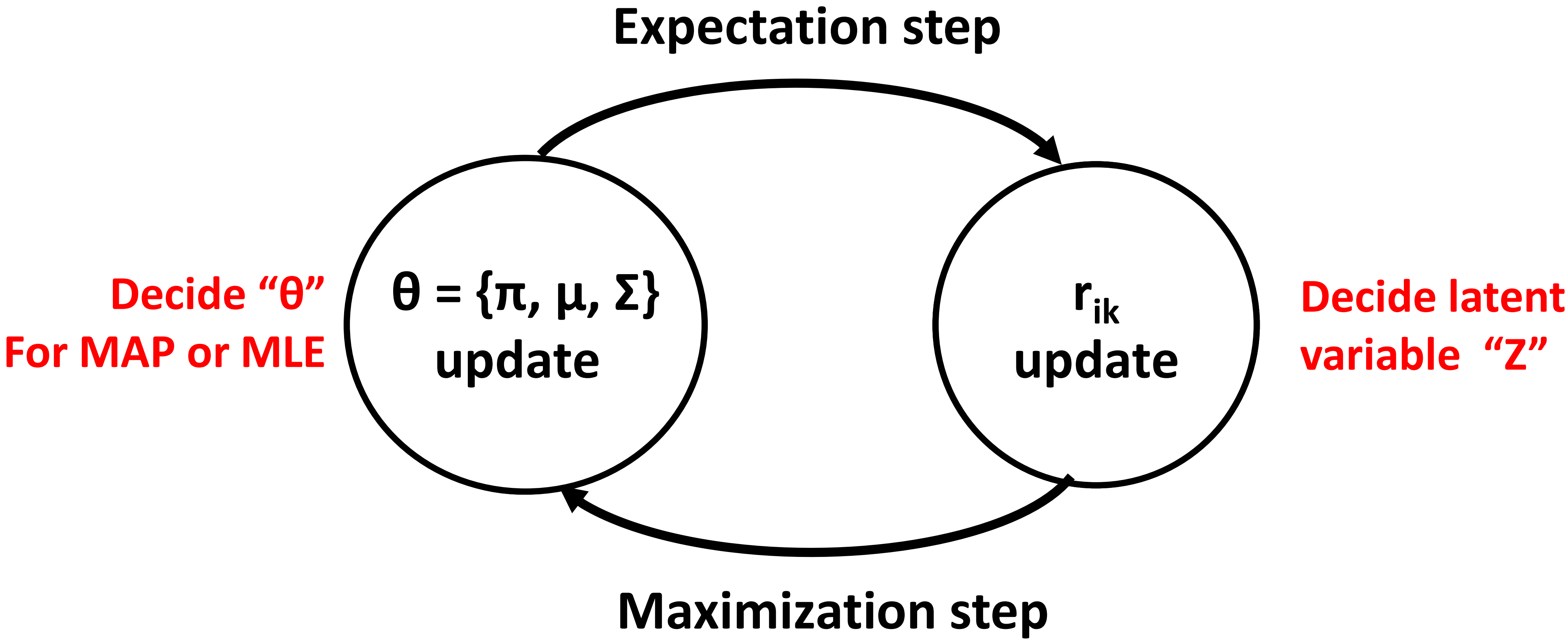
MLE in GMM Clustering

Hanwool Jeong

hwjeong@kw.ac.kr

Revisit EM in GMM Clustering

- Focus on M step!



$$\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$

EM for GMM Clustering; E step

- We already see this!
- Deriving r_{ik} = the posterior probability that point i belongs to cluster k.

$$
\begin{aligned}
r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= \frac{p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k', \boldsymbol{\theta})} \\
&= \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i|\boldsymbol{\theta}_{k'}^{(t-1)})}
\end{aligned}
$$

- The above term is called **responsibility**. How does look like?

