Problem 3

In [1]:
```
import numpy as np
from scipy.special import logsumexp
from scipy.stats import multivariate_normal
from random import randint
import pandas as pd
```

In [2]:
```
def gaussian_data():
    with open("2gaussian.txt", 'r') as f:
        lines = f.readlines()
    data = []
    for line in lines:
        x, y = line.strip().split()
        data.append([float(x), float(y)])
    return np.array(data)
```

In [3]:
```
data = gaussian_data()
print(data)
```

```
[[7.57104365 3.53027417]
 [7.33721752 4.26271316]
 [3.07182783 1.11801871]
 ...
 [5.61639331 3.77793239]
 [8.59215378 3.6349037 ]
 [3.02221288 3.78337346]]
```

In [4]:
```
def gaussian(X, MU, Covariance) -> np.array:
    n = X.shape[1]
    difference = (X - MU).T

    base = 1 / ((2 * np.pi) ** (n / 2) * np.linalg.det(Covariance) ** 0.5)
    exponent_value = -0.5 * np.dot(np.dot(difference.T, np.linalg.inv(Covariance)), difference)
    exponent = np.exp(exponent_value)

    return np.diagonal( base *  exponent).reshape(-1, 1)
```

In [5]:
```
def initialize_clusters(X, k) -> np.array:

    PI = [ 1/k for i in range(0,k) ]
    MU = [ X[randint(0,len(X)-1),:] for i in range(0,k) ]
    Covariance = [ [ np.identity(X.shape[1] ,dtype = np.float64) ] for i in range(0,k) ]

    clusters = []
    for i in range(k):
        cluster = {}
        cluster['PI'] = PI[i]
        cluster['MU'] = MU[i]
        cluster['Covariance'] = Covariance[i]
        clusters.append(cluster)

    return clusters
```

In [6]:
```
def E_step(X, clusters) -> dict:
    expectation = np.zeros((X.shape[0], 1), dtype = np.float64)

    for cluster in clusters:
        PI = cluster['PI']
        MU = cluster['MU']
        Covariance = cluster['Covariance']

        weight = (PI * gaussian(X, MU, Covariance)).astype(np.float64)

        for i in range(X.shape[0]):
            expectation[i] += weight[i]

        cluster['weight'] = weight
        cluster['expectation'] = expectation

    for cluster in clusters:
        cluster['weight'] /= cluster['expectation']

    return cluster
```

In [7]:
```
def M_step(X, clusters) -> dict:
    X_len = float(X.shape[0])

    for cluster in clusters:
        weight = cluster['weight']
        Covariance = np.zeros((X.shape[1], X.shape[1]))
        sum_weights = np.sum(weight, axis=0)
        PI = sum_weights / X_len
        MU = np.sum(weight * X, axis=0) / sum_weights

        for i in range(X.shape[0]):
            difference = (X[i] - MU).reshape(-1, 1)
            Covariance += weight[i] * np.dot(difference, difference.T)
        Covariance = Covariance / sum_weights

        cluster['PI'] = PI
        cluster['MU'] = MU
        cluster['Covariance'] = Covariance

    return clusters
```

In [8]:
```
def get_likelihood(X, clusters) -> list:

    likelihoods = np.log(np.array([cluster['expectation'] for cluster in clusters]))
    sum_log_likelihood = np.sum(likelihoods)

    return [sum_log_likelihood, likelihoods]
```

