Gaussian Mixture Models (GMM)
Maximum Likelihood & EM

Model: Finite Gaussian Mixture

Data: $\{x_n\}_{n=1}^N$, $x_n \in \mathbb{R}^d$ (IID).
GMM density:

$$p(x \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \qquad \theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K,$$

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \Sigma_k \succ 0.$$

Notation: $\phi_k(x) := \mathcal{N}(x \mid \mu_k, \Sigma_k)$.

Log-Likelihood and a Direct Gradient Attempt

Log-likelihood:

$$\ell(\theta) := \sum_{n=1}^{N} \log\Big(\sum_{k=1}^{K} \pi_k \, \phi_k(x_n) \Big).$$

Derivative wrt μ_k (outline):

$$\frac{\partial \ell}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \, \phi_k(x_n)}{\sum_j \pi_j \phi_j(x_n)} \, \Sigma_k^{-1}(x_n - \mu_k) = \sum_{n=1}^{N} r_{nk} \, \Sigma_k^{-1}(x_n - \mu_k),$$

where **responsibilities**

$$r_{nk} := \Pr(z_{nk} = 1 \mid x_n, \theta) = \frac{\pi_k \, \phi_k(x_n)}{\sum_{j=1}^{K} \pi_j \, \phi_j(x_n)}.$$

Stationary condition $(\frac{\partial \ell}{\partial \mu_k} = 0) \Rightarrow \sum_n r_{nk}(x_n - \mu_k) = 0 \Rightarrow \mu_k = \dfrac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}.$

Issue: r_{nk} depends on θ (circular).
Fix: Expectation–Maximization (EM).

Latent Variables and Complete Data

Introduce one-hot latent $z_{nk} \in \{0, 1\}$ with $\sum_k z_{nk} = 1$.
Complete-data log-likelihood:

$$\ell_c(\theta) := \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big[\log \pi_k + \log \phi_k(x_n) \Big],$$

$$\log \phi_k(x_n) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^\top \Sigma_k^{-1} (x_n - \mu_k).$$

EM idea: iterate E-step: $r_{nk} := \mathbb{E}[z_{nk} \mid x_n, \theta^{\text{old}}]$; M-step: $\theta^{\text{new}} = \arg\max_\theta \mathbb{E}[\ell_c(\theta) \mid X, \theta^{\text{old}}]$.

E-Step: Responsibilities (Bayes Rule)

Posterior over components:

$$r_{nk} = \Pr\big(z_{nk} = 1 \mid x_n, \theta^{\text{old}}\big) = \frac{\pi_k^{\text{old}}\, \mathcal{N}(x_n \mid \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^{K} \pi_j^{\text{old}}\, \mathcal{N}(x_n \mid \mu_j^{\text{old}}, \Sigma_j^{\text{old}})}.$$

Define effective counts $N_k := \sum_{n=1}^{N} r_{nk}$.

M-Step Objective: $Q(\theta \mid \theta^{\text{old}})$

Expected complete-data log-likelihood:

$$Q(\theta \mid \theta^{\text{old}}) = \mathbb{E}_{Z|X,\theta^{\text{old}}}[\ell_c(\theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\Big[\log \pi_k + \log \phi_k(x_n)\Big].$$

Maximize Q w.r.t. $\{\pi_k, \mu_k, \Sigma_k\}$ *separately for each* k.

M-Step for Mixing Weights π_k

Constraint: $\sum_k \pi_k = 1$, $\pi_k \geq 0$.
Lagrangian:

$$\mathcal{L}(\pi, \lambda) = \sum_{k=1}^{K} \Big(\sum_{n=1}^{N} r_{nk} \Big) \log \pi_k + \lambda \Big(\sum_{k=1}^{K} \pi_k - 1 \Big).$$

FOC:

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda = 0 \;\Rightarrow\; \pi_k = -\frac{N_k}{\lambda}.$$

Sum over k and enforce $\sum_k \pi_k = 1$:

$$1 = \sum_k \pi_k = -\frac{1}{\lambda} \sum_k N_k = -\frac{N}{\lambda} \;\Rightarrow\; \lambda = -N \;\Rightarrow\; \boxed{\pi_k^{\text{new}} = \frac{N_k}{N}}.$$

M-Step for Means μ_k

Extract the μ_k terms of Q:

$$Q_\mu = -\frac{1}{2}\sum_{n=1}^{N} r_{nk}\,(x_n - \mu_k)^\top \Sigma_k^{-1}(x_n - \mu_k) + \text{const.}$$

Gradient wrt μ_k:

$$\frac{\partial Q_\mu}{\partial \mu_k} = \sum_{n=1}^{N} r_{nk}\,\Sigma_k^{-1}(x_n - \mu_k).$$

Set to zero (use $\Sigma_k^{-1} \succ 0$):

$$\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \;\Rightarrow\; \boxed{\mu_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\,x_n}.$$

M-Step for Covariances Σ_k (1/2)

Extract the Σ_k terms:

$$Q_\Sigma = -\frac{1}{2}\sum_{n=1}^{N} r_{nk}\Big[\log|\Sigma_k| + (x_n-\mu_k)^\top \Sigma_k^{-1}(x_n-\mu_k)\Big] + \text{const.}$$

Use matrix differentials:

$$\mathrm{d}\log|\Sigma_k| = \mathrm{tr}(\Sigma_k^{-1}\mathrm{d}\Sigma_k), \quad \mathrm{d}\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big) = -\mathrm{tr}\big(\Sigma^{-1}(x-\mu)(x-\mu)^\top\Sigma^{-1}\mathrm{d}\Sigma\big).$$

M-Step for Covariances Σ_k (2/2)

Gradient wrt Σ_k:

$$\frac{\partial Q_\Sigma}{\partial \Sigma_k} = -\frac{1}{2}\sum_{n=1}^{N} r_{nk}\left[\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^\top \Sigma_k^{-1}\right].$$

Set to zero and multiply on both sides by Σ_k:

$$\sum_{n=1}^{N} r_{nk}\left[I - (x_n - \mu_k)(x_n - \mu_k)^\top \Sigma_k^{-1}\right] = 0$$

$$\implies \sum_{n=1}^{N} r_{nk}(x_n - \mu_k)(x_n - \mu_k)^\top = \Big(\sum_{n=1}^{N} r_{nk}\Big)\Sigma_k.$$

$$\boxed{\Sigma_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\,(x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^\top}.$$

EM Algorithm for GMM (Summary)

Given initial $\{\pi_k, \mu_k, \Sigma_k\}$, repeat until convergence:

1. **E-step:**

$$r_{nk} = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}; \quad N_k = \sum_n r_{nk}.$$

2. **M-step:**

$$\pi_k \leftarrow \frac{N_k}{N}, \qquad \mu_k \leftarrow \frac{1}{N_k} \sum_n r_{nk} x_n, \qquad \Sigma_k \leftarrow \frac{1}{N_k} \sum_n r_{nk} (x_n - \mu_k)(x_n - \mu_k)^\top.$$

Guarantee: Each EM iteration does not decrease $\ell(\theta)$.