Exercise 1.2

The complete log-likelihood of this HMM model writes as:

$$
\begin{aligned}
\ell_c(\theta) &= \log\left(p(q_0)\prod_{t=0}^{T-1} p(q_{t+1}|q_t)\prod_{t=0}^{T} p(u_t|q_t)\right) \\
&= \sum_{i=1}^{K}\delta(q_0=i)\log\pi_i + \sum_{t=0}^{T-1}\sum_{i,j}^{K}\delta(q_{t+1}=i,q_t=j)\log A_{i,j} + \sum_{t=0}^{T}\sum_{i=1}^{K}\delta(q_t=i)\log\left(\mathcal{N}(u_i,\mu_i,\Sigma_i)\right) \\
&= \sum_{i=1}^{K}\delta(q_0=i)\log\pi_i + \sum_{t=0}^{T-1}\sum_{i,j}^{K}\delta(q_{t+1}=i,q_t=j)\log A_{i,j} \\
&\quad -\frac{1}{2}\sum_{t=0}^{T}\sum_{i=1}^{K}\delta(q_t=i)\Big(\log|\Sigma_i| + (u_t-\mu_i)^T\Sigma_i^{-1}(u_t-\mu_i)\Big) - \frac{TKd}{2}\log 2\pi
\end{aligned}
$$

For the k-th expectation step, note that

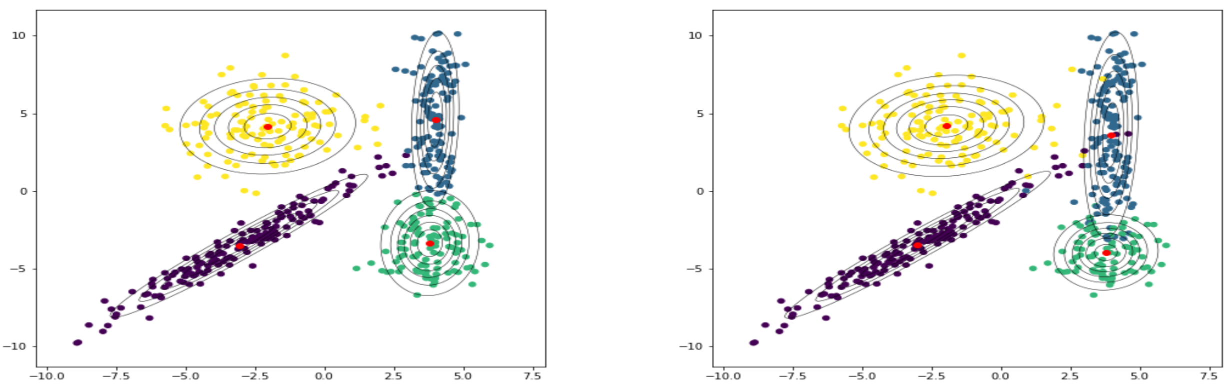
$$
\begin{aligned}
E(\ell_c(\theta)|u_0,\dots,u_T,\theta^{k-1}) = &\sum_{i=1}^{K}P(q_0=i|u;\theta^{k-1})\log\pi_i + \sum_{t=0}^{T-1}\sum_{i,j}^{K}P(q_{i+1}=i,q_t=i|u;\theta^{k-1})\log A_{i,j} \\
&-\frac{1}{2}\sum_{t=0}^{T}\sum_{i=1}^{K}P(q_t=i|u;\theta^{k-1})\Big(\log|\Sigma_i| + (u_t-\mu_i)^T\Sigma_i^{-1}(u_t-\mu_i)\Big) + C
\end{aligned}
$$

Note that this quantity is separable in each of the parameters $\mu_i, \Sigma_i, \pi_i, A_{i,j}$ The maximization step with respect to μ_i and Σ_i can be carried out directly by setting the gradients to 0, yielding

$$
\boxed{\Sigma_i^k = \frac{\sum_{t=0}^{T}P(q_t=i|u;\theta^{k-1})(u_t-\mu_i^k)(u_t-\mu_i^k)^T}{\sum_{t=0}^{T}P(q_t=i|u;\theta^{k-1})} \qquad \mu_i^k = \frac{\sum_{t=0}^{T}P(q_t=i|u;\theta^{k-1})u_t}{\sum_{t=0}^{T}P(q_t=i|u;\theta^{k-1})}}
$$

Maximization for the other parameters can be done with Lagrange multipliers and we find

$$
\boxed{\pi_i^k = P(q_0=i|u;\theta^{k-1}) \qquad A_{i,j}^k = \frac{\sum_{t=0}^{T-1}P(q_{t+1}=j,q_t=i|u;\theta^{k-1})}{\sum_{t=0}^{T-1}\sum_{j'=1}^{K}P(q_{t+1}=j',q_t=i|u;\theta^{k-1})}}
$$



(a) EM-GMM

(b) EM-HMM

1 Exercise 1.5

The results of EM-GMM and EM-HMM[1] are reported in the table below, here the log-likelihood divided by the sample size. They perform similarly on the train and test datasets. Nonetheless, the clusters proposed by EM-GMM seem to be more relevant than EM-HMM. The big difference lies in the set-up of the green and blue clusters owing to the transition kernel A. Roughly for EM-GMM, things stand as $A = I_K$, but with our estimates we get the following transition kernel:

Therefore many transitions took place between clusters green and blue so that the algorithm ends up by mixing them often.

	train	test
GMM	-4.662112	-4.839995
HMM	-3.825067	-3.923919

Table 1: normalized log-likelihood

	yellow	blue	purple	green
yellow	0.906553	0.072861	0.020584	0.000002
blue	0.032437	0.022553	0.011531	**0.933479**
purple	0.034126	0.046542	0.877638	0.041694
green	0.063033	**0.873919**	0.047294	0.015754

Table 2: estimated transition kernel

[1] EM-HMM was initialized with a uniform distribution