In [9]:
```
k = 2
cycles = 4000
X_partition = []
X = data

clusters = initialize_clusters(X, k = 2)
likelihoods = np.zeros((cycles, ))

updated_likelihood = 0
for i in range(cycles):
  E_step(X, clusters)
  M_step(X, clusters)

  result = get_likelihood(X, clusters)
  likelihood, sample_likelihoods = result[0], result[1]

  if likelihood == updated_likelihood: break
  else:
    updated_likelihood = likelihood
    print('Cycle: ', i + 1, '  |  Likelihood: ', likelihood)


n = 0
clusters
for cluster in clusters:
  n += 1
  print('\nCluster  :  ', n )
  PI =  cluster['PI']
  MU =  cluster['MU']
  Covariance = cluster['Covariance']

  print('PI : ', PI)
  print('Mean : ', MU)
  print('Covariance Matrix : \n', np.array(Covariance))
```

```
Cycle:  1   |  Likelihood:  -21759.283746140267
Cycle:  2   |  Likelihood:  -45391.91427481979
Cycle:  3   |  Likelihood:  -44986.21124153352
Cycle:  4   |  Likelihood:  -44640.75026177596
Cycle:  5   |  Likelihood:  -44361.739777584706
Cycle:  6   |  Likelihood:  -44079.794823029406
Cycle:  7   |  Likelihood:  -43705.95194052593
Cycle:  8   |  Likelihood:  -43221.08633291509
Cycle:  9   |  Likelihood:  -42789.717951574545
Cycle:  10   |  Likelihood:  -42531.28534627899
Cycle:  11   |  Likelihood:  -42373.647344868565
Cycle:  12   |  Likelihood:  -42262.504019602355
Cycle:  13   |  Likelihood:  -42180.60659853055
Cycle:  14   |  Likelihood:  -42122.50013982032
Cycle:  15   |  Likelihood:  -42084.62906798239
Cycle:  16   |  Likelihood:  -42062.12165213494
Cycle:  17   |  Likelihood:  -42049.75389655398
Cycle:  18   |  Likelihood:  -42043.34857348301
Cycle:  19   |  Likelihood:  -42040.16694358488
Cycle:  20   |  Likelihood:  -42038.6307812887
Cycle:  21   |  Likelihood:  -42037.9030107265
Cycle:  22   |  Likelihood:  -42037.56254138199
Cycle:  23   |  Likelihood:  -42037.40459251016
Cycle:  24   |  Likelihood:  -42037.331727191086
Cycle:  25   |  Likelihood:  -42037.29823874928
Cycle:  26   |  Likelihood:  -42037.2828864106
Cycle:  27   |  Likelihood:  -42037.27586025005
Cycle:  28   |  Likelihood:  -42037.27264832117
Cycle:  29   |  Likelihood:  -42037.27118115374
Cycle:  30   |  Likelihood:  -42037.270511318304
Cycle:  31   |  Likelihood:  -42037.270205611916
Cycle:  32   |  Likelihood:  -42037.270066123434
Cycle:  33   |  Likelihood:  -42037.27000248743
Cycle:  34   |  Likelihood:  -42037.269973459195
Cycle:  35   |  Likelihood:  -42037.26996021862
Cycle:  36   |  Likelihood:  -42037.26995417953
Cycle:  37   |  Likelihood:  -42037.26995142516
Cycle:  38   |  Likelihood:  -42037.26995016895
Cycle:  39   |  Likelihood:  -42037.26994959602
Cycle:  40   |  Likelihood:  -42037.269949334725
Cycle:  41   |  Likelihood:  -42037.26994921556
Cycle:  42   |  Likelihood:  -42037.269949161215
Cycle:  43   |  Likelihood:  -42037.26994913643
Cycle:  44   |  Likelihood:  -42037.269949125126
Cycle:  45   |  Likelihood:  -42037.26994911997
Cycle:  46   |  Likelihood:  -42037.269949117624
Cycle:  47   |  Likelihood:  -42037.26994911655
Cycle:  48   |  Likelihood:  -42037.26994911605
Cycle:  49   |  Likelihood:  -42037.269949115835
Cycle:  50   |  Likelihood:  -42037.26994911574
Cycle:  51   |  Likelihood:  -42037.26994911568
Cycle:  52   |  Likelihood:  -42037.26994911567
Cycle:  53   |  Likelihood:  -42037.26994911566
Cycle:  54   |  Likelihood:  -42037.26994911565

Cluster  :   1
PI :  [0.66520423]
Mean :  [7.01314832 3.98313419]
Covariance Matrix :
 [[0.97475892 0.4974703 ]
 [0.4974703  1.00114259]]

Cluster  :   2
PI :  [0.33479577]
Mean :  [2.99413183 3.0520966 ]
Covariance Matrix :
 [[1.01023427 0.02719139]
 [0.02719139 2.93782296]]
```

