

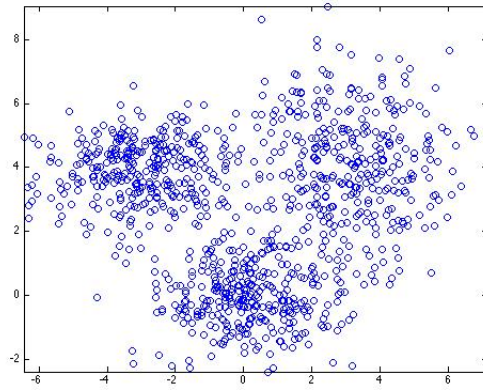
Expectation Maximization

Richard Xu

November 15, 2022

1 Motivation - Mixture Density models

When you have data that looks like:



Can you fit them using a single-mode Gaussian distribution, i.e.,:

$$\begin{aligned} p(X) &= \mathcal{N}(X|\mu, \Sigma) \\ &= (2\pi)^{-k/2} |\Sigma|^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \end{aligned} \quad (1)$$

Clearly Not! This is typically modelling using Mixture Densities, in the case of Gaussian Mixture Model (k-mixture) (GMM):

$$p(X) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \quad \text{s.t.} \quad \sum_{l=1}^k \alpha_l = 1 \quad (2)$$

1.1 Gaussian Mixture model result

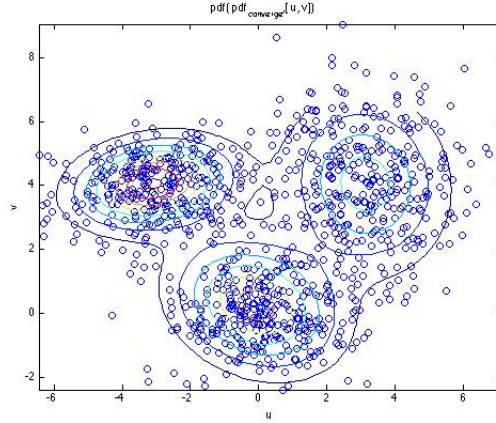


Figure 1: gmm fitting result

Let $\Theta = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$

$$\begin{aligned} \Theta_{\text{MLE}} &= \arg \max_{\Theta} \mathcal{L}(\Theta|X) \\ &= \arg \max_{\Theta} \left(\sum_{i=1}^n \log \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \right) \quad \sum_{l=1}^k \alpha_l = 1 \end{aligned} \quad (3)$$

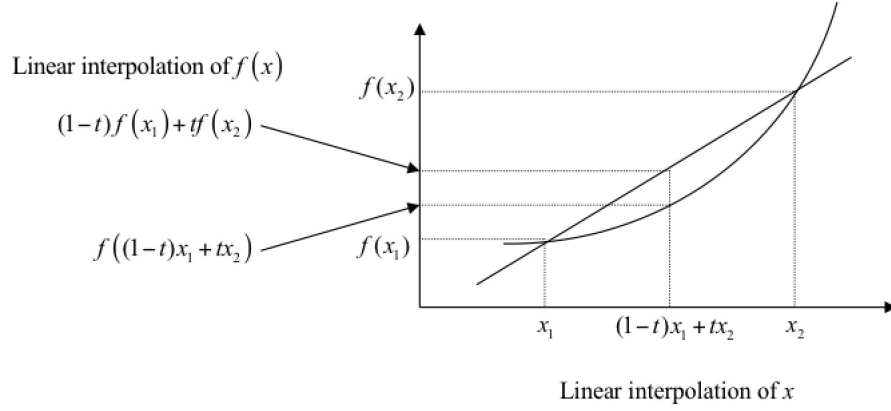
1. Unlike single mode Gaussian, we can't just take derivatives and let it equal zero easily, i.e., optimize it analytically.
2. In terms of optimizing its parameters: the problem is **non-concave**. So traditional gradient ascend is **not** suitable for it, i.e., many local maximums.

The **goal** is to find an iterative process that ensures that $\log(p(X|\Theta^{(g)}))$ remains non-decreasing for $g = 1, \dots$

To do so, for data $X = \{x_1, \dots, x_n\}$, we introduce “latent” variables $Z = \{z_1, \dots, z_n\}$, each z_i indicating which mixture x_i belongs to. The core problem then simply boils down to a single Gaussian fitting.

