EM Theory: (Statistician's View)

Data: $\{X_i\}_{i=1}^n$ is the sample data points; $\{y_i\}_{i=1}^n$ is the corresponding labels of which Gaussian X_i is fro
$(X_i \in \mathbb{R}^m)$ (observed) $(y_i \in \mathbb{R}^k)$ (missing)

Model: $y_i \overset{iid}{\sim} \text{Categorical}(\pi_1, \pi_2 \cdots \pi_K)$; (i.e. $y_i = (y_i^1 \cdots y_i^k)$, exact one $y_i^k = 1$, all others are 0,

$p(y_i = (0, 0 \cdots \underset{k\text{th}}{1}, 0 \cdots 0)) = p(y_i^k = 1) = \pi_k$)

$X_i \mid y_i^k = 1 \sim N(\mu_k, \Sigma_k)$; (i.e. pdf. $f(x_i \mid y_i^k = 1) = N(x_i, \mu_k, \Sigma_k) = (2\pi)^{-\frac{m}{2}} |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)}$)

Inference: parameters $\theta = (\pi_1 \cdots \pi_K, \mu_1 \cdots \mu_K, \Sigma_1 \cdots \Sigma_K)$; $X = \{X_i\}_{i=1}^n$; $Y = \{y_i\}_{i=1}^n$;

Bayes Rule: $p(X_i, y_i \mid \theta) = p(X_i \mid \theta)\, p(y_i \mid X_i, \theta)$, So $p(X_i \mid \theta) = \dfrac{p(X_i, y_i \mid \theta)}{p(y_i \mid X_i, \theta)}$, $\ln p(X_i \mid \theta) = \ln p(X_i, y_i \mid \theta) - \ln p(y_i \mid X_i, \theta)$;

$\Rightarrow$ log-likelihood: $l(\theta \mid X) = \sum_{i=1}^n \ln p(X_i \mid \theta) = \sum_{i=1}^n \left[\ln p(X_i, y_i \mid \theta) - \ln p(y_i \mid X_i, \theta)\right]$;

Calculate expectation w.r.t. $y_i \mid X_i, \theta_t$, i.e. $\int \sim\sim\; p(y_i \mid X_i, \theta_t)\, dy_i$, we have:

$E[\ln p(X_i \mid \theta) \mid X_i, \theta_t] = \int \underbrace{\ln p(X_i \mid \theta)}_{\text{no } y_i} p(y_i \mid X_i, \theta_t)\, dy_i = \ln p(X_i \mid \theta)$;

$E[\ln p(X_i, y_i \mid \theta) \mid X_i, \theta_t] = Q_i(\theta \mid \theta_t, X_i)$, $E[\ln p(y_i \mid X_i, \theta) \mid X_i, \theta_t] = H_i(\theta \mid \theta_t, X_i)$;

$$\Rightarrow l(\theta \mid X) = \sum_{i=1}^n \ln p(X_i \mid \theta) = \sum_{i=1}^n E[\ln p(X_i \mid \theta) \mid X_i, \theta_t] = \sum_{i=1}^n \Big(E[\ln p(X_i, y_i \mid \theta) \mid X_i, \theta_t] - E[\ln p(y_i \mid X_i, \theta) \mid X_i, \theta_t]\Big)$$

$$= \sum_{i=1}^n Q_i(\theta \mid \theta_t, X_i) - \sum_{i=1}^n H_i(\theta \mid \theta_t, X_i);$$

For H_i: $H_i(\theta \mid \theta_t, X_i) - H_i(\theta_t \mid \theta_t, X_i) = E\left[\ln \dfrac{p(y_i \mid X_i, \theta)}{p(y_i \mid X_i, \theta_t)} \,\Big|\, X_i, \theta_t\right]$, by Jensen's inequality: $E(\ln X) \le \ln EX$

$$\le \ln E\left[\frac{p(y_i \mid X_i, \theta)}{p(y_i \mid X_i, \theta_t)} \,\Big|\, X_i, \theta_t\right]$$

$$= \ln \int \frac{p(y_i \mid X_i, \theta)}{p(y_i \mid X_i, \theta_t)} \cdot p(y_i \mid X_i, \theta_t)\, dy_i = \ln \underbrace{\int p(y_i \mid X_i, \theta_t)\, dy_i}_{=1} = 0;$$

$\Rightarrow \forall \theta$, $H_i(\theta \mid \theta_t, X_i) \le H_i(\theta_t \mid \theta_t, X_i)$, So $\sum_{i=1}^n H_i(\theta \mid \theta_t, X_i) \le \sum_{i=1}^n H_i(\theta_t \mid \theta_t, X_i)$;

If we set $\theta_{t+1} = \arg\max_\theta \sum_{i=1}^n Q_i(\theta \mid \theta_t, X_i)$, then:
$$\begin{cases} \sum_{i=1}^n Q_i(\theta_{t+1} \mid \theta_t, X_i) \ge \sum_{i=1}^n Q_i(\theta_t \mid \theta_t, X_i) \\ \sum_{i=1}^n H_i(\theta_{t+1} \mid \theta_t, X_i) \le \sum_{i=1}^n H_i(\theta_t \mid \theta_t, X_i) \end{cases}$$

