Machine Learning

Expectation Maximisation

Hiroshi Shimodaira

March 2026

Ver. 1.0.1

Adapted from Kia Nazarpour's slides

Recap - GMM

- Hard boundaries are exchanged for flexible and probabilistic soft boundaries
- Immense flexibility: $p(\mathbf{x}_n \mid \cdots)$ can take the form of any probability density including Bernoulli distribution
- Expectation Maximisation instead of Maximum Likelihood

Learning Outcomes

1. Move from Gaussian Mixture Models to Latent Variable Models (abstraction)
2. Understand the key motivation behind Expectation Maximisation (EM).
3. Review observed and latent variables.
4. Study the EM formula

References:

1. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2008. (Section 9.4)
2. [M] K. Murphy (Section 8.7)
3. Rogers and Girolami, *A First Course in Machine Learning*, CRC Press, 2016. (Section 6.3)

General latent Variable model

- Two sets of random variables $\mathbf{X}$ and $\mathbf{Z}$.
- $\mathbf{X}$ captures all observed variables.
- $\mathbf{Z}$ captures all unseen/hidden/latent/unobserved variables
- Joint probability model is parameterised by $\boldsymbol{\theta} \in \Theta$ as

$$p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$$

EM - Key motivation

- It is hard to optimise for marginal log-likelihood

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X} \mid \boldsymbol{\theta})$$

- Typically, it is easier to optimise the log-likelihood for the complete data

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

Preparation – Jensen's Inequality

Theorem

If $f : \mathcal{R} \to \mathcal{R}$ is a convex function and x is a random variable, then

$$\mathbb{E}\left[f(x)\right] \geq f(\mathbb{E}\left[x\right])$$

- For example $f(x) = x^2$ is a convex function and $\mathbb{E}\left[x^2\right] \geq (\mathbb{E}\left[x\right])^2$

$$\sigma^2(x) = \mathbb{E}\left[x^2\right] - (\mathbb{E}\left[x\right])^2 \geq 0$$

- for $log()$, which is a concave function (thus the opposite inequality),

$$\log\left(\frac{\sum_{n=1}^{N} x_n}{N}\right) \geq \frac{\sum_{n=1}^{N} \log(x_n)}{N} \tag{1}$$

Preparation – Kullback-Leibler Divergence

- For discrete probability distributions p and q on the same probability space $\mathcal{X}$
- The KL-divergence (of p from q) is defined by

$$D_{\mathsf{KL}}(p \,\|\, q) = \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)}\right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- The KL-divergence measures the "distance" between p and q but
- The KL-divergence **is not a metric**.
- The KL-divergence **is not a symmetric**.

$$\begin{aligned} D_{\mathsf{KL}}(p \,\|\, q) &\geq 0 \\ D_{\mathsf{KL}}(p \,\|\, q) &\neq D_{\mathsf{KL}}(q \,\|\, p) \\ D_{\mathsf{KL}}(p \,\|\, p) &= 0 \end{aligned}$$

Finding the lower bound of the log-likelihood $\ell(\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \;=\; \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \;=\; \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) \tag{2}$$

Finding the lower bound of the log-likelihood $\ell(\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) \tag{2}$$

$$= \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \quad \cdots \; q_n(\boldsymbol{z}_n) \text{: arbitraty distributions over } \boldsymbol{z}_n \tag{3}$$

Finding the lower bound of the log-likelihood $\ell(\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) \tag{2}$$

$$= \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \quad \cdots \; q_n(\boldsymbol{z}_n) \text{: arbitraty distributions over } \boldsymbol{z}_n \tag{3}$$

$$\geq \sum_{n=1}^{N} \underbrace{\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}}_{L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n)) \quad \cdots \quad \text{Evidence Lower Bound (ELBO)}} \quad \cdots \text{ surrogate function} \tag{4}$$

Finding the lower bound of the log-likelihood $\ell(\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) \tag{2}$$

$$= \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \quad \cdots \; q_n(\boldsymbol{z}_n) \text{: arbitraty distributions over } \boldsymbol{z}_n \tag{3}$$

$$\geq \sum_{n=1}^{N} \underbrace{\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}}_{L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n)) \quad \cdots \quad \text{Evidence Lower Bound (ELBO)}} \quad \cdots \text{ surrogate function} \tag{4}$$

$$= \sum_{n=1}^{N} \left\{ \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} + \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \right\} \tag{5}$$