In [10]:

```
n1 = n2 = 0

for i in range(X.shape[0]):
    if clusters[0]['weight'][i][0] >= clusters[1]['weight'][i][0]:
        n1 += 1
    else:
        n2 += 1

print("Number of data points in Cluster 1:", n1)
print("Number of data points in Cluster 2:", n2)
```

```
Number of data points in Cluster 1: 4009
Number of data points in Cluster 2: 1991
```

In [11]:
```
def gaussian_data():
    with open("3gaussian.txt", 'r') as f:
        lines = f.readlines()
    data = []
    for line in lines:
        x, y = line.strip().split()
        data.append([float(x), float(y)])
    return np.array(data)
```

In [12]:
```
data = gaussian_data()
print(data)
```

```
[[2.94693347 3.16222499]
 [5.98399602 4.84671738]
 [5.30142995 8.16811309]
 ...
 [6.27055168 2.83700248]
 [5.27935185 7.87197636]
 [7.26196796 4.58568396]]
```

In [13]:
```
k = 3
cycles = 4000
X_partition = []
X = data

clusters = initialize_clusters(X, k = 3)
likelihoods = np.zeros((cycles, ))

updated_likelihood = 0
for i in range(cycles):
  E_step(X, clusters)
  M_step(X, clusters)

  result = get_likelihood(X, clusters)
  likelihood, sample_likelihoods = result[0], result[1]

  if likelihood == updated_likelihood: break
  else:
    updated_likelihood = likelihood
    print('Cycle: ', i + 1, '  |  Likelihood: ', likelihood)

n = 0
clusters
for cluster in clusters:
  n += 1
  print('\nCluster  :  ', n )
  PI =  cluster['PI']
  MU =  cluster['MU']
  Covariance = cluster['Covariance']

  print('PI : ', PI)
  print('Mean : ', MU)
  print('Covariance Matrix : \n', np.array(Covariance))
```

```
Cycle:  1   |  Likelihood:  -125169.26396770614
Cycle:  2   |  Likelihood:  -117363.83200370363
Cycle:  3   |  Likelihood:  -116336.36189120801
Cycle:  4   |  Likelihood:  -115839.98478240208
Cycle:  5   |  Likelihood:  -115510.99791576609
Cycle:  6   |  Likelihood:  -115284.41768730324
Cycle:  7   |  Likelihood:  -115124.56867900268
Cycle:  8   |  Likelihood:  -115007.55604407785
Cycle:  9   |  Likelihood:  -114918.5030418912
Cycle:  10   |  Likelihood:  -114848.45150494791
Cycle:  11   |  Likelihood:  -114791.80806195675
Cycle:  12   |  Likelihood:  -114744.83680019964
Cycle:  13   |  Likelihood:  -114704.90419575069
Cycle:  14   |  Likelihood:  -114670.10720811533
Cycle:  15   |  Likelihood:  -114639.06936005372
Cycle:  16   |  Likelihood:  -114610.81226993594
Cycle:  17   |  Likelihood:  -114584.66712532956
Cycle:  18   |  Likelihood:  -114560.20923136428
Cycle:  19   |  Likelihood:  -114537.20339013917
Cycle:  20   |  Likelihood:  -114515.5516953353
Cycle:  21   |  Likelihood:  -114495.24212379252
Cycle:  22   |  Likelihood:  -114476.30225769812
Cycle:  23   |  Likelihood:  -114458.76324385204
Cycle:  24   |  Likelihood:  -114442.63603403815
Cycle:  25   |  Likelihood:  -114427.89904286804
Cycle:  26   |  Likelihood:  -114414.49518481645
Cycle:  27   |  Likelihood:  -114402.33615576813
Cycle:  28   |  Likelihood:  -114391.31198657924
Cycle:  29   |  Likelihood:  -114381.3039441348
Cycle:  30   |  Likelihood:  -114372.19859110659
Cycle:  31   |  Likelihood:  -114363.90018000347
Cycle:  32   |  Likelihood:  -114356.338069299
Cycle:  33   |  Likelihood:  -114349.46673717414
Cycle:  34   |  Likelihood:  -114343.25885009079
Cycle:  35   |  Likelihood:  -114337.69505166504
