
gmm_model3.fit(X3_scaled)
plt.figure()
plt.scatter(X3[:,0], X3[:,1], c=gmm_model3.labels_, cmap='jet')
plt.title('Generated Data - Poorly separated data  - GMM clustered')
print "--- Poorly Seperated Data: Run Time ---", (time.time() - start_time)
```

```
### old-faithful data
X4 = np.loadtxt("old-faithful.txt")
X4_scaled = preprocessing.scale(X4)


plt.figure()
plt.scatter(X4[:,1], X4[:,2])
plt.title("Scatter plot of old-faithful data")
plt.xlabel("Duration of old faithful geyser eruptions")
plt.ylabel("waiting time between eruptions")

print "Old Faithful Data --- GMM Clustering --- "
start_time = time.time()
gmm_model4 = GMMClustering(n_clusters, 50, "Random")
gmm_model4.fit(X4_scaled[:,1:])
plt.figure()
plt.scatter(X4[:,1], X4[:,2], c=gmm_model4.labels_, cmap='jet')
plt.title("Scatter plot of old-faithful data - GMM clustered")
plt.xlabel("Duration of old faithful geyser eruptions")
plt.ylabel("waiting time between eruptions")
print "--- Old Faithful Data: Run Time ---", (time.time() - start_time)
```

PROBLEM 3:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from scipy.spatial.distance import cdist
from sklearn.metrics import silhouette_score

df = pd.read_csv("nips-87-92.csv")
del df["Unnamed: 0"]
df.index = df.doc_id
del df["doc_id"]

kmeans_model_dict = {}
sse_dict = {}
silhouette_score_dict = {}
cluster_size_dict = [10,50,100,300,500,600,650]
#cluster_size_dict = [100]

for k in sorted(cluster_size_dict):
    print "Cluster :",k
```

```
    kmeans_model = KMeans(n_clusters=k)
    kmeans_model.fit(df)
    kmeans_model_dict[k] = kmeans_model
    sse_dict[k] = sum(np.min(cdist(df, kmeans_model.cluster_centers_, 'euclidean'), axis=1)) /
df.shape[0]
    silhouette_score_dict[k] = silhouette_score(df, kmeans_model.labels_)

plt.figure()
plt.plot(*zip(*sorted(sse_dict.items())))
plt.xlabel("Cluster size")
plt.ylabel("sum of squared error (SSE)")
plt.title("Elbow method to determine no of cluster(k)")
plt.savefig('sse.png')

plt.figure()
plt.plot(*zip(*sorted(silhouette_score_dict.items())))
plt.xlabel("Cluster size")
plt.ylabel("Silhouette Score")
plt.title("Elbow method to determine no of cluster(k)")
plt.savefig('silh.png')


best_cluster_size = 100
df['Cluster-Labels'] = kmeans_model_dict[best_cluster_size].labels_

for i in range(0,best_cluster_size):
    print "Document Id's belonging to Cluster: ", i
    print df.loc[df['Cluster-Labels'] == i].index
    print "-------------------------------------------"
```