Finding the lower bound of the log-likelihood $\ell(\boldsymbol{\theta})$

$$\ell(\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) \tag{2}$$

$$= \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \quad \cdots \; q_n(\boldsymbol{z}_n) \text{: arbitraty distributions over } \boldsymbol{z}_n \tag{3}$$

$$\geq \sum_{n=1}^{N} \underbrace{\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}}_{L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n)) \quad \cdots \quad \text{Evidence Lower Bound (ELBO)}} \quad \cdots \text{ surrogate function} \tag{4}$$

$$= \sum_{n=1}^{N} \left\{ \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} + \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \right\} \tag{5}$$

$$= \sum_{n=1}^{N} \left\{ -D_{\mathsf{KL}}(q_n(\boldsymbol{z}_n) \,\|\, p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})) + \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \right\} \tag{6}$$

Making the lower band closer to $\ell(\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n)) = \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \tag{7}$$

$$= -D_{\mathsf{KL}}(q_n(\boldsymbol{z}_n) \,\|\, p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})) + \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \tag{8}$$

If we let $q_n(\boldsymbol{z}_n) = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})$, denoted as $q_n^*(\boldsymbol{z}_n)$, $D_{\mathsf{KL}}(q_n(\boldsymbol{z}_n) \,\|\, p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})) = 0$. As a result,

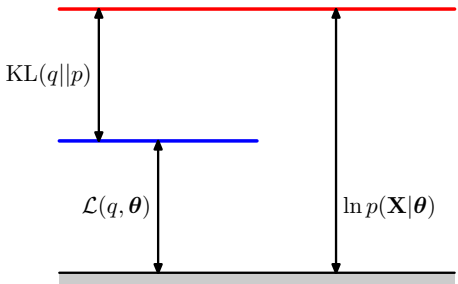
$$L(\boldsymbol{\theta}, q_n^*(\boldsymbol{z}_n)) = \log p(\mathbf{x}_n \,|\, \boldsymbol{\theta}) \;=\; \ell_n(\boldsymbol{\theta}) \tag{9}$$

The new lower bound can be rewritten as:

$$L(\boldsymbol{\theta}, q_n^*(\boldsymbol{z}_n)) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) \log p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}) + H(q_n^*), \tag{10}$$

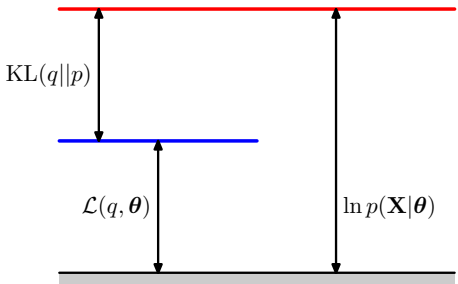
where $H(q) = -\sum_q q \log q$.

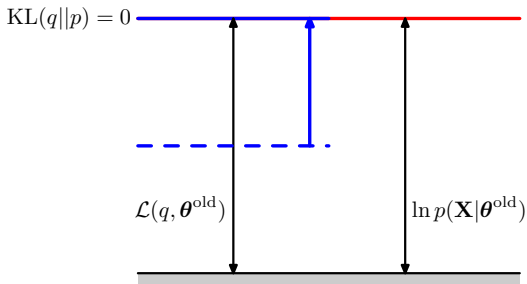
$L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n))$ **vs** $\log p(\mathbf{x}_n \,|\, \boldsymbol{\theta})$



(Credit: Bishop, Figures 9.11 and 9.12)

$L(\boldsymbol{\theta}, q_n(\boldsymbol{z}_n))$ **vs** $\log p(\mathbf{x}_n \,|\, \boldsymbol{\theta})$





(Credit: Bishop, Figures 9.11 and 9.12)

Maximising the lower band

$$L(\boldsymbol{\theta}, q_n^*(\boldsymbol{z}_n)) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) \log p(\mathbf{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta}) + H(q_n^*)$$

Letting

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{old}) \log p(\mathbf{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta}), \tag{11}$$

we have:

$$L(\boldsymbol{\theta}^{new}, q_n^*(\boldsymbol{z}_n)) \geq L(\boldsymbol{\theta}^{old}, q_n^*(\boldsymbol{z}_n)). \tag{12}$$

Expectation Maximisation steps

1. Letting iteration counter $t = 1$,

$$\ell_n(\boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, q_n)$$

Expectation Maximisation steps

1. Letting iteration counter $t = 1$,
$$\ell_n(\boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, q_n)$$
2. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^t)$,
$$\ell_n(\boldsymbol{\theta}^t) = L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$$