EM for GMM Clustering; M step

- M step, first, which estimates θ or potential output based on the latent variables
- First, for π_k :

$$\pi_k \quad = \quad \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$

- Maximizing the expected complete data log likelihood defined as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) \quad \triangleq \quad \mathbb{E}\left[\sum_i \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta})\right] \quad = \quad \sum_i \mathbb{E}\left[\log\left[\prod_{k=1}^{K}(\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k))^{\mathbb{I}(z_i=k)}\right]\right]$$

$$= \quad \sum_i \sum_k \mathbb{E}\left[\mathbb{I}(z_i = k)\right] \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)]$$

$$= \quad \sum_i \sum_k p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)]$$

EM for GMM Clustering; M step

- That is, for GMM, the following should be maximized

$$\begin{aligned}\ell(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \sum_k \sum_i r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= -\frac{1}{2} \sum_i r_{ik} \left[\log|\boldsymbol{\Sigma}_k| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right]\end{aligned}$$

- And it can be easily proved with the above term is maximized when

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k} \\ \boldsymbol{\Sigma}_k &= \frac{\sum_i r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{r_k} = \frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\end{aligned}$$

GMM Clustering

- Pseudo-code is shown

Initialize θ

while(until converge)

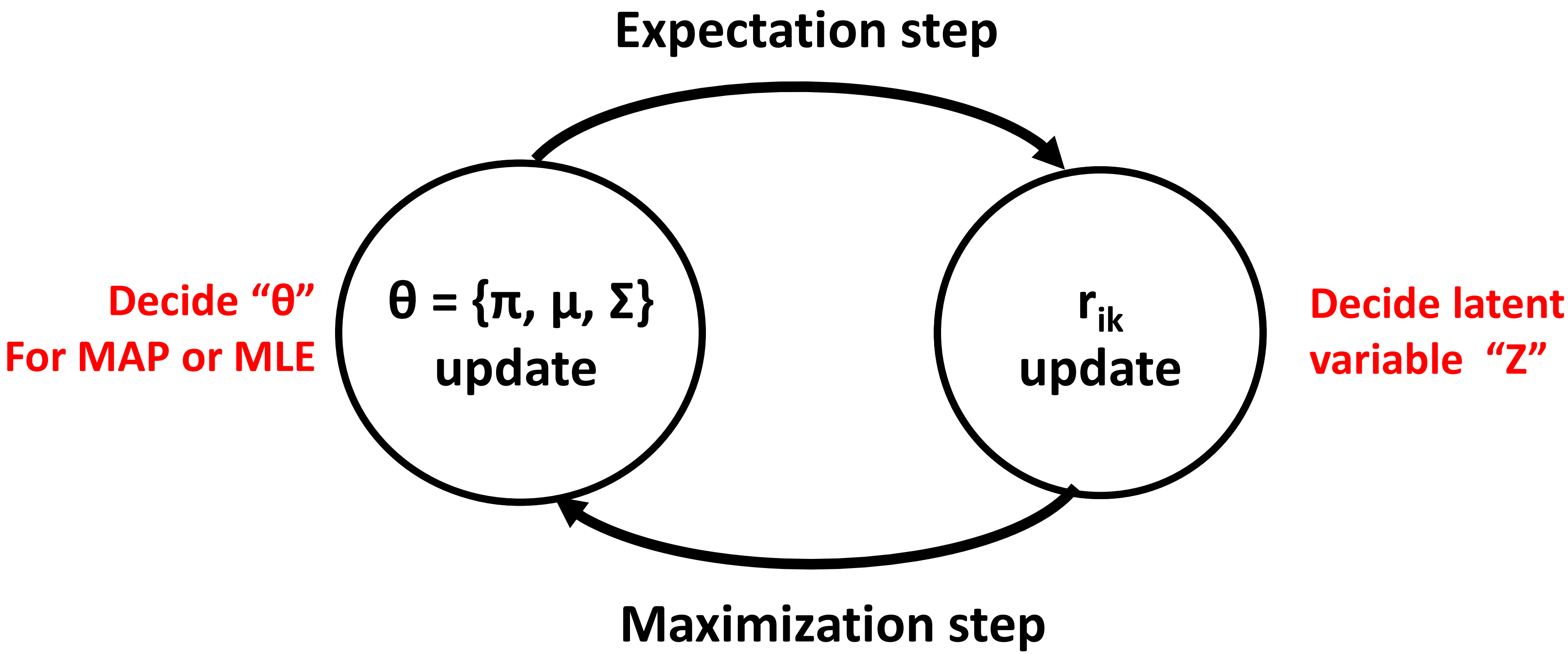
Estimate $\mathbf{r}_{ij}$ based on **θ**

