HWK 3 - Graphical Models

Alejandro de la Concha (alexdavidhalo@gmail.com)

Tong ZHAO (tong.zhao@eleves.enpc.fr)

Hidden Markov Model

Question 1.2

Let $\{z_t\}_{t=0}^T$ denote the hidden states of a Hidden Markov Model, $\{y_t\}_{t=0}^T$ the observed variables and θ the parameters of the model. In this exercise we suppose $\{z\}_{t=0}^T$ takes k different values and $P(y_t|z_t = i)$ is the density function of a multivariate normal distribution with mean vector μ_i and covariance matrix Σ_i.

The likelihood of the model is:

$$L(\theta) = p(z_0) \prod_{t=0}^{T-1} p(z_{t+1}|z_t) \prod_{t=0}^{T} p(y_t|z_t)$$

Then the loglikelihood can be written as:

$$\begin{aligned} l(\theta) &= \log(p(z_0)) + \sum_{t=0}^{T-1} \log(p(z_{t+1}|z_t)) + \sum_{t=0}^{T} \log(p(y_t|z_t)) \\ &= \sum_{i=1}^{K} \delta(z_0 = i) \log((\pi_0)_i) + \sum_{t=0}^{T-1} \sum_{i,j=1}^{K} \delta(z_{t+1} = i, z_t = j) \log(A_{i,j}) + \sum_{t=0}^{T} \sum_{i=1}^{K} \delta(z_t = i) \log(P(y_t|z_t = i)) \end{aligned}$$

where π_0 is the initial distribution of the Markov chain, and $A_{i,j}$ is its transition matrix.

The E-step during the iteration n will be:

$$\begin{aligned} E_{z|x}[l(\theta^n)] &= \sum_{i=1}^{K} E_{z|x}[\delta(z_0 = i)] \log((\pi_0)_i) + \sum_{t=0}^{T-1} \sum_{i,j=1}^{K} E_{z|x}[\delta(z_{t+1} = i, z_t = j)] \log(A_{i,j}) \\ &+ \sum_{t=0}^{T} \sum_{i=1}^{K} E_{z|x}[\delta(z_t = i)] \log(P(y_t|z_t = i)) \\ &= \sum_{i=1}^{K} P(z_0 = i|y, \theta^{(n)}) \log((\pi_0)_i) + \sum_{t=0}^{T-1} \sum_{i,j=1}^{K} P(z_{t+1} = i, z_t = j|y, \theta^{(n)}) \log(A_{i,j}) \\ &+ \sum_{t=0}^{T} \sum_{i=1}^{K} P(z_t = i|y, \theta^{(n)}) \log(P(y_t|z_t = i)) \end{aligned}$$

The estimation of the parameters $P(z_t|y,\theta^{(n)})$ and $P(z_{t+1},z_t|y,\theta^{(n)})$ needs the SPA. In this way we get the expressions:

$$\begin{aligned} P(z_t|y,\theta^{(n)}) &= \frac{\alpha_t(z_t)\beta_t(z_t)}{\sum_{z_t}\alpha_t(z_t)\beta_t(z_t)} \\ for\ \forall\ t < T & \\ P(z_{t+1},z_t|y,\theta^{(n)}) &= \frac{\alpha_t(z_t)\beta_t(z_{t+1})P(z_{t+1}|z_t)P(y_t|z_{t+1})}{\sum_{z_t}\alpha_t(z_t)\beta_t(z_t)} \end{aligned}$$

where $\alpha_t(z_t)$ $\beta_t(z_t)$ are the alpha and beta messages defined in the lectures, valuated using the parameters $\theta^{(n)}$.

The M-step consists in solving the following optimization problem:

$$\begin{aligned} \max_{\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K} & E_{z|x}[l(\theta^{(n)})] \\ s.t\ & \sum_{i=1}^K \pi_i = 1 \\ & \sum_{j=1}^K A_{i,j} = 1\ i = 1,...,K \end{aligned}$$

By introducing Lagrangian multipliers λ, $\{\nu\}_{i=1}^K$, the Lagrangian can be written as:

$$L(\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K,\lambda,\{\nu_i\}_{i=1}^K) = E_{z|x}[l(\theta^{(n)}] + \lambda(1-\sum_{i=1}^K \pi_i) + \sum_{i=1}^K \nu_i(1-\sum_{j=1}^K A_{i,j})$$

