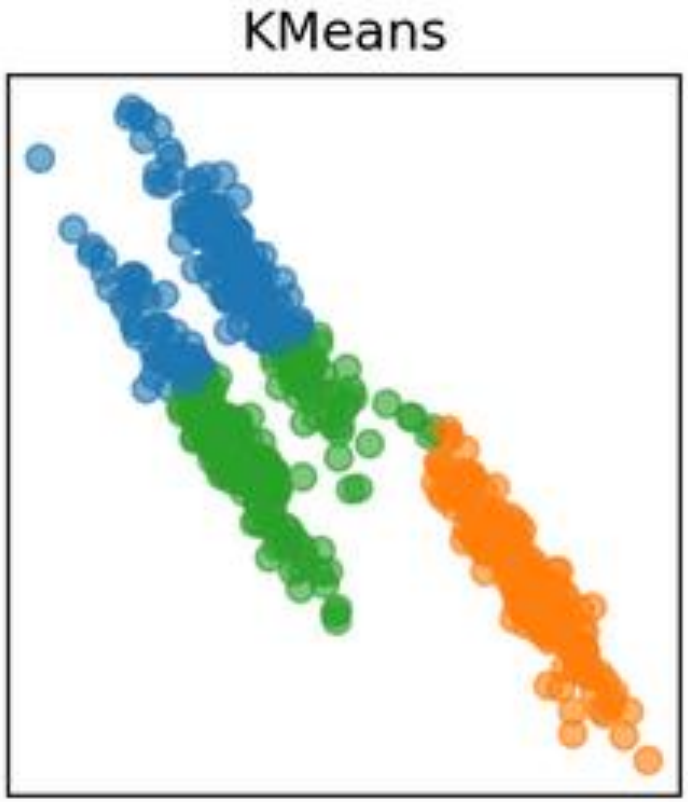
Gaussian Mixture Model
Clustering

Hanwool Jeong

hwjeong@kw.ac.kr

Limitation of K-Means



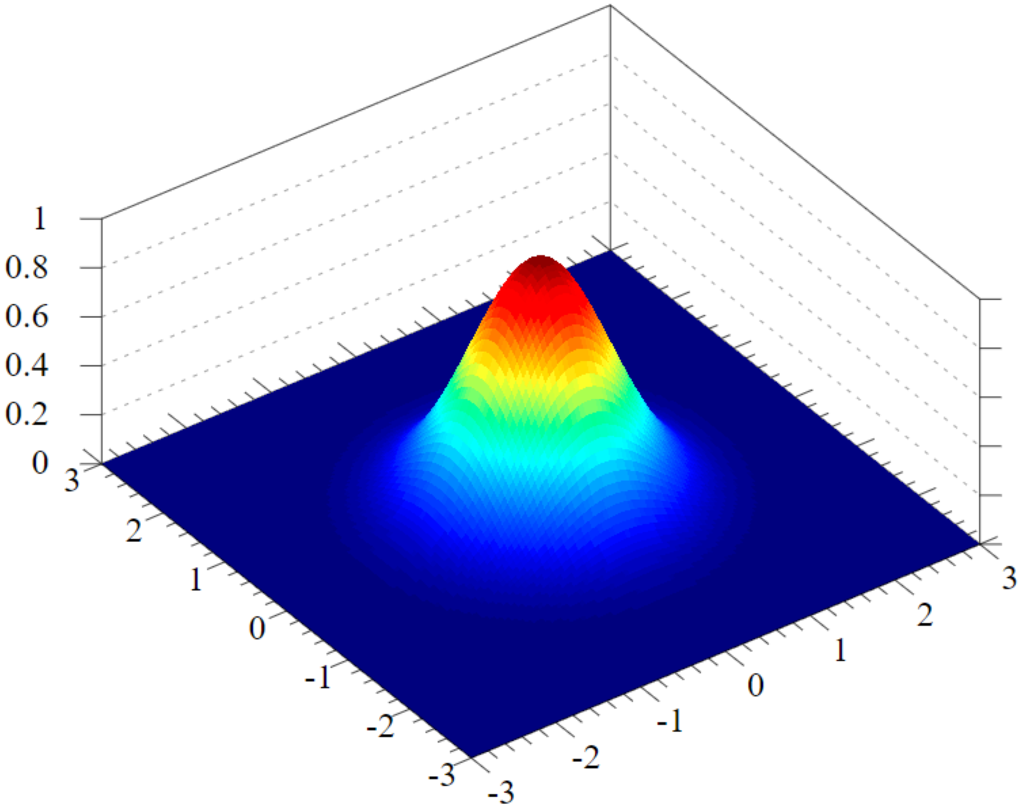
Categorial Distribution

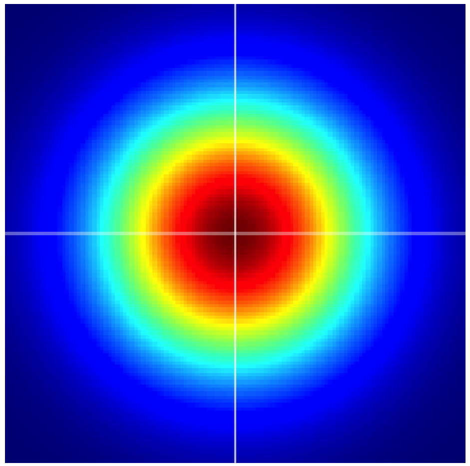
- Binomial distribution vs. Bernoulli distribution

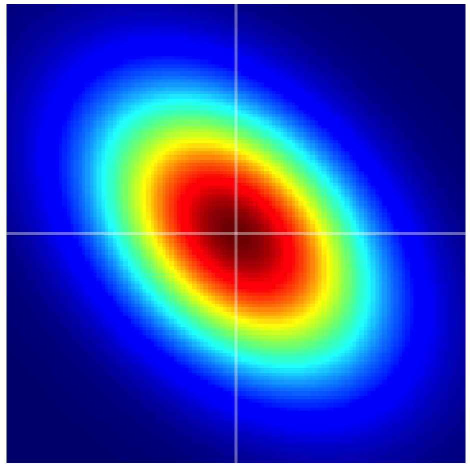
- Multinomial distribution vs. Categorial distribution

Multivariate(Joint) Gaussian Distribution

- Can be independent or correlated







How Can Embody the Correlation?

- There can be correlation between x_1 & x_2
- We will use "covariance" instead of "variance" matrix.

$$\begin{aligned}\operatorname{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] \\ &= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1, X_2] & \cdots & \operatorname{cov}[X_1, X_d] \\ \operatorname{cov}[X_2, X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d, X_1] & \operatorname{cov}[X_d, X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix}\end{aligned}$$

- Correlation coefficient :

$$\operatorname{corr}[X, Y] \triangleq \frac{\operatorname{cov}[X, Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}$$

Application of Mixture Model

- Can you figure out the meaning of latent variable?
- We typically use Z for latent variable.

The Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- $\mu = E[x] \in R^D$ is the mean vector, and $\Sigma = \mathrm{cov}[x]$ is the $D \times D$ covariance matrix.

Mixture Model

- The simplest form of latent variable model (LVM) is when $z_i \in \{1, . . . , K\}$, representing a discrete latent state.
- With prior $p(z_i) = \text{Cat}(\pi)$ and likelihood $p(x_i | z_i = k) = p_k(x_i)$, the overall model shown below is mixture **model**:

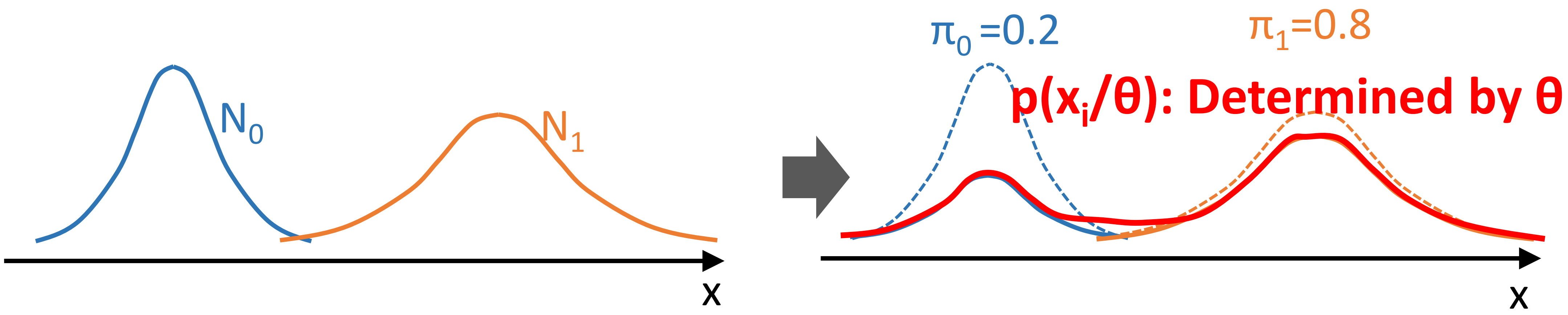
$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta})$$

Gaussian Mixture Model (GMM)

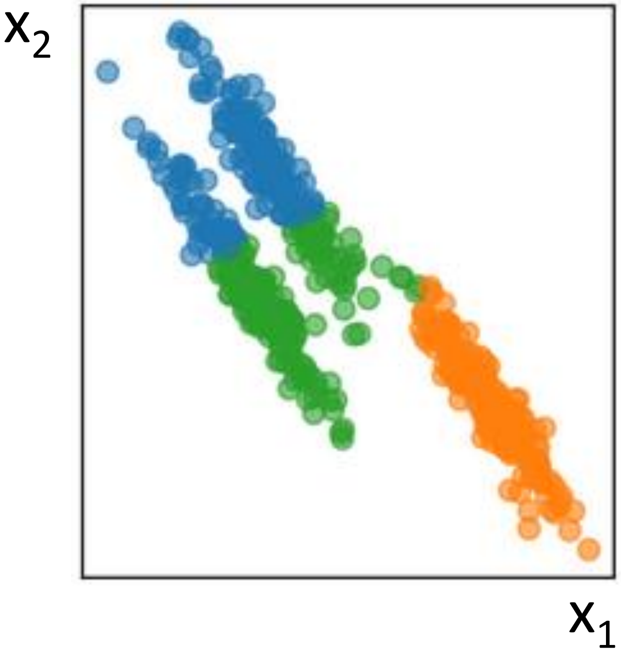
- The most widely used mixture model is the mixture of Gaussians (MOG), also called a Gaussian mixture model (GMM), in the form of:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

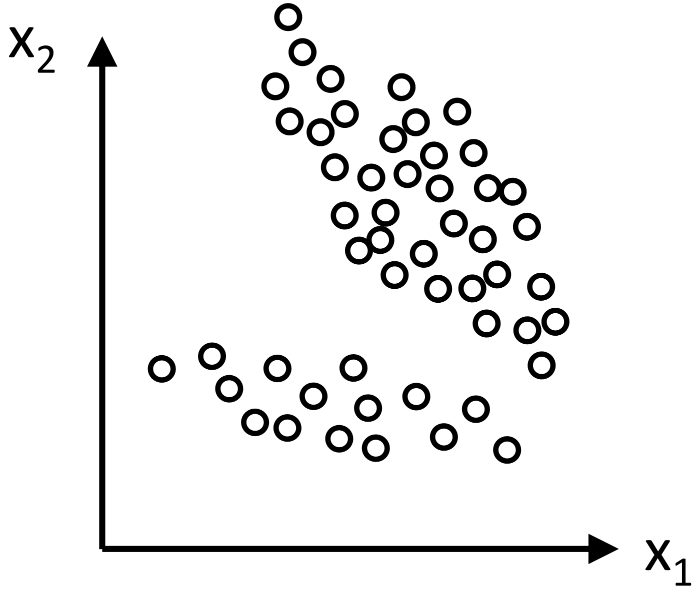
We can say $\theta = (\pi_1, \pi_2, \ldots$
$\mu_1, \mu_2, \ldots$
$\Sigma_1, \Sigma_2, \ldots)$



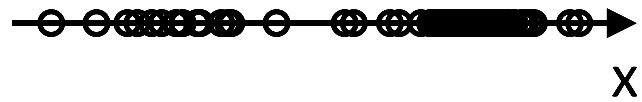
Now We Are Ready for Clustering Using GMM



or more simply,



or even more simply,



Latent Variable for GMM Clustering

- We say we need determine

$$\theta = (\pi_1, \pi_2, \ldots, \mu_1, \mu_2, \ldots, \Sigma_1, \Sigma_2, \ldots)$$

- To maximize

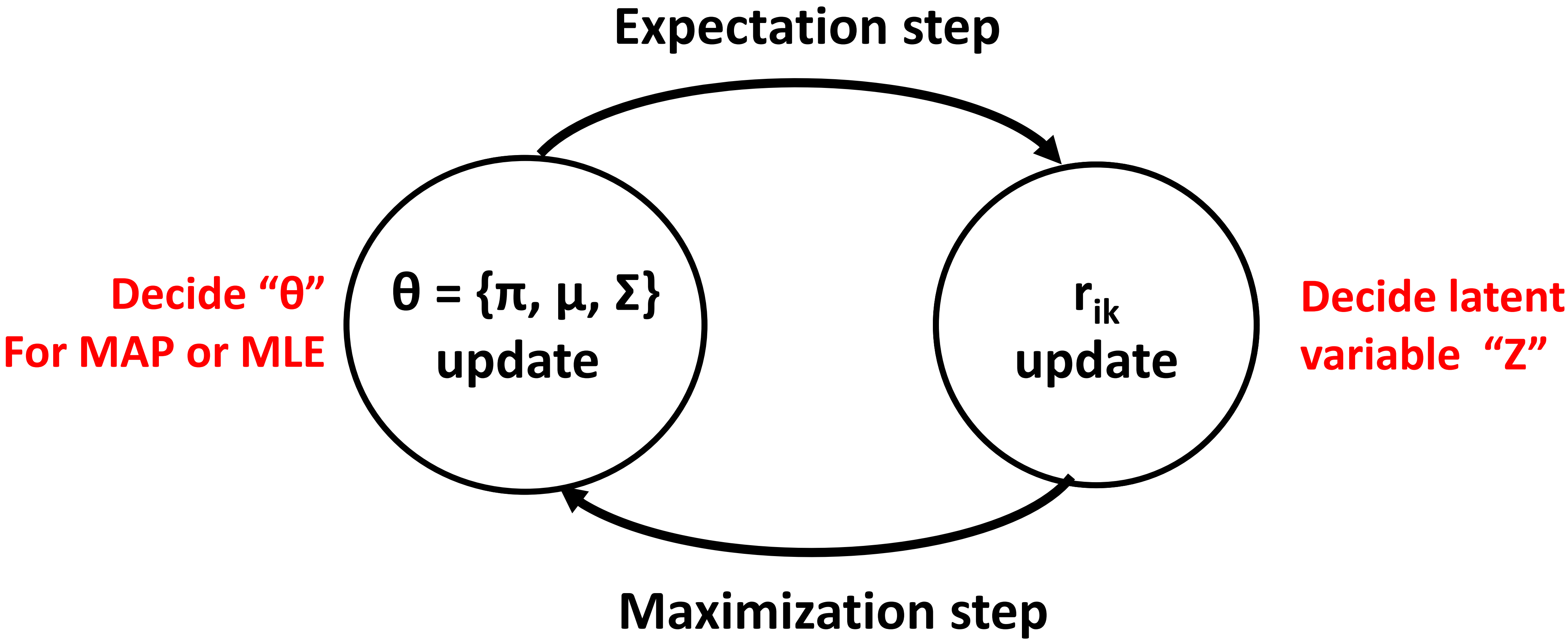
$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Then, what is the latent variable?
- We define the responsibility as the latent variable to be updated as follows,

$$r_{ik} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$$

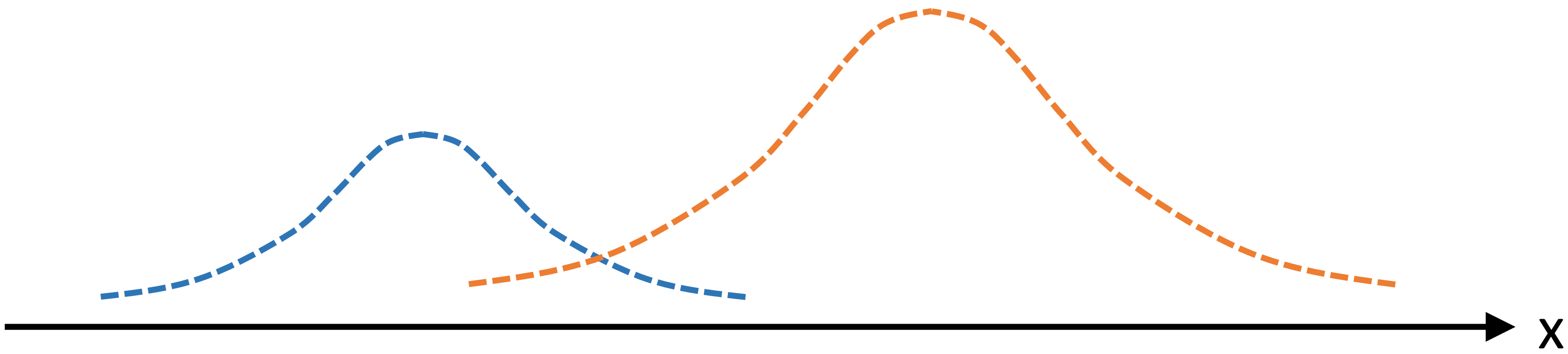
- which means the probability that x_i is clustered as k for given θ

Thus, EM in GMM Clustering is



Responsibility?

- Suppose that we have the mixture disturb determined by the sum of two distribution shown below
- Then, how r_{ik} is determined?



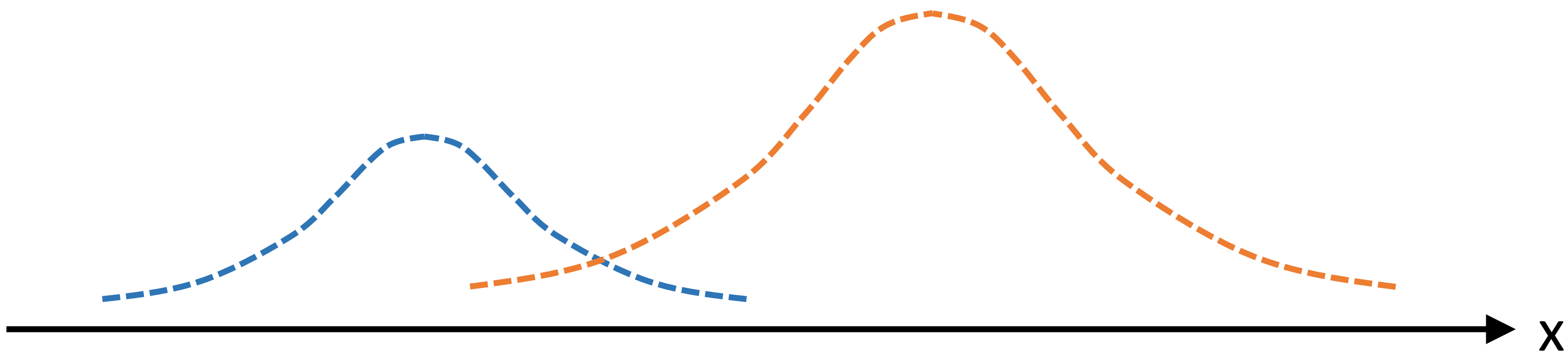
- Every time θ is updated, r_{ik} should be updated. (Intuitive)
- How does r_{ik} change affect θ? (We will see)