2 Preliminaries

2.1 Convex function



$$f((1-t)x_1 + tx_2) \leq (1-t)f(x_1) + tf(x_2) \quad t \in (0 \dots 1) \quad (4)$$

2.2 Jensen's inequality

Using notation ϕ instead of f :

$$\phi((1-t)x_1 + tx_2) \leq (1-t)\phi(x_1) + t\phi(x_2) \quad t \in (0 \dots 1) \quad (5)$$

Can be generalized further for any convex combination, i.e., let $\sum_{i=1}^n p_i = 1$:

$$\begin{aligned} \phi(p_1x_1 + p_2x_2 + \dots p_nx_n) &\leq p_1\phi(x_1) + p_2\phi(x_2) \dots p_n\phi(x_n) \quad \sum_{i=1}^n p_i = 1 \\ \implies \phi\left(\sum_{i=1}^n p_ix_i\right) &\leq \sum_{i=1}^n p_i\phi(x_i) \\ \implies \phi\left(\sum_{i=1}^n p_i f(x_i)\right) &\leq \sum_{i=1}^n p_i\phi(f(x_i)) \quad \text{by replacing } x_i \text{ with } f(x_i) \end{aligned} \quad (6)$$

generalize to the continuous case:

$$\phi\left(\int_x f(x)p(x)\right) \leq \int_x \phi(f(x))p(x) \implies \phi(\mathbb{E}[f(x)]) \leq \mathbb{E}[\phi(f(x))] \quad (7)$$

2.2.1 Jensen's inequality example: $-\log(x)$

$\phi(x) = -\log(x)$ is a convex function:

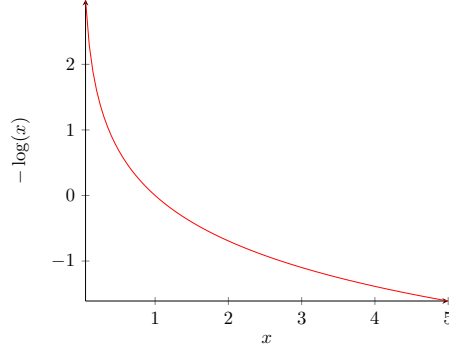


Figure 2: example of convex function $-\log(x)$

1. when $\phi(\cdot)$ is convex:

$$\begin{aligned} \phi(\mathbb{E}[f(\mathbf{x})]) &\leq \mathbb{E}[\phi(f(\mathbf{x}))] \\ \text{e.g. } -\log(\mathbb{E}[f(\mathbf{x})]) &\leq \mathbb{E}[-\log(f(\mathbf{x}))] \end{aligned} \quad (8)$$

2. when $\phi(\cdot)$ is concave:

$$\begin{aligned} \phi(\mathbb{E}[f(\mathbf{x})]) &\geq \mathbb{E}[\phi(f(\mathbf{x}))] \\ \text{e.g. } \log(\mathbb{E}[f(\mathbf{x})]) &\geq \mathbb{E}[\log(f(\mathbf{x}))] \end{aligned} \quad (9)$$

3 Expectation-Maximization Algorithm

Instead of perform:

$$\begin{aligned} \Theta^{\text{MLE}} &= \arg \max_{\Theta} \mathcal{L}(\Theta|X) \\ &= \arg \max_{\Theta} (\log[p(X|\Theta)]) \end{aligned} \quad (10)$$

The trick is to assume some “latent” variable Z to the model
For each iteration of the E-M algorithm, we perform:

$$\Theta^{(g+1)} = \arg \max_{\Theta} \left(\int_z \log(p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) \right) dZ \quad (11)$$

However, before we apply it, we must ensure convergence, i.e.,

$$\begin{aligned}
\log(p(X|\Theta^{(g+1)})) &= \mathcal{L}(\Theta^{(g+1)}|X) \\
&\geq \mathcal{L}(\Theta^{(g)}|X) \\
&= \log(p(X|\Theta^{(g)})) \quad \forall g
\end{aligned} \tag{12}$$

Note the difference between variable Θ and the constant $\Theta^{(g)}$. Also note that gradient ascend is **not** suitable for many E-M problems, as they can be **non-concave**.

4 Proof of convergence: Maximization-Maximization

We have seen it from variational inference literature, for the ELBO-KL decomposition:

$$\begin{aligned}
\mathcal{L}(\Theta|X) &= \log(p(X|\Theta)) \\
&= \log\left(\frac{p(X, Z|\Theta)}{p(Z|X, \Theta)}\right) \\
&= \log\left(\frac{p(X, Z|\Theta)}{q(Z)} \times \frac{q(Z)}{p(Z|X, \Theta)}\right) \\
&= \log\left(\frac{p(X, Z|\Theta)}{q(Z)}\right) + \log\left(\frac{q(Z)}{p(Z|X, \Theta)}\right) \\
&= \int_Z \log\left(\frac{p(X, Z|\Theta)}{q(Z)}\right) q(Z) + \int_Z \log\left(\frac{q(Z)}{p(Z|X, \Theta)}\right) q(Z) \\
&= \text{ELBO}(\Theta, q) + \text{KL}(q(Z)||p(Z|X, \Theta))
\end{aligned} \tag{13}$$

or to use Jensen's inequality:

$$\begin{aligned}
\mathcal{L}(\Theta|X) &= \log p(X|\Theta) = \log \int_Z p(X, Z|\Theta) \\
&= \log \left(\int_Z \frac{p(X, Z|\Theta)}{q(Z)} q(Z) \right) \\
&\geq \int_Z \log \left(\frac{p(X, Z|\Theta)}{q(Z)} \right) q(Z) \\
&= \text{ELBO}(\Theta, q)
\end{aligned} \tag{14}$$