$\Rightarrow l(\theta_{t+1} \mid X) \ge l(\theta_t \mid X)$;

Therefore if we iteratively update θ_t as above, then $l(\theta_t \mid X)$ will be nondecreasing,

So in this way we can obtain the (local) maximum log-likelihood;

How to update

$$\theta_{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} Q_i(\theta|\theta_t, X_i) \quad , \quad Q_i(\theta|\theta_t, X_i) = \int \ln p(X_i, Y_i|\theta)\, p(Y_i|X_i,\theta_t)\, dY_i \ ;$$

$$= \int [\ln p(X_i|Y_i,\theta) + \ln p(Y_i|\theta)]\, p(Y_i|X_i,\theta_t)\, dY_i \ ;$$

denote $Z_i = k$ if $Y_i^k = 1$,

which is the distribution X_i is from;

$$= \sum_{k=1}^{K} p(Z_i=k|X_i,\theta_t) \cdot [\ln N(X_i, \mu_k, \Sigma_k) + \ln \pi_k] \ ; \quad (Y_i \text{ discrete})$$

denote $\alpha_{ik} = p(Z_i=k|X_i,\theta_t)$;

$$= \sum_{k=1}^{K} \alpha_{ik} \left[-\frac{m}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(X_i-\mu_k)^T \Sigma_k^{-1}(X_i-\mu_k) + \ln \pi_k\right] \ ;$$

Here: $\alpha_{ik} = p(Z_i=k|X_i,\theta_t) = \dfrac{p(X_i|Z_i=k,\theta_t)\, p(Z_i=k|\theta_t)}{\sum_{r=1}^{K} p(X_i|Z_i=r,\theta_t)\, p(Z_i=r|\theta_t)}$ (Bayesian Rule: $p(Z_i=k|X_i,\theta_t) = \dfrac{p(X_i|Z_i=k,\theta_t)\,p(Z_i=k|\theta_t)}{p(X_i|\theta_t)}$)

$$= \frac{\pi_{kt} \cdot N(X_i, \mu_{kt}, \Sigma_{kt})}{\sum_{r=1}^{K} \pi_{rt} N(X_i, \mu_{rt}, \Sigma_{rt})} \ ,$$

$\theta_t = (\pi_{1t} .. \pi_{Kt}, \mu_{1t} .. \mu_{Kt}, \Sigma_{1t} .. \Sigma_{Kt})$, the previous θ values, known ;

So α_{ik} is known;

This calculation is the E-step for all $i = 1, 2 \dots n$;

$$\Rightarrow \theta_{t+1} = \arg\max_{\theta} \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{K} \alpha_{ik}\left[\ln \pi_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(X_i-\mu_k)^T\Sigma_k^{-1}(X_i-\mu_k)\right]}_{=Q(\theta)} \ ,$$

$\alpha_{ik} \cdot -\frac{m}{2}\ln 2\pi$ is constant, removed ;

$\left(\sum_{k=1}^{K}\alpha_{ik} = \sum_{k=1}^{K} p(Z_i=k|X_i,\theta_t) = 1\right)$

Usual way: let derivative to be 0 ;

① $\pi_{k\,t+1}$: $\sum_{k=1}^{K}\pi_k = 1$, so $Q(\theta) = Q(\theta) - \lambda(\sum_{k=1}^{K}\pi_k - 1)$, $\dfrac{dQ}{d\pi_k} = \dfrac{1}{\pi_k}\sum_{i=1}^{n}\alpha_{ik} - \lambda$, $\hat{\pi}_k = \dfrac{1}{\lambda}\sum_{i=1}^{n}\alpha_{ik}$; $\sum_{k=1}^{K}\hat{\pi}_k = 1$ gives $\lambda = \dfrac{1}{\sum_{i=1}^{n}\sum_{k=1}^{K}\alpha_{ik}} = \dfrac{1}{n}$;

$$\Rightarrow \pi_{k\,t+1} = \hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\alpha_{ik} \ ;$$

② $\mu_{k\,t+1}$: $\dfrac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{n}\alpha_{ik}\Sigma_k^{-1}(X_i-\mu_k) = \Sigma_k^{-1}\sum_{i=1}^{n}\alpha_{ik}(X_i-\mu_k)$, $\Rightarrow \mu_{k\,t+1} = \dfrac{\sum_{i=1}^{n}\alpha_{ik}X_i}{\sum_{i=1}^{n}\alpha_{ik}}$;

③ $\Sigma_{k\,t+1}$: $\dfrac{\partial Q}{\partial \Sigma_k} = \sum_{i=1}^{n}\alpha_{ik}\left[-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}(X_i-\mu_k)(X_i-\mu_k)^T\Sigma_k^{-1}\right]$ $\Rightarrow \Sigma_{k\,t+1} = \dfrac{\sum_{i=1}^{n}\alpha_{ik}(X_i-\mu_k)(X_i-\mu_k)^T}{\sum_{i=1}^{n}\alpha_{ik}}$;

$$= \frac{1}{2}\Sigma_k^{-1}\sum_{i=1}^{n}\alpha_{ik}[(X_i-\mu_k)(X_i-\mu_k)^T\Sigma_k^{-1} - I]$$

(Matrix differentiation can be searched online, which is not required)

These update-step is the M-step for all $k = 1, 2 .. K$;