Calculation of r_{ik} for given θ
Mixture Model for Clustering

- Fit the mixture model (how?), then compute $p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$
  - $p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$ = Posterior probability that x_i belongs to cluster k.
- This responsibility of cluster k for x_i, and can be computed using Bayes rule as follows:

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

is **soft clustering**



Soft? There is Also Hard Clustering

- If you pick one cluster for x_i, as below,

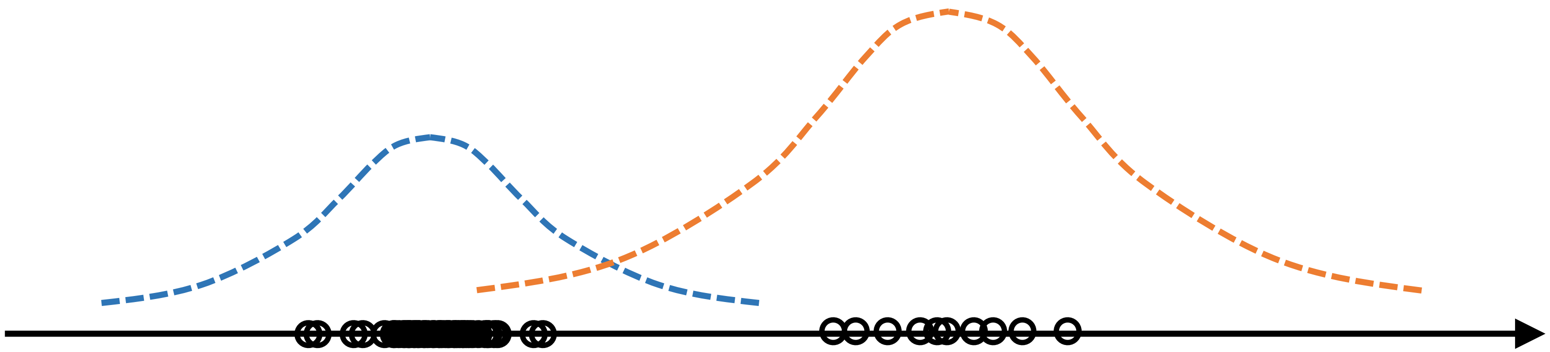
$$z_i^* = \arg\max_k r_{ik} = \arg\max_k \log p(\mathbf{x}_i | z_i = k, \boldsymbol{\theta}) + \log p(\mathbf{z}_i = k | \boldsymbol{\theta})$$

is **hard clustering**

Think How We Can "Refine" our θ

- To think how we can exploit r_{ik} to improve θ, imagine the situation θ is not well determined (extreme is better!)
- Derived r_{ik}, how can we improve θ
  (or it would be better if you think how we can improve θ without the assumption we have r_{ik}!)

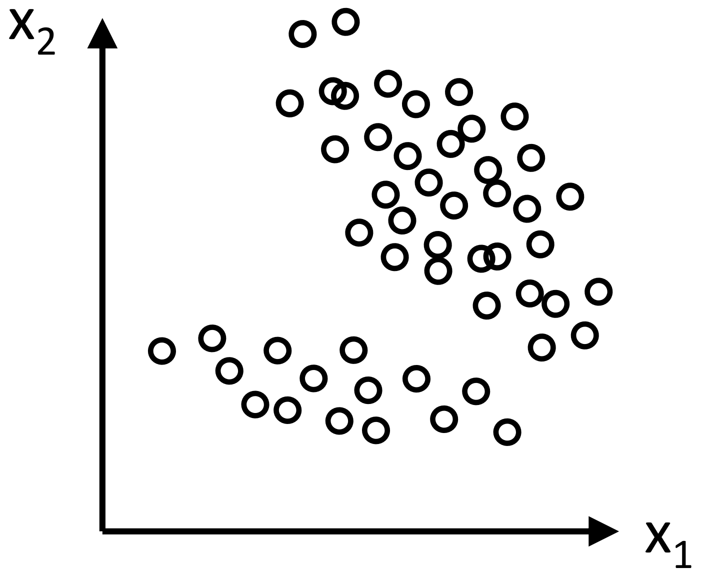
$$\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$



Hard Clustering using a GMM

- Formulating two Gaussian ➔ Mixture Gaussian
- With K=2, can you imagine how r_{i1} and r_{i2} would be derived for each x_i?