4.1 Maximization of ELBO

When applying E-M as a Maximization-Maximization algorithm, we only need to optimize ELBO, let's revisit the ELBO-KL decomposition again:

$$\begin{aligned}
\mathcal{L}(\Theta|X) &= \int_Z \log\left(\frac{p(X, Z|\Theta)}{q(Z)}\right) q(Z) + \int_Z \log\left(\frac{q(Z)}{p(Z|X, \Theta)}\right) q(Z) \\
&= \text{ELBO}(\Theta, q) + \text{KL}(q(Z)||p(Z|X, \Theta))
\end{aligned} \tag{15}$$

STEP 1: fix $\Theta = \Theta^{(g)}$, maximize $q(Z)$ for ELBO

$$\begin{aligned}
q^*(Z) &= \arg \max_q \{ \text{ELBO}(\Theta^{(g)}, q) \} \\
&= \arg \max_q \left\{ \int_Z \log \left(\frac{p(X, Z | \Theta^{(g)})}{q(Z)} \right) q(Z) \right\} \quad (16) \\
&= p(Z | X, \Theta^{(g)}) \\
\implies \text{ELBO}(\Theta^{(g)}, q^*) &= \mathcal{L}(\Theta^{(g)} | X)
\end{aligned}$$

This can be done with the realization that when fixing $\Theta = \Theta^{(g)}$, ELBO is upper-bounded by $\mathcal{L}(\Theta^{(g)} | X)$, and maximum occurs, i.e., when $\text{KL}(q^* \| p(Z | X, \Theta^{(g)})) = 0$

STEP 1 is invisible to the algorithm, but needs to be considered for interpretation convergence.

STEP 2 Fix $q(Z) = p(Z | X, \Theta^{(g)})$, maximize Θ

substitute $q^* = p(Z | X, \Theta^{(g)})$ back into the ELBO:

$$\begin{aligned}
\Theta^{(g+1)} &= \arg \max_{\Theta} \{ \text{ELBO}(\Theta, q^* = p(Z | X, \Theta^{(g)})) \} \\
&= \arg \max_{q(Z)} \left\{ \int_Z \log \left(\frac{p(X, Z | \Theta^{(g)})}{p(Z | X, \Theta^{(g)})} \right) p(Z | X, \Theta^{(g)}) \right\} \\
&= \arg \max_{\Theta} \left(\int_Z \log(p(X, Z | \Theta)) p(Z | X, \Theta^{(g)}) dZ \right) \quad \text{remove constant terms} \quad (17)
\end{aligned}$$

after obtaining $\Theta^{(g+1)}$, it will introduce a new likelihood function $\mathcal{L}(\Theta^{(g+1)} | X)$ for which the upper bound of the $\text{ELBO}(\Theta^{(g+1)}, q)$ is increased for **STEP 1**

4.2 Proof of convergence without ELBO

let's decompose $\mathcal{L}(\Theta | X)$ into:

$$\begin{aligned}
\mathcal{L}(\Theta | X) &= \log(p(X | \Theta)) \\
&= \log(p(Z, X, \Theta)) - \log(p(Z | X, \Theta)) \quad (18)
\end{aligned}$$

take expectation with respect to both sides with respect to $p(Z | X, \Theta^{(g)})$:

$$\begin{aligned}
\mathcal{L}(\Theta | X) &= \underbrace{\int_Z \log(p(Z, X, \Theta)) p(Z | X, \Theta^{(g)}) dZ}_{Q(\Theta | \Theta^{(g)})} - \underbrace{\int_Z \log(p(Z | X, \Theta)) p(Z | X, \Theta^{(g)}) dZ}_{H(\Theta | \Theta^{(g)})} \quad (19)
\end{aligned}$$

where $H(\Theta|\Theta^{(g)}) = \text{Cross-Entropy}(p(Z|X, \Theta^{(g)})||p(Z|X, \Theta))$.

Eq.(??) can be considered as a simpler version of the ELBO-KL decomposition of $\mathcal{L}(\Theta|X)$ we have seen previously:

$$\mathcal{L}(\Theta|X) = \int_Z \log \left(\frac{p(Z, X|\Theta)}{p(Z|X, \Theta^{(g)})} \right) p(Z|X, \Theta^{(g)}) dZ - \int_Z \log \left(\frac{p(Z|X, \Theta)}{p(Z|X, \Theta^{(g)})} \right) p(Z|X, \Theta^{(g)}) dZ \quad (20)$$

i.e., the added term $\log(p(Z|X, \Theta^{(g)}))p(Z|X, \Theta^{(g)})$ of the two terms cancel out and equivalent to Eq.(??).