Expectation Maximisation steps

1. Letting iteration counter $t = 1$,
$$\ell_n(\boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, q_n)$$
2. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^t)$,
$$\ell_n(\boldsymbol{\theta}^t) = L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$$
3. Letting $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_n L(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$,
$$L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \ell_n(\boldsymbol{\theta}^t)$$

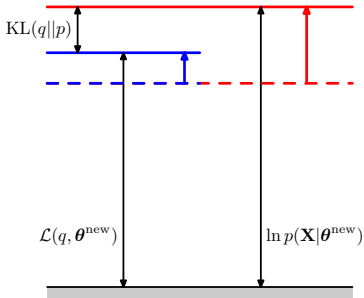
Expectation Maximisation steps

1. Letting iteration counter $t = 1$,
$$\ell_n(\boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, q_n)$$
2. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^t)$,
$$\ell_n(\boldsymbol{\theta}^t) = L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$$
3. Letting $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_n L(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$,
$$L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \ell_n(\boldsymbol{\theta}^t)$$
4. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^{t+1})$,
$$\ell_n(\boldsymbol{\theta}^{t+1}) = L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+1})$$

Expectation Maximisation steps

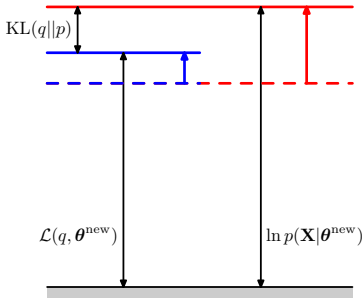
1. Letting iteration counter $t = 1$,
$$\ell_n(\boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, q_n)$$
2. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^t)$,
$$\ell_n(\boldsymbol{\theta}^t) = L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$$
3. Letting $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_n L(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$,
$$L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq L(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \ell_n(\boldsymbol{\theta}^t)$$
4. Letting $q_n = p(\boldsymbol{z}_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}^{t+1})$,
$$\ell_n(\boldsymbol{\theta}^{t+1}) = L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+1})$$
5. Letting $\boldsymbol{\theta}^{t+2} = \arg\max_{\boldsymbol{\theta}} \sum_n L(\boldsymbol{\theta}, \boldsymbol{\theta}^{t+1})$,
$$L(\boldsymbol{\theta}^{t+2}, \boldsymbol{\theta}^{t+1}) \geq L(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+1}) = \ell_n(\boldsymbol{\theta}^{t+1})$$

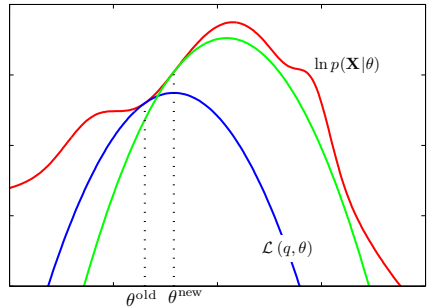
Maximisation in EM



(Credit: Bishop, Figures 9.13 and 9.14)

Maximisation in EM





(Credit: Bishop, Figures 9.13 and 9.14)

Recap - GMM training with EM

1. Initialise the model parameters: $K, \{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}$

Recap - GMM training with EM

1. Initialise the model parameters: $K, \{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}$
2. Calculate $\{r_{nk}\}$ based on the current model [Expectation].

$$r_{nk} = p(z_{nk} = 1 \,|\, \mathbf{x}_n, \boldsymbol{\theta}) \;=\; \frac{\pi_k \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

Recap - GMM training with EM

1. Initialise the model parameters: $K, \{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}$
2. Calculate $\{r_{nk}\}$ based on the current model [Expectation].

$$r_{nk} = p(z_{nk} = 1 \,|\, \mathbf{x}_n, \boldsymbol{\theta}) \;=\; \frac{\pi_k \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

3. Update the model parameters [Maximisation]

$$\pi_k^{(\text{new})} = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \tag{13}$$

$$\boldsymbol{\mu}_k^{(\text{new})} = \frac{\sum_{n=1}^{N} r_{nk}\, \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{14}$$

$$\boldsymbol{\Sigma}_k^{(\text{new})} = \frac{\sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} r_{nk}} \tag{15}$$

Recap - GMM training with EM

1. Initialise the model parameters: $K, \{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}$
2. Calculate $\{r_{nk}\}$ based on the current model [Expectation].

$$r_{nk} = p(z_{nk} = 1 \,|\, \mathbf{x}_n, \boldsymbol{\theta}) \;=\; \frac{\pi_k \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

