We derive the Expectize-Maximize algorithm (EM) for Gaussian Mixture modeling (GMM). In general, any random space can be modeled by a mixture (weighed linear combination) of normal densities. Then, such an arbitrary density can be represented as:

$$f(\boldsymbol{\phi}; \boldsymbol{y}) = \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{\theta}_k; \boldsymbol{y}) \tag{1}$$

where

- K is the total number of densities
- $\boldsymbol{y}$ is the data vector
- $\boldsymbol{\phi} = [\lambda_1 \dots \lambda_K, \boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_K]$ and $\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ represent the mean vector and covariance matrix of the kth density
- $f_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $\{\lambda_k\}$ represent the mixing weights with constraint $\sum_{k=1}^{K} \lambda_k = 1$

From now onwards, we drop the bold face notation for vectors and matrices for convenience. To derive the EM algorithm, we need to identify the missing information. Often identifying this missing information can lead to different results affecting convergence. In the mixture modeling problem, if know the label of the data, i.e., if we know the density index that could have possibly produced the observation, we could easily estimate the mixture parameters by simply collecting all the data corresponding to that particular density. In other words, if we associate each observation to one of the K densities, estimating each density parameters is trivial. Hence, we choose the label as the missing information.

Once we identify the missing information, we need to define the complete data likelihood and marginal unobserved data density conditional on the observed data. A very good explanation of the EM algorithm is given in[1, pp.66-87]

Let $u(n) \in \{1 \dots K\}$ be the missing label corresponding to the nth observation $y(n)$. The marginal density (in fact pmf) of this unobserved data is then given as:

$$p(u(n) = k) = \lambda_k \tag{2}$$

The density of the observed data given the unobserved data is:

$$p(y(n)|u(n)) = f_{u(n)}(y(n); \theta_{u(n)}) \tag{3}$$

Then we obtain the complete data density as:

$$p(y(n), u(n); \phi) = f_{u(n)}(y(n); \theta_{u(n)}) \lambda_{u(n)} \tag{4}$$

The complete data density of N independent observations is then:

$$p(y, u; \phi) = \prod_{n=1}^{N} p(y(n), u(n); \phi) \tag{5}$$

Marginal unobserved data density given the observed data is:

$$p(u|y; \phi) = p(y, u; \phi)/p(y; \phi) \tag{6}$$

where $p(y; \phi)$ is the incomplete or observed data density (likelihood function of the observed data).

The EM algorithm proceeds as follows:

- E-Step: Given an initial estimate of the parameter $\hat{\phi}$, we calculate the expectation of the complete log-likelihood. The expectation is taken over the marginal density of the unobserved data given the observed data. We use $E_{u|y}$ to emphasize the distribution over which the expectation is taken.
- M-step: We maximize this function w.r.t ϕ to obtain an improved (hopefully) estimate of ϕ

E-Step

Since we have independent observations

$$E_{u|y}\left[\ln p(u, y; \phi)\right] = \sum_{n=1}^{N} E_{u(n)|y(n)}\left[\ln p(u(n), y(n); \theta_{u(n)}))\right] \tag{7}$$

Upon expanding the expectation operator and letting $u(n) = k$,

$$E_{k|y(n)}\left[\ln p(k, y(n); \theta_k))\right] = \sum_{k=1}^{K} \left[\ln p(k, y(n); \theta_k))\right] \frac{p(y(n), k; \hat{\theta}_k)}{p(y(n); \hat{\phi})} \tag{8}$$

The dependence of $u(n)$ on k becomes obvious if we recollect that $u(n) \in \{1 \dots K\}$. Also, the denominator is only a function of the data and hence can be ignored while performing the M-step. Hence, we restate the above equation as:

$$E_{k|y(n)}\left[\ln p(k, y(n); \theta_k))\right] \propto \sum_{k=1}^{K} \left[\ln p(k, y(n); \theta_k))\right] p(y(n), k; \hat{\theta}_k) \tag{9}$$

Then,

$$E_{u|y}\left[\ln p(u, y; \phi)\right] \propto \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{\lambda}_k f_k(y(n); \hat{\theta}_k) \left[\ln \lambda_k - \frac{1}{2} \exp\left[(y(n) - \mu_k)^T \Sigma_k^{-1} (y(n) - \mu_k)\right] - \ln |\Sigma_k|\right] \tag{10}$$

The above function has to be maximized subject to the constraint $\sum_{k=1}^{K} \lambda_k = 1$ and define it as $Q(\phi; \hat{\phi})$.

M-Step

$$\hat{\phi}_{new} = \arg\max_{\phi} Q(\phi; \hat{\phi}) \tag{11}$$

Let φ be the Lagrange multiplier. The then function to be maximized is:

$$U(\phi; \hat{\phi}) = Q(\phi; \hat{\phi}) - \varphi \left(\sum_{k=1}^{K} \lambda_k - 1 \right) \tag{12}$$

To obtain $\hat{\lambda}_{k,new}$, we set the derivative of $U(\phi; \hat{\phi})$ w.r.t λ_k to zero and solve for the roots and similarly for μ_k and Σ_k. After some straightforward algebraic manipulations, we obtain the expressions for the estimates as [2, Chapter 5]:

$$\hat{\lambda}_{k,new} = \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{\lambda}_k f_k(y(n); \hat{\theta}_k)}{f(y(n); \hat{\phi})} \tag{13}$$

$$\hat{\mu}_{k,new} = \frac{1}{N\hat{\lambda}_{k,new}} \sum_{n=1}^{N} \frac{\hat{\lambda}_k f_k(y(n); \hat{\theta}_k)}{f(y(n); \hat{\phi})} y(n) \tag{14}$$

$$\hat{\Sigma}_{k,new} = \frac{1}{N\hat{\lambda}_{k,new}} \sum_{n=1}^{N} \frac{\hat{\lambda}_k f_k(y(n); \hat{\theta}_k)}{f(y(n); \hat{\phi})} \left[(y(n) - \hat{\mu}_k)^T (y(n) - \hat{\mu}_k) \right] \tag{15}$$

References

[1] Aleksander Dogandžić, "Detection and estimation theory," Spring, 2005, class notes #3, ECpE Dept., Iowa State University, url: home.eng.iastate.edu/ ald/EE527.html.

[2] Soma Sekhar Dhavala, *Time-frequency representations: Analysis, Synthesis and Implementation*, M. S, Indian Institute of Technology, Madras, March 2000, url: www.asankhya.org/me/msthesisiitm.html.