4.2.1 why only $Q(\Theta|\Theta^{(g)})$ needs to be maximized

In E-M, we only maximize non Θ part of ELBO:

$$\begin{aligned} \Theta^{(g+1)} &= \arg \max_{\Theta} Q(\Theta|\Theta^{(g)}) \\ &= \arg \max_{\Theta} \left(\int_Z \log(p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) dZ \right) \end{aligned} \quad (21)$$

instead of maximizing the entire Eq.(??), i.e., $Q(\Theta|\Theta^{(g)}) + H(\Theta|\Theta^{(g)})$

the **trick** is, if we can prove:

$$H(\Theta|\Theta^{(g)}) \geq H(\Theta^{(g)}|\Theta^{(g)}) \quad \forall \Theta \quad (22)$$

then we can show:

$$\begin{aligned} \mathcal{L}(\Theta^{(g+1)}) &= Q(\Theta^{(g+1)}|\Theta^{(g)}) + H(\Theta^{(g+1)}|\Theta^{(g)}) \\ &\geq \underbrace{Q(\Theta^{(g)}|\Theta^{(g)})}_{\text{Eq.(??)}} + \underbrace{H(\Theta^{(g)}|\Theta^{(g)})}_{\text{Eq.(??)}} \\ &= \mathcal{L}(\Theta^{(g)}) \end{aligned} \quad (23)$$

it is obvious that:

$$\begin{aligned} \bar{\Theta} &= \arg \max_{\Theta} \{Q(\Theta|\Theta^{(g)}) + H(\Theta|\Theta^{(g)})\} \\ \implies \mathcal{L}(\Theta^{(g+1)}) &\neq \mathcal{L}(\bar{\Theta}) \end{aligned} \quad (24)$$

$$\mathbf{4.2.2} \quad H(\Theta|\Theta^{(g)}) \geq H(\Theta^{(g)}|\Theta^{(g)}) \quad \forall \Theta$$

1. cross entropy

$$\begin{aligned} H(\Theta|\Theta^{(g)}) &= - \int_Z \log(p(Z|X, \Theta)) p(Z|X, \Theta^{(g)}) dZ \\ &= \text{Cross-Entropy}(p(Z|X, \Theta^{(g)}) \| p(Z|X, \Theta)) \\ \implies \arg \min_{\Theta} \{H(\Theta|\Theta^{(g)})\} &= \Theta^{(g)} \\ \implies \min_{\Theta} \{H(\Theta|\Theta^{(g)})\} &= \text{Entropy}(p(Z|X, \Theta^{(g)})) \end{aligned} \tag{25}$$

2. directly

$$\begin{aligned} &H(\Theta|\Theta^{(g)}) - H(\Theta^{(g)}|\Theta^{(g)}) \\ &= \int_Z -\log(p(Z|X, \Theta)) p(Z|X, \Theta^{(g)}) dz - \int_Z -\log(p(Z|X, \Theta^{(g)})) p(Z|X, \Theta^{(g)}) dZ \\ &= \int_Z \log\left(\frac{p(Z|X, \Theta^{(g)})}{p(Z|X, \Theta)}\right) p(Z|X, \Theta^{(g)}) dZ \\ &= \int_Z -\log\left(\frac{p(Z|X, \Theta)}{p(Z|X, \Theta^{(g)})}\right) p(Z|X, \Theta^{(g)}) dZ \\ &\geq -\log\left(\int_Z \frac{p(Z|X, \Theta)}{p(Z|X, \Theta^{(g)})} p(Z|X, \Theta^{(g)}) dZ\right) \\ &= 0 \quad \because \phi = -\log \text{ a convex unction} \end{aligned} \tag{26}$$

5 E-M Example: Gaussian Mixture Model

Gaussian Mixture Model (k-mixture) (GMM):

$$p(X|\Theta) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) \quad \sum_{l=1}^k \alpha_l = 1 \quad (27)$$

and $\Theta = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$

For data $X = \{x_1, \dots, x_n\}$ we introduce “latent” variable $Z = \{z_1, \dots, z_n\}$, each z_i indicates which mixture component x_i belong to.

Looking at the E-M algorithm:

$$\Theta^{(g+1)} = \arg \max_{\Theta} \left[q(\Theta, \Theta^{(g)}) \right] = \arg \max_{\Theta} \left(\int_z \log(p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) dz \right) \quad (28)$$

All we need to do is to define both $p(X, Z|\Theta)$ and $p(Z|X, \Theta)$

5.1 Gaussian Mixture Model in action

$$p(X|\Theta) = \sum_{l=1}^k \alpha_l \mathcal{N}(X|\mu_l, \Sigma_l) = \prod_{i=1}^n \sum_{l=1}^k \alpha_l \mathcal{N}(x_i|\mu_l, \Sigma_l) \quad (29)$$

How to define $p(X, Z|\Theta)$

$$p(X, Z|\Theta) = \prod_{i=1}^n p(x_i, z_i|\Theta) = \prod_{i=1}^n \underbrace{p(x_i|z_i, \Theta)}_{\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})} \underbrace{p(z_i|\Theta)}_{\alpha_{z_i}} = \prod_{i=1}^n \alpha_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i}) \quad (30)$$

Notice that $p(X, Z|\Theta)$ is actually simple than $p(X|\Theta)$.

How to define $p(Z|X, \Theta)$

$$p(Z|X, \Theta) = \prod_{i=1}^n p(z_i|x_i, \Theta) = \prod_{i=1}^n \frac{\alpha_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i})}{\sum_{l=1}^k \alpha_l \mathcal{N}(x_i|\mu_l, \Sigma_l)} \quad (31)$$

5.2 The E-Step:

$$\begin{aligned} q(\Theta, \Theta^{(g)}) &= \int_z \log(p(X, Z|\Theta)) p(Z|X, \Theta^{(g)}) dz \\ &= \int_{z_1} \dots \int_{z_n} \left(\sum_{i=1}^n \log p(z_i, x_i|\Theta) \prod_{i=1}^n p(z_i|x_i, \Theta^{(g)}) \right) dz_1, \dots, dz_n \end{aligned} \quad (32)$$

5.3 Some derivation to help

let $p(Y)$ be the joint pdf: $P(y_1, \dots, y_n)$, also let $F(Y)$ be a linear function, where each term involves only one variable y_i , i.e.,

$$F(Y) = f_1(x_1) + \dots + f_n(x_n) = \sum_{i=1}^n f_i(y_i) \quad (33)$$

then,

$$\int_{y_1} \dots \int_{y_n} \left(\sum_{i=1}^n f_i(y_i) \right) P(Y) dY = \sum_i \left(\int_{y_i} f_i(y_i) P_i(y_i) dy_i \right) \quad (34)$$

5.3.1 Proof

$$\int_Y F(Y) p(Y) dY = \int_{y_1} \int_{y_2} \dots \int_{y_N} \left(\sum_{i=1}^N f_i(y_i) \right) p(y_1, \dots, y_N) dy_1, \dots, dy_N \quad (35)$$

Expand it out, this equation has N sum terms. The first term is:

$$\begin{aligned} &= \int_{y_1} \int_{y_2} \dots \int_{y_N} f_1(y_1) p(y_1, \dots, y_N) \prod_{i=1}^N (dy_i) + \dots + \int_{y_1} \int_{y_2} \dots \int_{y_N} f_N(y_N) p(y_1, \dots, y_N) \prod_{i=1}^N (dy_i) \\ &= \int_{y_1} f_1(y_1) dy_1 \left(\int_{y_2} \dots \int_{y_N} p(y_1, \dots, y_N) \prod_{i=2}^N (dy_i) \right) + \dots + \int_{y_N} f_N(y_N) dy_N \int_{y_1} \dots \int_{y_{N-1}} p(y_1, \dots, y_N) \prod_{i=1}^{N-1} (dy_i) \end{aligned} \quad (36)$$

inside the first big bracket becomes the marginal probability density of $p(y_1)$, therefore, the first term becomes:

$$\int_{y_1} f_1(y_1) p(y_1) dy_1 \quad (37)$$

Apply this to each of the N terms, therefore:

$$\int_Y (F(Y)) P(Y) dY = \int_{y_1} f_1(y_1) P_1(y_1) dy_1 + \dots + \int_{y_n} f_n(y_n) P_n(y_n) dy_n \quad (38)$$

now apply Eq.(??), we have:

$$\begin{aligned}
q(\Theta, \Theta^{(g)}) &= \int_{z_1} \cdots \int_{z_n} \left(\sum_{i=1}^n \log p(z_i, x_i | \Theta) \prod_{i=1}^n p(z_i | x_i, \Theta^{(g)}) \right) dz_1, \dots, dz_n \\
&= \sum_{i=1}^n \left(\int_{z_i} \log p(z_i, x_i | \Theta) p(z_i | x_i, \Theta^{(g)}) dz_i \right) \quad z_i \in \{1, \dots, k\} \\
&= \sum_{z_i=1}^k \sum_{i=1}^n \log p(z_i, x_i | \Theta) p(z_i | x_i, \Theta^{(g)}) \quad \text{swap the summation terms} \\
&= \sum_{l=1}^k \sum_{i=1}^n \log[\alpha_l \mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \quad \text{substitute Gaussian and replace } z_i \rightarrow l
\end{aligned} \tag{39}$$

5.4 The M-Step objective function

$$\begin{aligned}
q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \log[\alpha_l \mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)}) \\
&= \sum_{l=1}^k \sum_{i=1}^n \log(\alpha_l) p(l | x_i, \Theta^{(g)}) + \sum_{l=1}^k \sum_{i=1}^n \log[\mathcal{N}(x_i | \mu_l, \Sigma_l)] p(l | x_i, \Theta^{(g)})
\end{aligned} \tag{40}$$

5.4.1 computing responsibilities

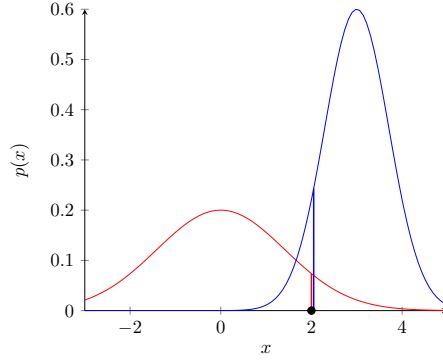


Figure 3: compute responsibility probabilities, red and blue each indicate its own un-normalized responsibilities

$$p(l | x_i, \Theta^{(g)}) = \frac{\mathcal{N}(\mathbf{x}_i; \mu_l, \Sigma_l)}{\sum_{s=1}^k \mathcal{N}(\mathbf{x}_i; \mu_s, \Sigma_s)} \tag{41}$$

Eq.(??) shows that the first term contains only α and the second term contains only μ, Σ . So we can maximize both terms independently.

5.5 The M-Step: maximizing α

Maximizing α means that:

$$\frac{\partial \sum_{l=1}^k \sum_{i=1}^n \log(\alpha_l) p(l|x_i, \Theta^{(g)})}{\partial \alpha_1, \dots, \partial \alpha_k} = [0 \dots 0] \quad \text{subject to } \sum_{l=1}^k \alpha_l = 1 \quad (42)$$

This is to be solved using Lagrange Multiplier

$$\mathbb{LM}(\alpha_1, \dots, \alpha_k, \lambda) = \sum_{l=1}^k \log(\alpha_l) \underbrace{\left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right)}_{\text{contains no } \alpha} + \lambda \left(\sum_{l=1}^k \alpha_l - 1 \right) \quad (43)$$

taking derivative with respect to just one α_l :

$$\begin{aligned} \frac{\partial \mathbb{LM}}{\partial \alpha_l} &= \frac{1}{\alpha_l} \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) + \lambda = 0 \\ \implies -\lambda &= \frac{1}{\alpha_l} \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) \\ \implies -\lambda \alpha_l &= \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) \\ \implies -\lambda \sum_{l=1}^k \alpha_l &= \sum_{l=1}^k \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) \\ \implies \lambda &= -n \end{aligned} \quad (44)$$

substitute $\lambda = -n$ into Eq.(??):

$$\begin{aligned} \frac{1}{\alpha_l} \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) + \lambda &= 0 \\ \implies \frac{1}{\alpha_l} \left(\sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \right) - n &= 0 \\ \implies \alpha_l &= \frac{1}{n} \sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \end{aligned} \quad (45)$$

5.6 Optional The M-Step: maximizing μ, Σ

Maximizing μ, Σ means that:

$$\frac{\partial \sum_{l=1}^k \sum_{i=1}^n \log(\alpha_l) p(l|x_i, \Theta^{(g)})}{\partial \mu_1, \dots, \partial \mu_k, \partial \Sigma_1, \dots, \partial \Sigma_k} = [0 \dots 0] \quad (46)$$

- You will need some linear algebra identities to solve this. It's quite involved. For details, please refer:
- J. Bilmes. "A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models"

5.6.1 Some formulas to remember

- derivatives of log of determinant (**with** determinant)

$$\frac{\partial \log |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top \quad (47)$$

- Derivatives of Traces

$$\frac{\partial \text{Tr}(F(\mathbf{X}))}{\partial \mathbf{X}} = (f(\mathbf{X}))^\top \quad (48)$$

where $f(\cdot)$ is the **scalar derivative** of $F(\cdot)$

- Derivatives of Traces of inverse, fact 1

$$\frac{\partial \text{Tr}(\mathbf{A}\mathbf{X}\mathbf{B})}{\partial \mathbf{X}} = \mathbf{A}^\top \mathbf{B}^\top \quad (49)$$

- Derivatives of Traces of inverse, fact 2

$$\frac{\partial \text{Tr}((\mathbf{X} + \mathbf{A})^{-1})}{\partial \mathbf{X}} = -((\mathbf{X} + \mathbf{A})^{-1}(\mathbf{X} + \mathbf{A})^{-1})^\top \quad (50)$$

