Derivation of the Expectation-Maximization Algorithm for Gaussian Mixture Models

Miloš Tomić

Problem Setup

Given a dataset $\{y_n\}_{n=1}^N$ of D-dimensional vectors, we assume a latent variable $z_n \in \{1, \ldots, K\}$ for each y_n representing its origin cluster. The generative model:

- $z_n \sim \text{Categorical}(\pi_1, \ldots, \pi_K)$
- $y_n \mid z_n = k \sim \mathcal{N}(\mu_k, \Sigma_k)$

Log-Likelihood and Lower Bound

The marginal log-likelihood is given by:

$$\ell(\theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta) = \sum_n \log \sum_{z_n} p(y_n, z_n \mid \theta)$$

Introducing $q_n(z_n)$ (arbitrary distribution over z_n):

$$\ell(\theta) = \sum_n \log \sum_{z_n} q_n(z_n) \frac{p(y_n, z_n \mid \theta)}{q(z_n)} \overset{\text{Jensen's inq.}}{\geq} \sum_n \sum_{z_n} q_n(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n(z_n)}$$

RHS is a lower bound on the log-likelihood. We can further write it in the form:

$$\ell(\theta) \geq \sum_n \sum_{z_n} \log \frac{p(z_n \mid y_n; \theta) \times p(y_n \mid \theta)}{q_n(z_n)} = \sum_n \left[\sum_{z_n} q_n(z_n) \log \frac{p(z_n \mid y_n; \theta)}{q_n(z_n)} + \sum_{z_n} q_n(z_n) \log p(y_n \mid \theta) \right]$$

which can be rewritten as:

$$\sum_n \left[-D_{KL}(q_n(z_n) \| p(z_n \mid y_n; \theta)) + \log p(y_n \mid \theta) \right]$$

where D_{KL} is the Kullback-Leibler divergence. The first term is always non-negative, so we can write the lower bound as:

$$\ell(\theta) \geq \sum_n \log p(y_n \mid \theta)$$

for $q_n(z_n) = p(z_n \mid y_n; \theta)$, to minimize the KL divergence.

Expectation Step of EM Algorithm

In the E-step, we compute the posterior distribution of the latent variables given the data and the current parameters:

$$q_n(z_n)^{(t)} = p(z_n \mid y_n; \theta^{(t)}) = \frac{p(y_n, z_n \mid \theta^{(t)})}{p(y_n \mid \theta^{(t)})} = \frac{\pi_k^{(t)} \mathcal{N}(y_n \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(y_n \mid \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})}$$

where $\mathcal{N}(y_n \mid \mu_k^{(t)}, \Sigma_k^{(t)})$ is the Gaussian density function with mean $\mu_k^{(t)}$ and covariance $\Sigma_k^{(t)}$, and $\pi_k^{(t)}$ is the mixing coefficient for cluster k at iteration t. We will denote the posterior probability of $z_n = k$ as $\gamma_{nk}^{(t)}$ as in the literature:

$$\gamma_{nk}^{(t)} = p(z_n = k \mid y_n; \theta^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(y_n \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(y_n \mid \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})}$$

Maximization Step of EM Algorithm

In the M-step, we maximize the expected log-likelihood with respect to the parameters θ using the posterior probabilities computed in the E-step:

$$\ell^{(t)}(\theta) = \sum_n \sum_{z_n} q_n^{(t)}(z_n) \log \frac{p(y_n, z_n \mid \theta)}{q_n^{(t)}(z_n)}$$

since the denominator does not depend on θ, we can ignore it when maximizing. We can write the log-likelihood as:

$$\ell^{(t)}(\theta) = \sum_n \sum_{z_n} q_n^{(t)}(z_n) \log p(z_n \mid \theta) p(y_n \mid z_n; \theta) = \sum_n \sum_{k=1}^{K} \gamma_{nk}^{(t)} \log \pi_k \mathcal{N}(y_n \mid \mu_k, \Sigma_k)$$

rewriting the log-likelihood as a function of the parameters θ we get:

$$\begin{aligned} \text{maximize} \quad & \sum_n \sum_k \gamma_{nk}^{(t)} \log \pi_k - \frac{1}{2} \sum_n \sum_k \gamma_{nk}^{(t)} \left[(y_n - \mu_k)^\top \Sigma_k^{-1} (y_n - \mu_k) + \log |\Sigma_k| \right] \\ \text{s.t.} \quad & \sum_k \pi_k = 1 \end{aligned}$$

We get the updated parameters as zeros of the gradients of the log-likelihood wrt. the parameters. The gradients are given by:

$$\frac{\partial \ell^{(t)}}{\partial \pi_k} = \frac{1}{\pi_k} \sum_n \gamma_{nk}^{(t)} - \lambda \quad \text{for } k = 1, \dots, K$$

we get the Lagrange multiplier λ from the constraint $\sum_k \pi_k = 1$:

$$\sum_k \frac{\sum_n \gamma_{nk}^{(t)}}{\lambda} = 1 \implies \lambda = \sum_k \sum_n \gamma_{nk}^{(t)} = \sum_n \sum_k \gamma_{nk}^{(t)} = N$$

$$\pi_k^{(t+1)} = \frac{\sum_n \gamma_{nk}^{(t)}}{N} \quad \text{for } k = 1, \dots, K$$

The gradients for the covariance matrices are given by:

$$\frac{\partial \ell^{(t)}}{\partial \Sigma_k} = \frac{\partial \ell^{(t)}}{\partial \Sigma_k^{-1}} \frac{\partial \Sigma_k^{-1}}{\partial \Sigma_k} = -\frac{1}{2} \sum_n \gamma_{nk}^{(t)} \left[(y_n - \mu_k)(y_n - \mu_k)^\top - \Sigma_k \right] \frac{\partial \Sigma_k^{-1}}{\partial \Sigma_k} = 0$$

by setting the gradient to zero we get:

$$\sum_n \gamma_{nk}^{(t)} (y_n - \mu_k)(y_n - \mu_k)^\top = \sum_n \gamma_{nk}^{(t)} \Sigma_k \implies \Sigma_k^{(t+1)} = \frac{\sum_n \gamma_{nk}^{(t)} (y_n - \mu_k^{(t+1)})(y_n - \mu_k^{(t+1)})^\top}{\sum_n \gamma_{nk}^{(t)}} \quad \text{for } k = 1, \dots, K$$

The gradients for the means are given by:

$$\frac{\partial \ell^{(t)}}{\partial \mu_k} = -\frac{1}{2}\sum_n \gamma_{nk}^{(t)} \frac{\partial}{\partial \mu_k}\left[(y_n - \mu_k)^\top \Sigma_k^{-1}(y_n - \mu_k)\right] = 0$$

Using the fact that the covariance matrix is symmetric and the fact that $\frac{\partial}{\partial x} x^\top A x = 2Ax$ for a symmetric matrix A, we get:

$$\frac{\partial \ell^{(t)}}{\partial \mu_k} = -\frac{1}{2}\sum_n \gamma_{nk}^{(t)}\left[-2\Sigma_k^{-1}(y_n - \mu_k)\right] = 0 \implies \sum_n \gamma_{nk}^{(t)}(y_n - \mu_k) = 0$$

Thus, we have:

$$\mu_k^{(t+1)} = \frac{\sum_n \gamma_{nk}^{(t)} y_n}{\sum_n \gamma_{nk}^{(t)}} \quad \text{for } k = 1, \ldots, K$$