Cycle:  36   |  Likelihood:  -114332.75496473334
Cycle:  37   |  Likelihood:  -114328.4119517584
Cycle:  38   |  Likelihood:  -114324.63163281936
Cycle:  39   |  Likelihood:  -114321.37284295177
Cycle:  40   |  Likelihood:  -114318.58967395505
Cycle:  41   |  Likelihood:  -114316.23372648103
Cycle:  42   |  Likelihood:  -114314.25614940186
Cycle:  43   |  Likelihood:  -114312.60931154268
Cycle:  44   |  Likelihood:  -114311.24807470913
Cycle:  45   |  Likelihood:  -114310.130686447
Cycle:  46   |  Likelihood:  -114309.21933145834
Cycle:  47   |  Likelihood:  -114308.4803913475
Cycle:  48   |  Likelihood:  -114307.88446843742
Cycle:  49   |  Likelihood:  -114307.40623104674
Cycle:  50   |  Likelihood:  -114307.024134804
Cycle:  51   |  Likelihood:  -114306.72006812986
Cycle:  52   |  Likelihood:  -114306.47896137554
Cycle:  53   |  Likelihood:  -114306.28838976455
Cycle:  54   |  Likelihood:  -114306.13819142291
Cycle:  55   |  Likelihood:  -114306.02011413372
Cycle:  56   |  Likelihood:  -114305.92749833548
Cycle:  57   |  Likelihood:  -114305.85499931942
Cycle:  58   |  Likelihood:  -114305.79834840343
Cycle:  59   |  Likelihood:  -114305.75415082586
Cycle:  60   |  Likelihood:  -114305.7197169424
Cycle:  61   |  Likelihood:  -114305.69292278407
Cycle:  62   |  Likelihood:  -114305.67209593893
Cycle:  63   |  Likelihood:  -114305.65592289779
Cycle:  64   |  Likelihood:  -114305.64337433584
Cycle:  65   |  Likelihood:  -114305.63364520977
Cycle:  66   |  Likelihood:  -114305.62610697266
Cycle:  67   |  Likelihood:  -114305.62026962145
Cycle:  68   |  Likelihood:  -114305.61575166642
Cycle:  69   |  Likelihood:  -114305.61225644684
Cycle:  70   |  Likelihood:  -114305.60955350426
Cycle:  71   |  Likelihood:  -114305.60746397005
Cycle:  72   |  Likelihood:  -114305.6058491262
Cycle:  73   |  Likelihood:  -114305.60460146723
Cycle:  74   |  Likelihood:  -114305.603637728
Cycle:  75   |  Likelihood:  -114305.60289345236
Cycle:  76   |  Likelihood:  -114305.60231876782
Cycle:  77   |  Likelihood:  -114305.60187510174
Cycle:  78   |  Likelihood:  -114305.60153263198
Cycle:  79   |  Likelihood:  -114305.60126830894
Cycle:  80   |  Likelihood:  -114305.60106432265
Cycle:  81   |  Likelihood:  -114305.60090691503
Cycle:  82   |  Likelihood:  -114305.60078546028
Cycle:  83   |  Likelihood:  -114305.60069175341
Cycle:  84   |  Likelihood:  -114305.6006194597
Cycle:  85   |  Likelihood:  -114305.6005636891
Cycle:  86   |  Likelihood:  -114305.60052066734
```

```
Cycle:  87   |  Likelihood:  -114305.60048748151
Cycle:  88   |  Likelihood:  -114305.60046188386
Cycle:  89   |  Likelihood:  -114305.60044213993
Cycle:  90   |  Likelihood:  -114305.60042691158
Cycle:  91   |  Likelihood:  -114305.6004151663
Cycle:  92   |  Likelihood:  -114305.6004061077
Cycle:  93   |  Likelihood:  -114305.60039912131
Cycle:  94   |  Likelihood:  -114305.60039373321
Cycle:  95   |  Likelihood:  -114305.60038957783
Cycle:  96   |  Likelihood:  -114305.60038637317
Cycle:  97   |  Likelihood:  -114305.60038390175
Cycle:  98   |  Likelihood:  -114305.60038199581
Cycle:  99   |  Likelihood:  -114305.60038052601
Cycle:  100   |  Likelihood:  -114305.6003793925
Cycle:  101   |  Likelihood:  -114305.60037851839