- Derivatives of Traces of inverse, fact 3

$$\frac{\partial \text{Tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})}{\partial \mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1})^\top \quad (51)$$

5.6.2 Maximization μ_l

$$\begin{aligned}
\text{second part of } q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \log[\mathcal{N}(x_i|\mu_l, \Sigma_l)]p(l|x_i, \Theta^{(g)}) \\
&= \sum_{i=1}^n \sum_{l=1}^k \log \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right) \right) p(l|x_i, \Theta^{(g)})
\end{aligned} \tag{52}$$

Let \mathbf{Y} be zero-meaned data matrix, where each **column** of \mathbf{Y} is $y_i - \mu_l$:

$$\mathcal{L} \equiv \mathcal{L}(p(\mathbf{Y}|\mathcal{K})) = -\frac{DN}{2} \log(2\pi) - \frac{D}{2} \log |\mathcal{K}| - \frac{1}{2} \text{Tr}(\mathcal{K}^{-1} \mathbf{Y} \mathbf{Y}^\top) \tag{53}$$

$$\begin{aligned}
\text{second part of } q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \log[\mathcal{N}(x_i|\mu_l, \Sigma_l)]p(l|x_i, \Theta^{(g)}) \\
&= \sum_{i=1}^n \sum_{l=1}^k \log \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right) \right) p(l|x_i, \Theta^{(g)}) \\
\Rightarrow \mathcal{S}(\mu_l, \Sigma_l) &= \sum_{i=1}^n -\frac{1}{2} \log(|\Sigma_l|) p(l|x_i, \Theta^{(g)}) - \sum_{i=1}^n \frac{1}{2} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) p(l|x_i, \Theta^{(g)})
\end{aligned} \tag{54}$$

$$\begin{aligned}
\Rightarrow \mathcal{S}(\mu_l, \Sigma_l^{-1}) &= -\text{Tr} \left(\frac{\Sigma_l^{-1}}{2} \sum_{i=1}^n (x_i - \mu_l)(x_i - \mu_l)^\top p(l|x_i, \Theta^{(g)}) \right) + \text{Constant} \\
\Rightarrow \frac{\partial \mathcal{S}(\mu_l, \Sigma_l^{-1})}{\partial \mu_l} &= \frac{\Sigma_l^{-1}}{2} \sum_{i=1}^n 2(x_i - \mu_l) p(l|x_i, \Theta^{(g)}) = 0 \\
\Rightarrow \sum_{i=1}^n x_i p(l|x_i, \Theta^{(g)}) &= \mu_l \sum_{i=1}^n p(l|x_i, \Theta^{(g)}) \\
\Rightarrow \mu_l &= \frac{\sum_{i=1}^n x_i p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})}
\end{aligned} \tag{55}$$

5.6.3 Maximization Σ_l

$$\begin{aligned}
\text{second part of } q(\Theta, \Theta^{(g)}) &= \sum_{l=1}^k \sum_{i=1}^n \log[\mathcal{N}(x_i|\mu_l, \Sigma_l)]p(l|x_i, \Theta^{(g)}) \\
&= \sum_{i=1}^n \sum_{l=1}^k \log \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma_l|}} \exp \left(-\frac{1}{2} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right) \right) p(l|x_i, \Theta^{(g)})
\end{aligned} \tag{56}$$

- let \mathbf{Y} be zero-meaned data matrix, where each **column** of \mathbf{Y} is $x_i - \mu_l$
- let \mathbf{P} be diagonal matrix in which \mathbf{P}_{ii} correspond to $p(l|x_i, \Theta^{(g)})$

$$\mathcal{L} \equiv \mathcal{L}(p(\mathbf{Y}|\mu_l, \Sigma_l)) = -\frac{d \times \text{Tr}(\mathbf{P})}{2} \log(2\pi) - \frac{\text{Tr}(\mathbf{P})}{2} \log |\Sigma_l| - \frac{1}{2} \text{Tr}(\Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top) \quad (57)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Sigma_l} &= \Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} - \text{Tr}(\mathbf{P}) \Sigma_l^{-1} = \mathbf{0} \\ \implies \Sigma_l^{-1} \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} &= \text{Tr}(\mathbf{P}) \Sigma_l^{-1} \\ \implies \mathbf{Y} \mathbf{P} \mathbf{Y}^\top \Sigma_l^{-1} &= \text{Tr}(\mathbf{P}) \implies \Sigma_l^{-1} = \text{Tr}(\mathbf{P}) (\mathbf{Y} \mathbf{P} \mathbf{Y}^\top)^{-1} \\ \implies \Sigma_l &= \text{Tr}(\mathbf{P})^{-1} (\mathbf{Y} \mathbf{P} \mathbf{Y}^\top) = \frac{(\mathbf{Y} \mathbf{P} \mathbf{Y}^\top)}{\text{Tr}(\mathbf{P})} \\ &= \frac{\sum_{i=1}^n (x_i - \mu_l)(x - \mu_l)^T p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})} \end{aligned} \quad (58)$$

$$\mathcal{S}(\mu_l, \Sigma_l^{-1}) = \sum_{i=1}^n \left(-\frac{1}{2} \log(|\Sigma_l|) - \frac{1}{2} (x_i - \mu_l)^T \Sigma_l^{-1} (x - \mu_l) \right) p(l|x_i, \Theta^{(g)}) \quad (59)$$

Change Σ to Σ^{-1} , this is so that after taking derivative of $\log(X)$, the result is in terms of X^{-1}

$$\begin{aligned} &= \left(\sum_{i=1}^n \log(|\Sigma_l^{-1}|) p(l|x_i, \Theta^{(g)}) - \frac{1}{2} \text{tr} \left(\Sigma^{-1} \underbrace{\sum_{i=1}^n (x_i - \mu_l)(x - \mu_l)^T p(l|x_i, \Theta^{(g)})}_{M_l} \right) \right) \\ \implies \frac{\partial \mathcal{S}(\mu_l, \Sigma_l^{-1})}{\partial \Sigma_l^{-1}} &= \frac{2 \sum_{i=1}^n \Sigma_l p(l|x_i, \Theta^{(g)}) - \sum_{i=1}^n \text{diag}(\Sigma) p(l|x_i, \Theta^{(g)})}{2} - \frac{2M_l - \text{diag}(M_l)}{2} = 0 \\ \implies 2 \left(\sum_{i=1}^n \Sigma p(l|x_i, \Theta^{(g)}) - M_l \right) - \sum_{i=1}^n \text{diag}(\Sigma p(l|x_i, \Theta^{(g)}) - M_l) &= 0 \\ \implies \sum_{i=1}^n \Sigma p(l|x_i, \Theta^{(g)}) - M_l &= 0 \\ \implies \Sigma &= \frac{\sum_{i=1}^n M_l}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})} = \frac{\sum_{i=1}^n (x_i - \mu_l)(x - \mu_l)^T p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^n p(l|x_i, \Theta^{(g)})} \end{aligned} \quad (60)$$

5.7 Summary of Gaussian Mixture Model

Maximizing μ, Σ means that to update $\Theta^{(g)} \rightarrow \Theta^{(g+1)}$:

$$\alpha_l^{(g+1)} = \frac{1}{N} \sum_{i=1}^N p(l|x_i, \Theta^{(g)}) \quad (61)$$

$$\mu_l^{(g+1)} = \frac{\sum_{i=1}^N x_i p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^N p(l|x_i, \Theta^{(g)})} \quad (62)$$

$$\Sigma_l^{(g+1)} = \frac{\sum_{i=1}^N [x_i - \mu_l^{(i+1)}][x_i - \mu_l^{(i+1)}]^\top p(l|x_i, \Theta^{(g)})}{\sum_{i=1}^N p(l|x_i, \Theta^{(g)})} \quad (63)$$

One can verify that when the number of component $K = 1 \implies p(l = 1|x_i, \Theta^{(g)}) = 1$, then:

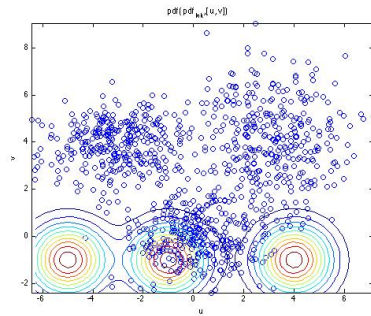
$$\alpha_l^{(g+1)} = 1 \quad (64)$$

$$\mu_l^{(g+1)} = \frac{1}{N} \sum_{i=1}^N x_i \quad (65)$$

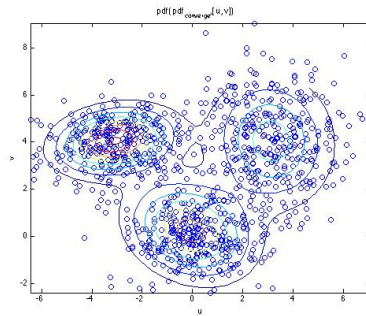
$$\Sigma_l^{(g+1)} = \frac{1}{N} \sum_{i=1}^N [x_i - \mu_l^{(i+1)}][x_i - \mu_l^{(i+1)}]^\top \quad (66)$$

which becomes the parameter update of a single Gaussian

5.8 To show the diagram again



(a) GMM at initialization: $\Theta^{(1)}$



(b) GMM at convergence: $\Theta^{(f)}$