Estimate **θ** based on $\mathbf{r}_{ij}$

$$\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$

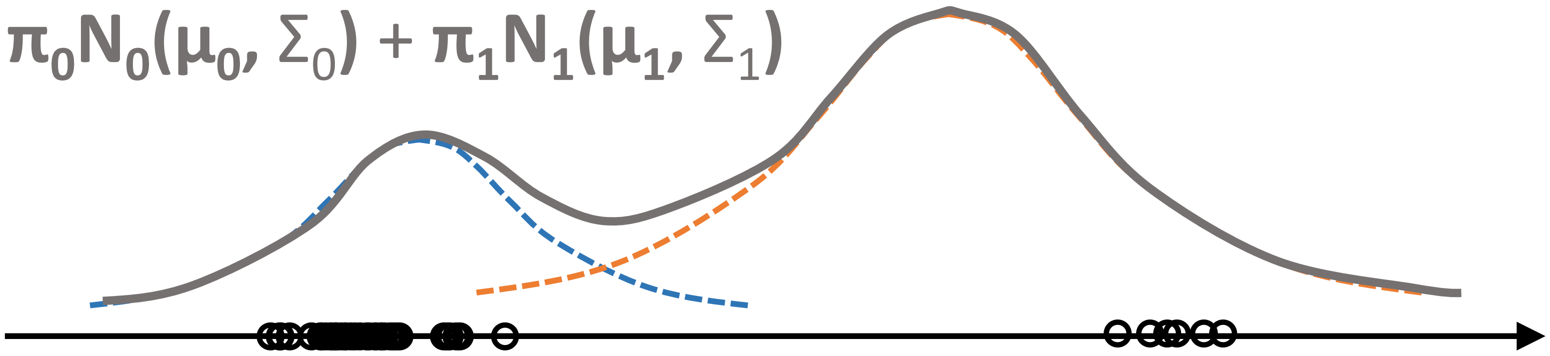
Imagine a Situation Where Something Went Wrong

- What do we have to determine?
- Extreme is better!
- Simplified situation is better!



You See Something Wrong

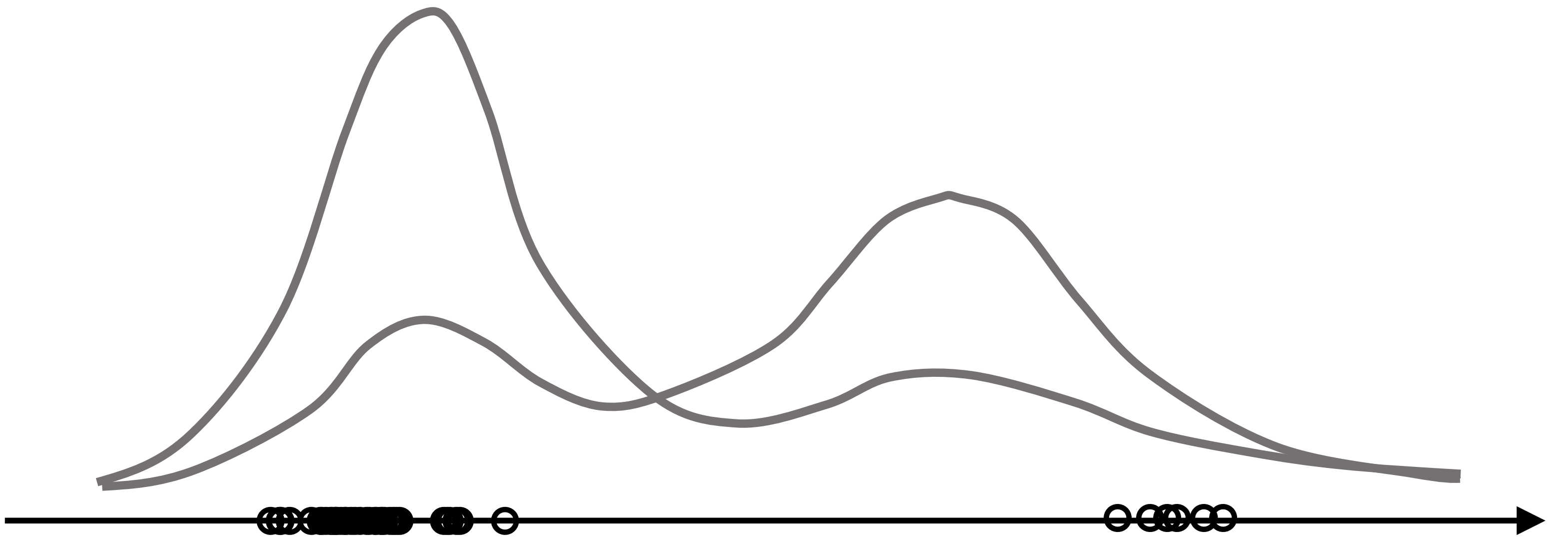
- $\boldsymbol{\theta} = \{\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}\}$, All should be updated properly
- Let's simplify the situation for better understanding, $\boldsymbol{\theta} = \{\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}\}$ is to be updated.