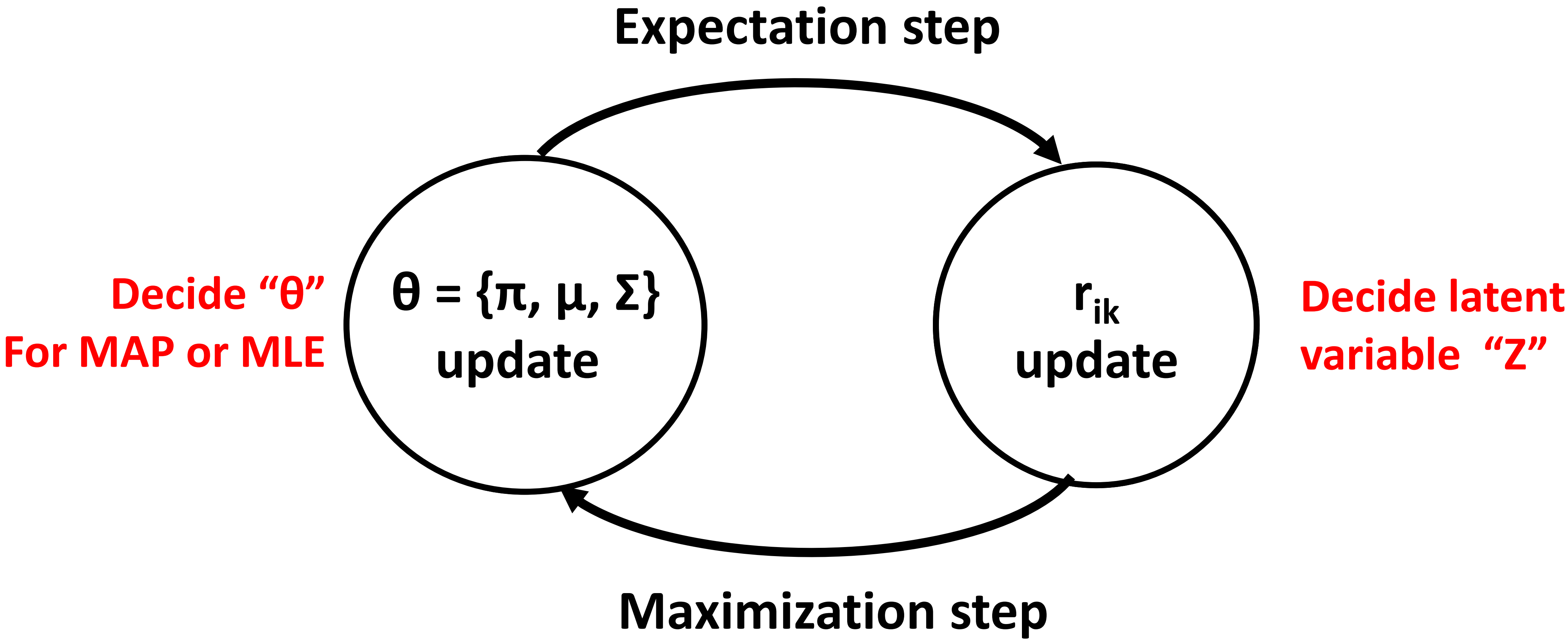
$$p(\mathbf{x_i}|\boldsymbol{\theta}^*) = (15/49)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + (34/49)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$$



$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

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.

