Results of EM algorithm on GMM

LIBIN N GEORGE

111501015

Clustering Performance using Rand index

For 2 dimensional data

1. K = 2
    1. N=100

        Rand Index = 0.8362
    2. N=1000

        Rand Index = 0.791848
    3. N=10000

        Rand Index = 0.79706632
2. K = 5
    1. N=100

        Rand Index = 0.8076
    2. N=1000

        Rand Index = 0.864072
    3. N=10000

        Rand Index = 0.9137696
3. K = 10
    1. N=100

        Rand Index = 0.891
    2. N=1000

        Rand Index = 0.886802
    3. N=10000

        Rand Index = 0.89645128

For 5 dimensional data

4. K = 2
    1. N=100

        Rand Index = 1.0
    2. N=1000

        Rand Index = 1.0
    3. N=10000

        Rand Index = 1.0
5. K = 5
    1. N=100

        Rand Index = 0.932
    2. N=1000

        Rand Index = 0.93395
    3. N=10000

        Rand Index = 0.79515272
6. K = 10
    1. N=100

        Rand Index = 0.8978
    2. N=1000

        Rand Index = 0.90309
    3. N=10000

        Rand Index = 0.95433612

For 10 dimensional data

7. K = 2
    1. N=100

        Rand Index = 0.52 (error due to small or large value (resulting inf in python))
    2. N=1000

        Rand Index = 1.0
    3. N=10000

Rand Index = 1.0

8. K = 5
    1. N=100

       Rand Index = 0.802
    2. N=1000

       Rand Index = 0.899348
    3. N=10000

       Rand Index = 0.9315452
9. K = 10
    1. N=100

       Rand Index = 0.9062
    2. N=1000

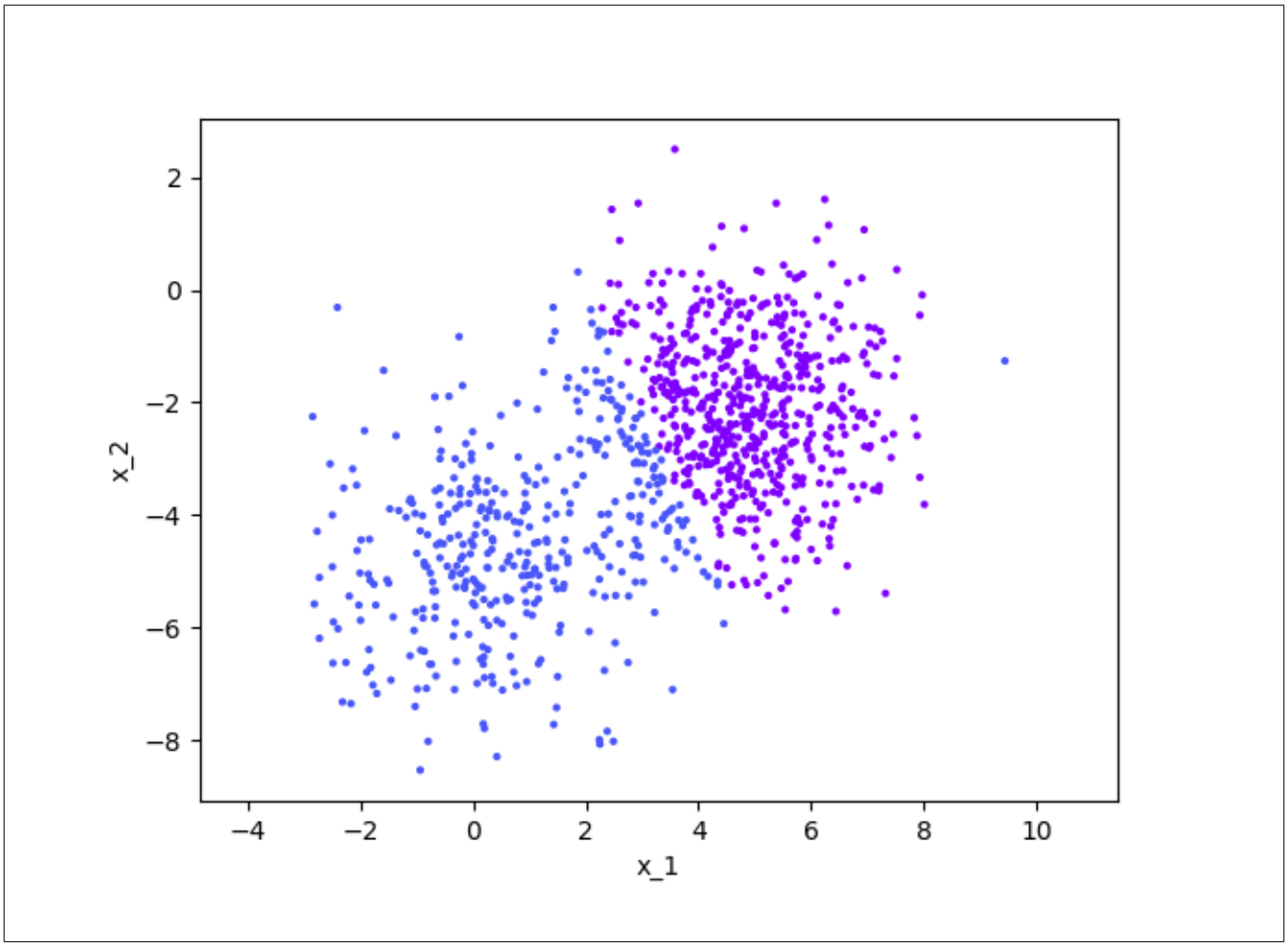
       Rand Index = 0.932594
    3. N=10000