Cycle:  102   |  Likelihood:  -114305.60037784431
Cycle:  103   |  Likelihood:  -114305.60037732449
Cycle:  104   |  Likelihood:  -114305.60037692363
Cycle:  105   |  Likelihood:  -114305.60037661449
Cycle:  106   |  Likelihood:  -114305.6003763761
Cycle:  107   |  Likelihood:  -114305.60037619228
Cycle:  108   |  Likelihood:  -114305.6003760505
Cycle:  109   |  Likelihood:  -114305.60037594117
Cycle:  110   |  Likelihood:  -114305.60037585687
Cycle:  111   |  Likelihood:  -114305.60037579187
Cycle:  112   |  Likelihood:  -114305.60037574175
Cycle:  113   |  Likelihood:  -114305.60037570307
Cycle:  114   |  Likelihood:  -114305.60037567327
Cycle:  115   |  Likelihood:  -114305.60037565028
Cycle:  116   |  Likelihood:  -114305.60037563257
Cycle:  117   |  Likelihood:  -114305.60037561889
Cycle:  118   |  Likelihood:  -114305.60037560835
Cycle:  119   |  Likelihood:  -114305.60037560022
Cycle:  120   |  Likelihood:  -114305.60037559396
Cycle:  121   |  Likelihood:  -114305.60037558911
Cycle:  122   |  Likelihood:  -114305.60037558539
Cycle:  123   |  Likelihood:  -114305.60037558252
Cycle:  124   |  Likelihood:  -114305.60037558031
Cycle:  125   |  Likelihood:  -114305.6003755786
Cycle:  126   |  Likelihood:  -114305.60037557727
Cycle:  127   |  Likelihood:  -114305.60037557626
Cycle:  128   |  Likelihood:  -114305.60037557546
Cycle:  129   |  Likelihood:  -114305.60037557487
Cycle:  130   |  Likelihood:  -114305.6003755744
Cycle:  131   |  Likelihood:  -114305.60037557405
Cycle:  132   |  Likelihood:  -114305.60037557378
Cycle:  133   |  Likelihood:  -114305.60037557354
Cycle:  134   |  Likelihood:  -114305.60037557338
Cycle:  135   |  Likelihood:  -114305.60037557327
Cycle:  136   |  Likelihood:  -114305.60037557317
Cycle:  137   |  Likelihood:  -114305.60037557309
Cycle:  138   |  Likelihood:  -114305.60037557302
Cycle:  139   |  Likelihood:  -114305.60037557298
Cycle:  140   |  Likelihood:  -114305.60037557295
Cycle:  141   |  Likelihood:  -114305.60037557292
Cycle:  142   |  Likelihood:  -114305.60037557289
Cycle:  143   |  Likelihood:  -114305.6003755729
Cycle:  144   |  Likelihood:  -114305.60037557289
Cycle:  145   |  Likelihood:  -114305.60037557286

Cluster  :   1
PI :  [0.29843661]
Mean :  [7.02156142 4.01546065]
Covariance Matrix :
 [[0.99041327 0.50095954]
 [0.50095954 0.99564873]]

Cluster  :   2
PI :  [0.49596835]
Mean :  [5.0117217  7.00146622]
Covariance Matrix :
 [[0.97972162 0.18516295]
 [0.18516295 0.97455232]]

Cluster  :   3
PI :  [0.20559504]
Mean :  [3.03968827 3.04847409]
Covariance Matrix :
 [[1.02849913 0.02681589]
 [0.02681589 3.38466417]]
```

In [21]:
```
n1 = n2 = n3 = 0

for i in range(X.shape[0]):
    if clusters[0]['weight'][i][0] >= clusters[1]['weight'][i][0] and clusters[0]['weight'][i][0] >= clusters[2]['weight'][i][0]:
        n2 += 1
    elif clusters[1]['weight'][i][0] >= clusters[0]['weight'][i][0] and clusters[1]['weight'][i][0] >= clusters[2]['weight'][i][0]
        n3 += 1
    else:
        n1 += 1

print("Number of data points in Cluster 1:", n1)
print("Number of data points in Cluster 2:", n2)
print("Number of data points in Cluster 3:", n3)
```

```
Number of data points in Cluster 1: 1964
Number of data points in Cluster 2: 3004
Number of data points in Cluster 3: 5032
```

In [ ]: