1 Estimation equations of EM

The model is composed of 3 unknowns : π_0 (the initial distribution), $A_{i,j}$ (the matrix of transition from state i to j) and f (the function of the conditional probability of the known variable according to the latent variable, i.e. $p(y_t|z_t) = f(y_t, z_t)$) We will group the 3 unknowns under the variable θ

The complete likelihood of the HMM model is :

$$l_c(\theta) = log\Big(p(z_0)\prod_{t=0}^{T-1} p(z_{t+1}|z_t)\prod_{t=0}^{T} f(\bar{y}_t|z_t)\Big)$$

$$l_c(\theta) = log(p(z_0)) + \sum_{t=0}^{T-1} log(p(z_{t+1}|z_t)) + \sum_{t=0}^{T} log(f(\bar{y}_t|z_t))$$

$$l_c(\theta) = \sum_{i=1}^{K} \delta(z_0 = i) log(\pi_0)_i + \sum_{t=0}^{T-1}\sum_{i,j=1}^{K} \delta(z_{t+1} = j, z_t = i) log(A_{i,j}) + \sum_{t=0}^{T}\sum_{i=1}^{K} \delta(z_t = i) log(f(\bar{y}_t|z_t))$$

As seen in the class notes, When applying E-M to estimate the parameters of this HMM, we use Jensen's inequality to obtain a lower bound on the log-likelihood:
$logp(\bar{y_0}, ..., \bar{y_T}) \geq E_q\Big[logp(z_0, ..., z_T, \bar{y_0}, ..., \bar{y_T})\Big] = E_q\Big[l_c(\theta)\Big]$

E-M algorithm :

k-th E-step:
we use $q(z_0, ..., z_T) = P(z_0, ..., z_T|\bar{y_0}, ..., \bar{y_T}; \theta_{k-1})$, and this boils down to applying the following rules :
$E[\delta(z_0 = i)|\bar{y}] = p(z_0 = i|\bar{y}; \theta_{k-1})$
$E[\delta(z_t = i)|\bar{y}] = p(z_t = i|\bar{y}; \theta_{k-1})$
$E[\delta(z_{t+1} = j, z_t = i)|\bar{y}] = p(z_{t+1} = j, z_t = i|\bar{y}; \theta_{k-1})$
Thus, in the former expression of the complete log-likelihood, we just have to replace $\delta(z_0 = i)$ by $p(z_0 = i|y; \theta_{k1})$, and similarly for the other terms.

k-th M-step:
We use the fact that we replaced the δ function by their corresponding terms. This leads to a decoupling of the terms. We can thus easily maximize the log-likelihood according to all elements of θ knowing their constraints which are :
$\sum_{i=1}^{K}(\pi_0)_i = 1$
$\sum_{i,j=1}^{K} A_{i,j} = 1$

This leads to :

$$(\hat{\pi}_0)_i = p(z_0 = i|\bar{y}; \theta_{k-1})$$

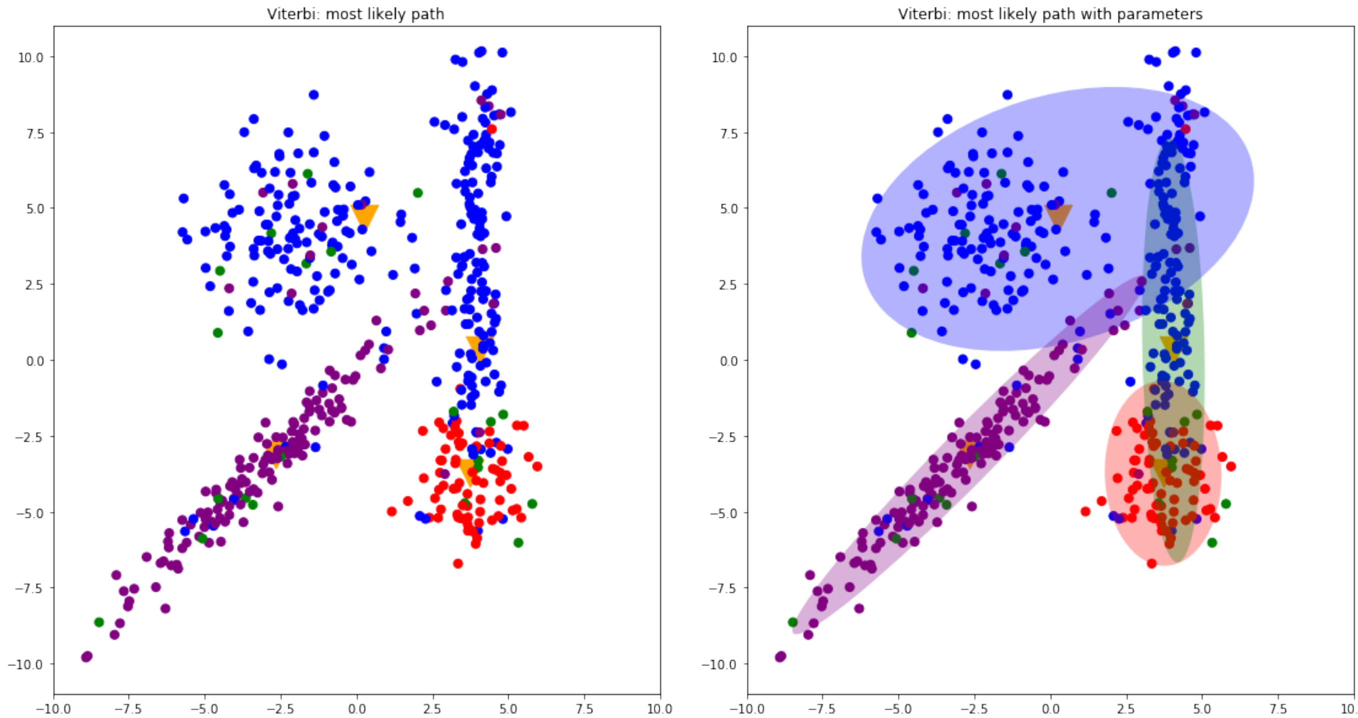
$$\hat{A}_{i,j} = \frac{\sum_{t=0}^{T-1} p(z_{t+1} = j, z_t = i|\bar{y}; \theta_{k-1})}{\sum_{t=0}^{T-1} p(z_t = i|\bar{y}; \theta_{k-1})}$$

In this model we consider that $y_t|z_t = i \sim$ N (μ_i, Σ_i)
Thus,

$$\hat{\mu}_i = \frac{\sum_{t=0}^{T} p(z_t = i|\bar{y}; \theta_{k-1})\bar{y}_t}{\sum_{t=0}^{T} p(z_t = i|\bar{y}; \theta_{k-1})}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=0}^{T} p(z_t = i|\bar{y}; \theta_{k-1})(\bar{y}_t - \mu_i)(\bar{y}_t - \mu_i)^T}{\sum_{t=0}^{T} p(z_t = i|\bar{y}; \theta_{k-1})}$$

These are information that we can access thanks to the α and β recursion that we implemented in the question 1.



From these graphics it appears that we might have a problem in our EM (or Viterbi) implementation. Indeed the "green" cluster is under-represented and even though its parameters on the right graph seem reasonable, our most likely path almost never give the green cluster value.

For 6 EM iterations, we obtain a log likelihood of -4600 for the training set, and -10000 for the testing set. It is much lower than values we had in HW2 (around -5 for the Gaussian Mixture), but it doesn't make sense to compare since these two values because the HMM model tries to take into account a new dimension which is temporality (thus dependence between subsequent data) and thus brings more complexity. Having lower likelihood values is thus expected.