How Should We Update "Portion"?

- $\pi_0 = 0.9$ and $\pi_1 = 0.1$
- More smoothly, (softly,) $\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$
- Where the responsibility is

$$r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_i = k'|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k', \boldsymbol{\theta})}$$



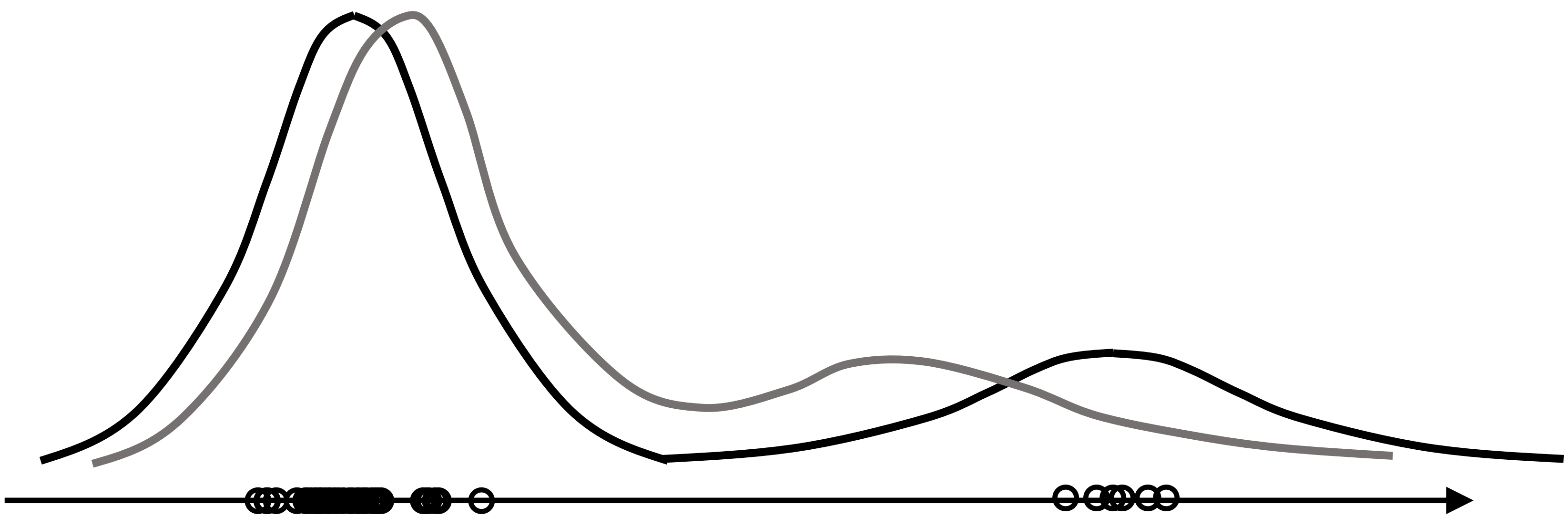
Now We Should Update Gaussian!

- Based on what? **MLE!**

Maximize $p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta})$ adjusting θ

- By how? **Formalize it and differentiate it!**

We can do it by focusing on each k



Focusing on k = 0

- That is, you should maximize what?
- For $\pi_0 N_0(\boldsymbol{\mu_0}, \Sigma_0)$, find a θ such that maximizes p(**x**/**θ**), that is,

π is already known

$$p_0(\mathbf{x}/\boldsymbol{\theta}) = \pi_0 \prod_i \frac{1}{\sqrt{2\pi}\, |\Sigma_0|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{\mu})\Sigma_0{}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})\right\}$$

