Question 2 : Estimation equations of EM

The log-likelihood of the model with parameters $\theta = (\mu_i, \Sigma_i, A, \pi)$ after the k^{th} E-step of the EM algorithm is (Calculation details in the course notes) :

$$l(\theta) = \sum_{i=1}^{K} \gamma_{1,i} \log(\pi_i) + \sum_{t=1}^{T-1} \sum_{i,j=1}^{K} \xi_{i,j}^{(t)} \log(A_{i,j}) + \sum_{t=1}^{T} \sum_{i=1}^{K} \gamma_{t,i} \log(\mathcal{N}(u_t; \mu_i, \Sigma_i))$$

$$\text{With} \quad \gamma_{t,i} = p(q_t = i|\bar{u}; \theta^{k-1}) \qquad \text{and} \quad \xi_{i,j}^{(t)} = p(q_{t+1} = i, q_t = j|\bar{u}; \theta^{k-1})$$

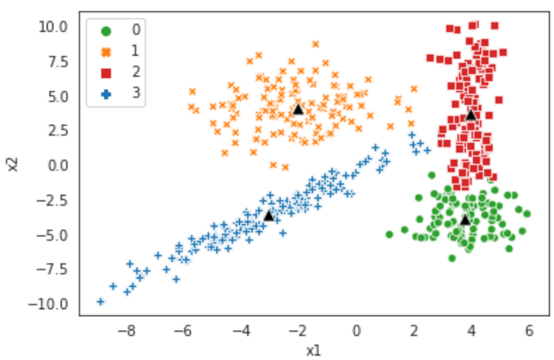
For the M-step, the Lagrangian of the problem is written in the form :

$$\mathscr{L}(\theta, \lambda, \delta) = \sum_{i=1}^{K} \gamma_{1,i} \log(\pi_i) + \sum_{t=1}^{T-1} \sum_{i,j=1}^{K} \xi_{i,j}^{(t)} \log(A_{i,j}) + \sum_{t=1}^{T} \sum_{i=1}^{K} \gamma_{t,i} \log(\mathcal{N}(u_t; \mu_i, \Sigma_i)) + \lambda(1 - \sum_{i=1}^{K} \pi_i) + \sum_{i=1}^{K} \delta_i (1 - \sum_{j=1}^{K} A_{i,j})$$

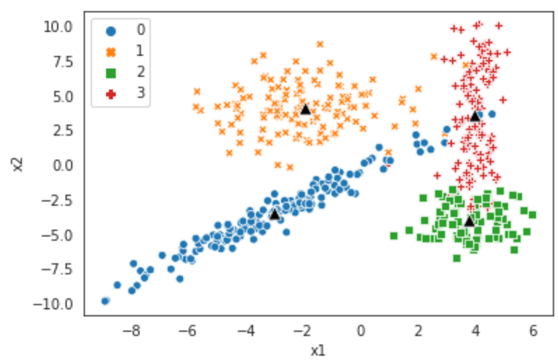
The log-likelihood is a strictly concave function *wrt* to the parameters (separately), in addition, it is clear that *Slater's constraint qualification* are verified, so the problem has strong duality property. Therefore, by setting the derivative equal to zero (method similar to the other HWs), we find :

$$\pi_i = \gamma_{1,i}, \quad A_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_{i,j}^{(t)}}{\sum_{i'=1}^{K} \sum_{t=1}^{T-1} \xi_{i',j}^{(t)}}, \quad \mu_i = \frac{\sum_{t=1}^{T} \gamma_{t,i} u_t}{\sum_{t=1}^{T} \gamma_{t,i}}, \quad \Sigma_i = \frac{\sum_{t=1}^{T} \gamma_{t,i} (u_t - \mu_i)(u_t - \mu_i)^T}{\sum_{t=1}^{T} \gamma_{t,i}}$$

Question 4 : Plots (HMM)



"EM - GMM"



"EM - HMM"

Question 5 : Comments

Log-likelihood

Inference Type	Train	Test
HMM	-1898.79	-1916.40
GMM	-2327.71	-2408.97

– It can be seen that HMM gives better log-likelihood than GMM on the train set and on the test set.
– This comparison makes no sense, because we have made different assumptions about the distribution (i.i.d for GMM and with temporal structure in HMM), and because we can maximize the log-likelihood just by making the model more complex (increase the number of states K).
– HMM's initialization with GMM allows it to converge quickly (20 iterations vs $\sim$ 150 for GMM), but even if we initialize randomly, it will converge faster than GMM.