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

$$= \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i|\boldsymbol{\theta}_{k'}^{(t-1)})}$$

- 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 :

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

- Maximizing the expected complete data log likelihood defined as

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

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

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

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

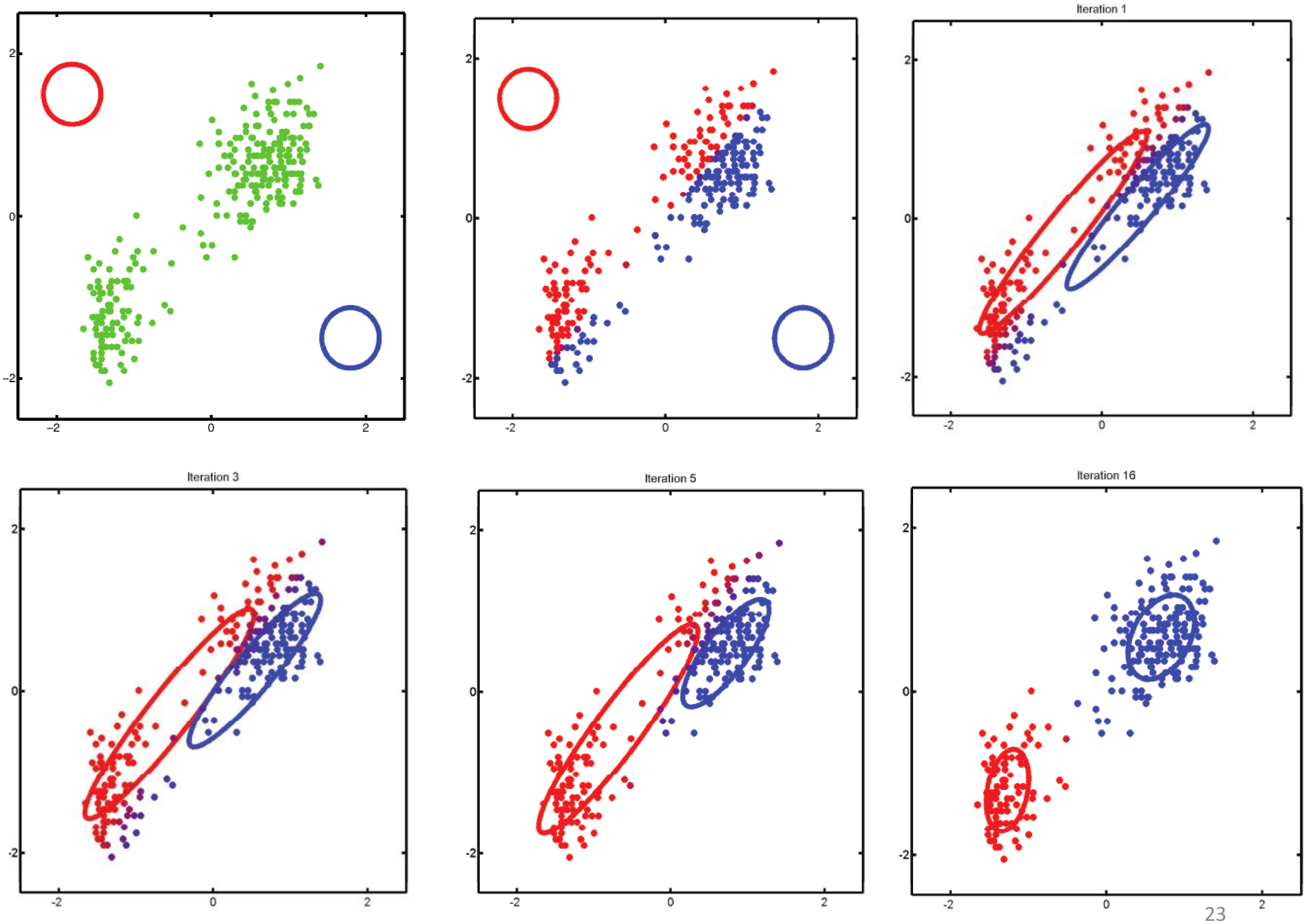
$$\pi_k = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$$

Initialize θ

while(until converge)

Estimate $\mathbf{r_{ij}}$ based on **θ**

Estimate **θ** based on $\mathbf{r_{ij}}$


Iteration 1
Iteration 3
Iteration 5
Iteration 16
2
0
-2
-2
0
2