CS 509 - Pattern recognition

Assignment 3

Problem: Programming exercise for EM algorithm

	ω_1			ω_2			ω_3		
point	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	0.42	-0.087	0.58	-0.4	0.58	0.089	0.83	1.6	-0.014
2	-0.2	-3.3	-3.4	-0.31	0.27	-0.04	1.1	1.6	0.48
3	1.3	-0.32	1.7	0.38	0.055	-0.035	-0.44	-0.41	0.32
4	0.39	0.71	0.23	-0.15	0.53	0.011	0.047	-0.45	1.4
5	-1.6	-5.3	-0.15	-0.35	0.47	0.034	0.28	0.35	3.1
6	-0.029	0.89	-4.7	0.17	0.69	0.1	-0.39	-0.48	0.11
7	-0.23	1.9	2.2	-0.011	0.55	-0.18	0.34	-0.079	0.14
8	0.27	-0.3	-0.87	-0.27	0.61	0.12	-0.3	-0.22	2.2
9	-1.9	0.76	-2.1	-0.065	0.49	0.0012	1.1	1.2	-0.46
10	0.87	-1.0	-2.6	-0.12	0.054	-0.063	0.18	-0.11	-0.49

Suppose we know that the ten data points in category ω_1 in the table above come from a three-dimensional Gaussian. Suppose, however, that we do not have access to the x_3 components for the even-numbered data points.

1. Write an EM program in Python to estimate the mean and covariance of the distribution. Start your estimate with $\mu_0 = 0$ and $\Sigma_0 = I$, the three-dimensional identity matrix. Display the obtained result in form of clusters.

   (**Hint:** for the missing x_3, you can simply attribute the value zero or $\frac{x_1+x_2}{2}$. In addition, I provided you the implementation of GMM)
2. Compare your final estimate with that for the case when there is no missing data. Display the obtained result in form of clusters.
3. Compare your final estimate with that for the case when there is no missing data in the case of diagonal covariance. Display the obtained result in form of clusters.
4. Akaike information criterion (AIC) and BIC Bayesian information criterion are both are used to choose the best k number of components to use in GMM. We choose the lowest values of AIC and BIC as the best number of components to use in GMM. In Figure 1, best number of components is 3. In general, their lowest values do not coincide, in this case we can choose the minimum of their lowest values, the maximum of their lowest values, or the average of their lowest values.

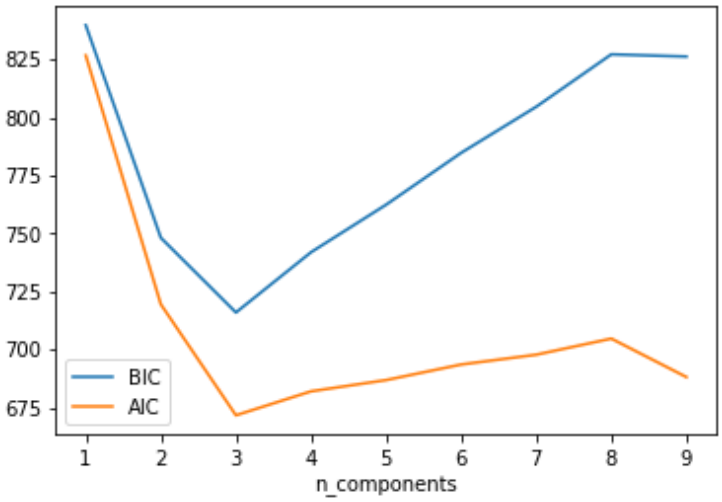

Figure 1

a. Do a search on internet for the definition of each one of them and explain their principles.

b. Implement AIC and BIC in Python.

   (**Hint:** I provided you the implementation in Python of AIC and BIC, you can find it in the method *testGMMsklearnBICAIC* in the file *estimate_gmm_sklearn.py*)

c. Draw the AIC and BIC curves for your dataset.

   (**Hint:** you can find how to draw their curves in the method *testGMMsklearnBICAIC*)

d. What is the best number of components to use in your GMM for your dataset? Justify your answer.

Submission

Please submit one pdf file and the Python files. ***Please do not submit your assignment in .zip nor in .rar***.