We take the derivatives with respect to each variable and then we get:

$$\begin{aligned} \frac{\partial L(\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K,\lambda,\{\nu_i\}_{i=1}^K)}{\partial(\pi_0)_i} &= \frac{P(z_0=i|y,\theta^{(k)})}{(\pi_0)_i} - \lambda \\ \frac{\partial L(\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K,\lambda,\{\nu_i\}_{i=1}^K)}{\partial A_{k,l}} &= \frac{\sum_{t=0}^{T-1} P(z_{t+1}=l,z_t=k|y,\theta^{(k)})}{A_{k,l}} - v_k \\ \Delta_{\mu_i} L(\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K,\lambda,\{\nu_i\}_{i=1}^K)) &= \sum_{t=0}^T P(z_t=i|y,\theta^{(k)})(\Sigma_i)^{-1}(\mu_i - y_t) \\ \Delta_{(\Sigma_i)^{-1}} L(\pi_0,A,\{\Sigma_i\}_{i=1}^K,\{\mu_i\}_{i=1}^K,\lambda,\{\nu_i\}_{i=1}^K)) &= \frac{1}{2}\sum_{t=0}^T P(z_t=i|y,\theta^{(k)})(\Sigma_i - (y_t-\mu_i)(y_t-\mu_i)^t) \end{aligned}$$

By using the KKT conditions, we deduce that the values of λ and v_k are:

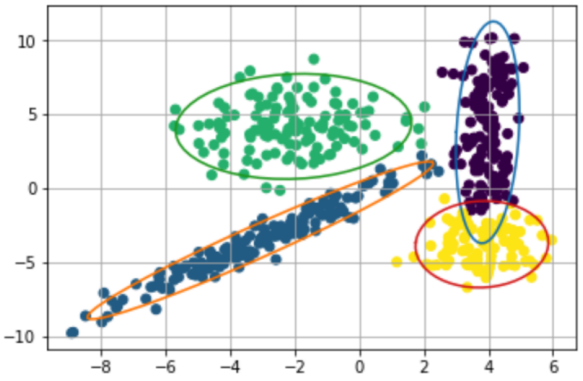
$$\hat{\lambda} = \sum_{i=1}^{K} P(z_0 = i|y, \theta^{(n)})$$
$$\hat{v_k} = \sum_{t=0}^{T-1} \sum_{l=1}^{K} P(z_{t+1} = l, z_t = k|y, \theta^{(n)})$$
$$= \sum_{t=0}^{T-1} P(z_t = k|y, \theta^{(n)})$$

Then we have:

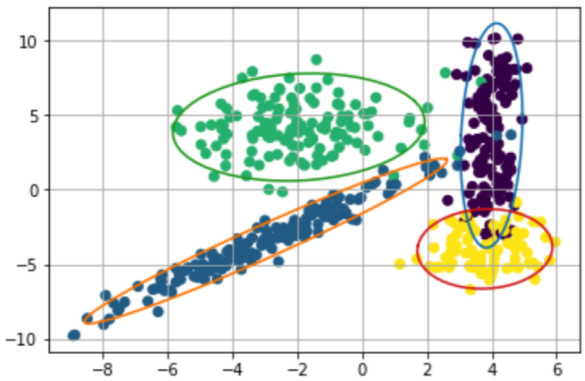
$$(\pi_0)_i^{(n+1)} = \frac{P(z_0 = i|y, \theta^{(n)})}{\sum_{i=1}^{K} P(z_0 = i|y, \theta^{(n)})}$$
$$A_{k,l}^{(n+1)} = \frac{\sum_{t=0}^{T-1} P(z_{t+1} = l, z_t = k|y, \theta^{(n)})}{\sum_{t=0}^{T-1} P(z_t = i|y, \theta^{(n)})}$$
$$\mu_i^{(n+1)} = \frac{\sum_{t=0}^{T} y_t P(z_t = i|y, \theta^{(n)})}{\sum_{t=0}^{T} P(z_t = i|y, \theta^{(n)})}$$
$$(\Sigma_i)^{(n+1)} = \frac{\sum_{t=0}^{T} (y_t - \mu_i)(y_t - \mu_i)^t P(z_t = i|y, \theta^{(n)})}{\sum_{t=0}^{T} P(z_t = i|y, \theta^{(n)})}$$

Question 1.4

Gaussian Mixture Model



Hidden Markov Model



We can observe from the figures that the HMM model produces some samples far from the assigned gaussian distribution, since it takes into consideration not only the probability densities, but also the transition probabilities. This means that the order in which each observation appears in the the data set matters and gives additional information about the cluster they belong.

Question 1.5

We compare the results of three models: the isotropic GMM, the anisotropic GMM and HMM. We use the average log-likelihood of the observations given the estimated parameters as criteria.

Log-likelihood	**Train**	**Test**
Isotropic GMM	-5.4432	5.5283
Anisotropic GMM	-4.6554	-4.8180
HMM	-3.7935	-3.9104

We conclude from the table that the HMM performs better than the other models in both data sets. We can see how , as expected, the more parameters a model has, the smaller the log-likelihood is. All models did not show overfitting issues as the log-likelihoods in both data sets are not very different. Nevertheless, we think that a better way to compare this models would be a criteria that penalizes the number of parameters like BIC or AIC.