3. Update the model parameters [Maximisation]

$$\pi_k^{(\text{new})} = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \tag{13}$$

$$\boldsymbol{\mu}_k^{(\text{new})} = \frac{\sum_{n=1}^{N} r_{nk}\, \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{14}$$

$$\boldsymbol{\Sigma}_k^{(\text{new})} = \frac{\sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} r_{nk}} \tag{15}$$

4. Repeat from step 2 until a stooping condition is met.

EM for GMM

E step:

$$r_{nk}^{(t)} = p(z_{nk} = 1 \,|\, \mathbf{x}_n, \boldsymbol{\theta}^t) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(\mathbf{x}_n \,|\, \boldsymbol{\mu}_{k'}, ^{(t)} \boldsymbol{\Sigma}_{k'}^{(t)})} \tag{16}$$

M step:

$$\sum_n L(\boldsymbol{\theta}^t, q) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n, \boldsymbol{z}_n \,|\, \boldsymbol{\theta}^t) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n \,|\, \boldsymbol{z}_n \boldsymbol{\theta}^t)\, p(\boldsymbol{z}_n \,|\, \boldsymbol{\theta}^t) \tag{17}$$

EM for GMM

E step:

$$r_{nk}^{(t)} = p(z_{nk} = 1 \mid \mathbf{x}_n, \boldsymbol{\theta}^t) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, ^{(t)} \boldsymbol{\Sigma}_{k'}^{(t)})} \tag{16}$$

M step:

$$\sum_n L(\boldsymbol{\theta}^t, q) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta}^t) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n \mid \boldsymbol{z}_n \boldsymbol{\theta}^t)\, p(\boldsymbol{z}_n \mid \boldsymbol{\theta}^t) \tag{17}$$

$$= \sum_n \sum_k r_{nk}^{(t)} \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) + \sum_n \sum_k r_{nk}^{(t)} \log \pi_k^{(t)} \tag{18}$$

EM for GMM

E step:

$$r_{nk}^{(t)} = p(z_{nk} = 1 \mid \mathbf{x}_n, \boldsymbol{\theta}^t) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, ^{(t)} \boldsymbol{\Sigma}_{k'}^{(t)})} \tag{16}$$

M step:

$$\sum_n L(\boldsymbol{\theta}^t, q) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta}^t) = \sum_n \sum_k r_{nk}^{(t)} \log p(\mathbf{x}_n \mid \boldsymbol{z}_n \boldsymbol{\theta}^t)\, p(\boldsymbol{z}_n \mid \boldsymbol{\theta}^t) \tag{17}$$

$$= \sum_n \sum_k r_{nk}^{(t)} \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) + \sum_n \sum_k r_{nk}^{(t)} \log \pi_k^{(t)} \tag{18}$$

$$= -\frac{1}{2} \sum_n \sum_k r_{nk}^{(t)} \left\{ \log |\boldsymbol{\Sigma}_k^{(t)}| + \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}^{(t)})^\top \boldsymbol{\Sigma}_k^{(t)^{-1}} (\mathbf{x}_n - \boldsymbol{\mu}^{(t)}) \right\}$$

$$+ \sum_n \sum_k r_{nk}^{(t)} \log \pi_k^{(t)} \qquad \text{(constat terms were ignored)} \tag{19}$$

EM for GMM (*cont.*)

$$\pi_k^{(t)} = \frac{1}{N} \sum_{n=1}^{N} r_{nk}^{(t)} \tag{20}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} r_{nk}^{(t)} \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}^{(t)}} \tag{21}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{n=1}^{N} r_{nk}^{(t)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^\top}{\sum_{n=1}^{N} r_{nk}^{(t)}} \tag{22}$$

See Rogers and Girolami (2016) [pp.218-222] for full derivations

EM - Summary - 1

1. Choose an initial $\boldsymbol{\theta}^{\text{old}}$
2. **Expectation Step** (E step)
   - Let $q^*(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, giving the best lower bound at $\boldsymbol{\theta}^{\text{old}}$
   - Let

$$J(\boldsymbol{\theta}) := (q^*, \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q^*(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^*(\mathbf{Z})} \right)}_{\text{Expectation}}$$

3. **Maximisation Step** (M step)

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

4. Go to step 2 until convergence

EM - Summary - 2

1. Maximum likelihood estimation is easy if we observe all the values of all the relevant random variables.
2. In case of missing data and/or latent variables, then Maximum likelihood estimation becomes hard.
3. In such cases, it is often simpler (but not always faster) to use the EM algorithm.
4. EM alternates between inferring the missing values given the parameters (E step), and then optimising the parameters given the *filled* in data (M step).
5. EM monotonically increases the observed data log likelihood.