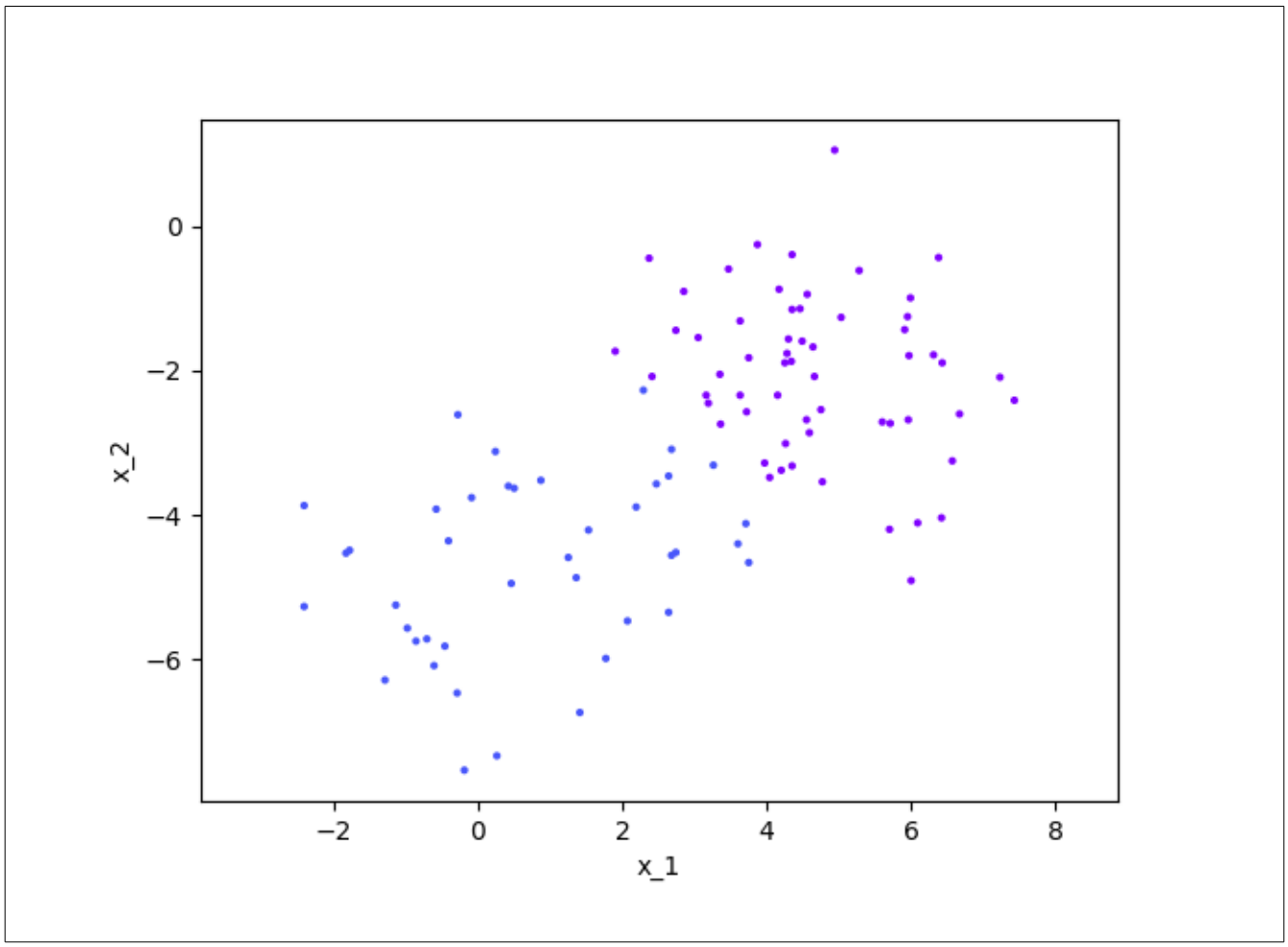
       Rand Index = 0.94650118

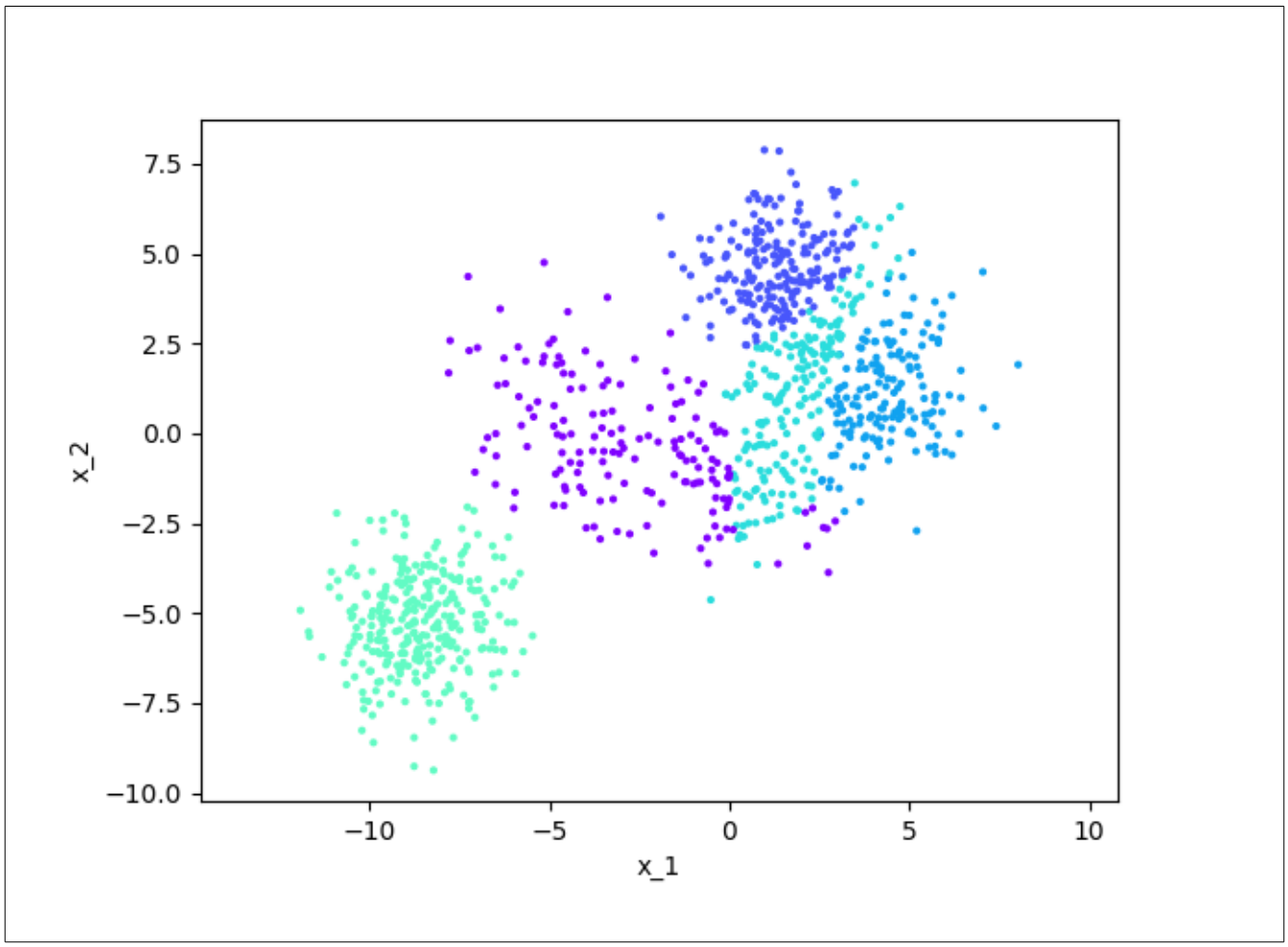
Observation

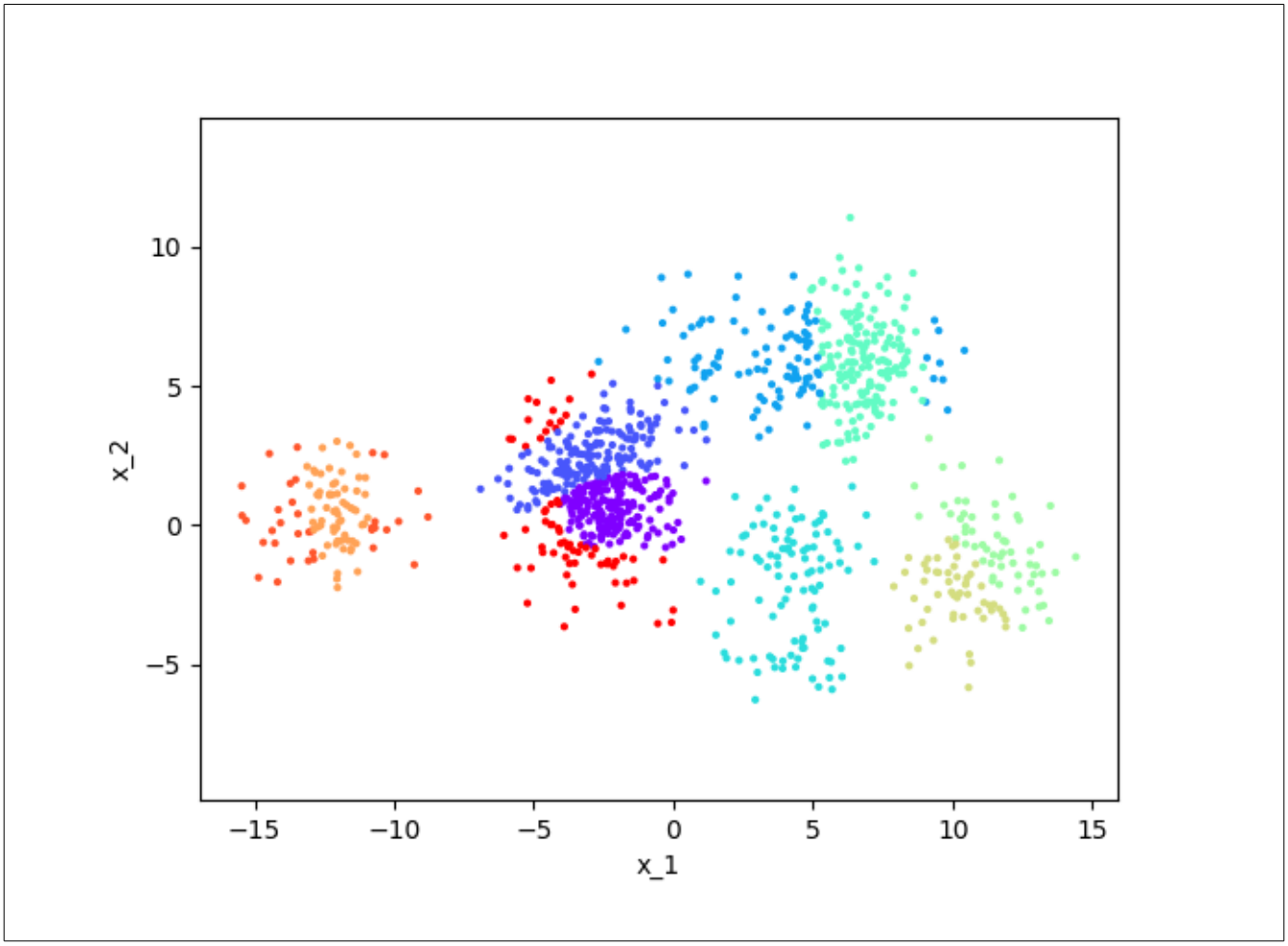
The Rand Index gets better when number of points increases and also when dimensions increase. As number of clusters increases the error also increases due to overlapping Gaussian distributions

Some Results plotted for 2 dimension




0
−2
−4
−6
x_2
−2
0
2
4
6
8
x_1


7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
x_2
−10
−5
0
5
10
x_1


10
5
0
−5
x_2
−15
−10
−5
0
5
10
15
x_1