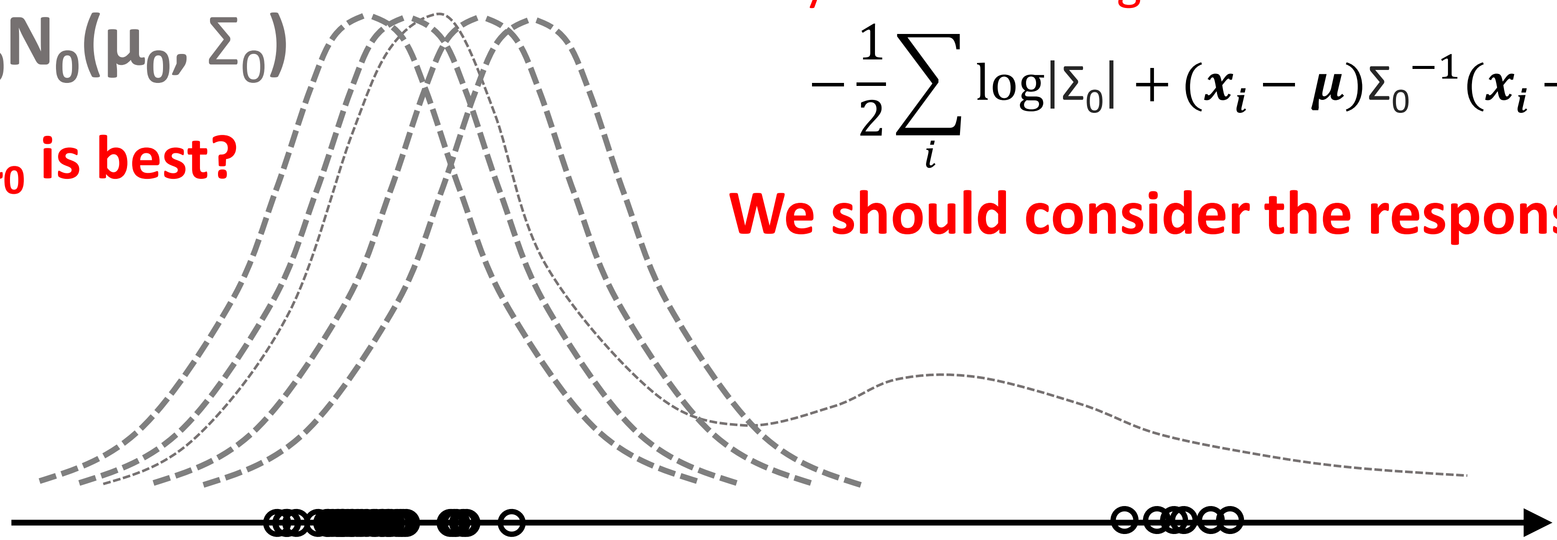
Can you find out log-likelihood term? Is it okay?

$$-\frac{1}{2}\sum_i \log|\Sigma_0| + (\boldsymbol{x_i} - \boldsymbol{\mu})\Sigma_0{}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})$$

We should consider the responsibility!

$\pi_0 N_0(\mu_0, \Sigma_0)$

Which μ_0 is best?



Thus,

- Instead of maximizing the below log likelihood

$$-\frac{1}{2}\sum_{i}\log|\Sigma_0| + (\boldsymbol{x_i} - \boldsymbol{\mu})\Sigma_0{}^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})$$

- But maximizing the “expected” log likelihood

$$\ell(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{1}{2}\sum_{i} r_{ik}\left[\log|\boldsymbol{\Sigma}_k| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right]$$

Revisit the Slide

- That is, for GMM, the following should be maximized

$$\begin{aligned}\ell(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \sum_k \sum_i r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= -\frac{1}{2} \sum_i r_{ik} \left[\log |\boldsymbol{\Sigma}_k| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right]\end{aligned}$$

- And it can be easily proved with the above term is maximized when

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{\sum_i r_{ik} \mathbf{x}_i}{r_k} \\ \boldsymbol{\Sigma}_k &= \frac{\sum_i r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{r_k} = \frac{\sum_i r_{ik} \mathbf{x}_i \mathbf{x}_i^T}{r_k} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\end{aligned}$$

What Did you Learn?

- What do we have to decide?
- Which criteria?
- How can you express “your thought” in mathematical term?