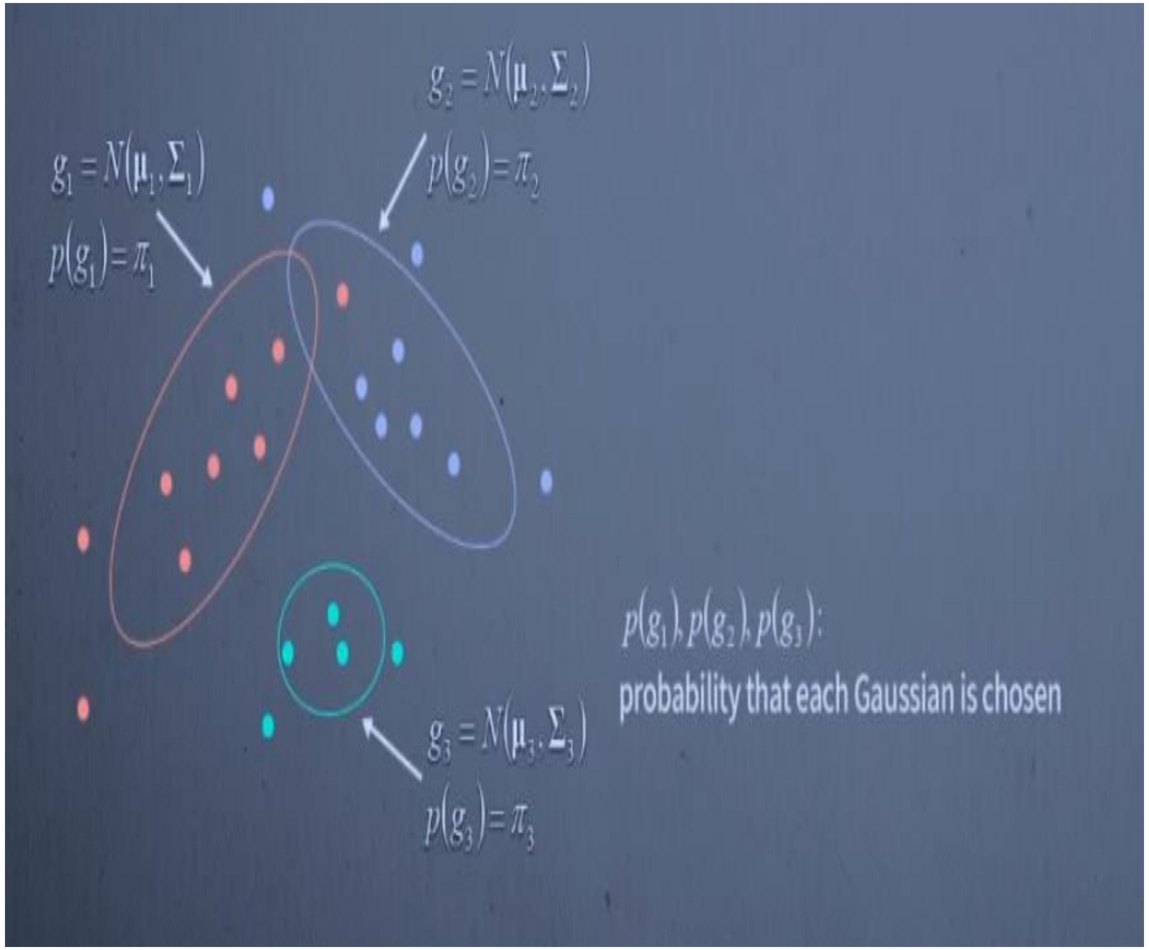
Gaussian Mixture Model

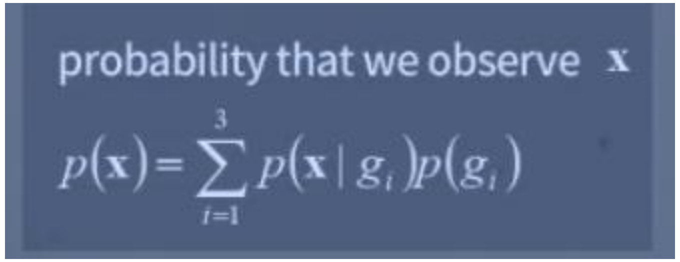
Gaussian Mixture Model, Clustering

- K-Means에 Soft한 Version(매우 비슷함)
- K-Means는 한 데이터가 한 Cluster에만 속할 수 있으나, Gaussian Mixture Model(GMM)에서는 데이터가 속할 확률을 결정한다.
- Expectation and Maximization Algorithm

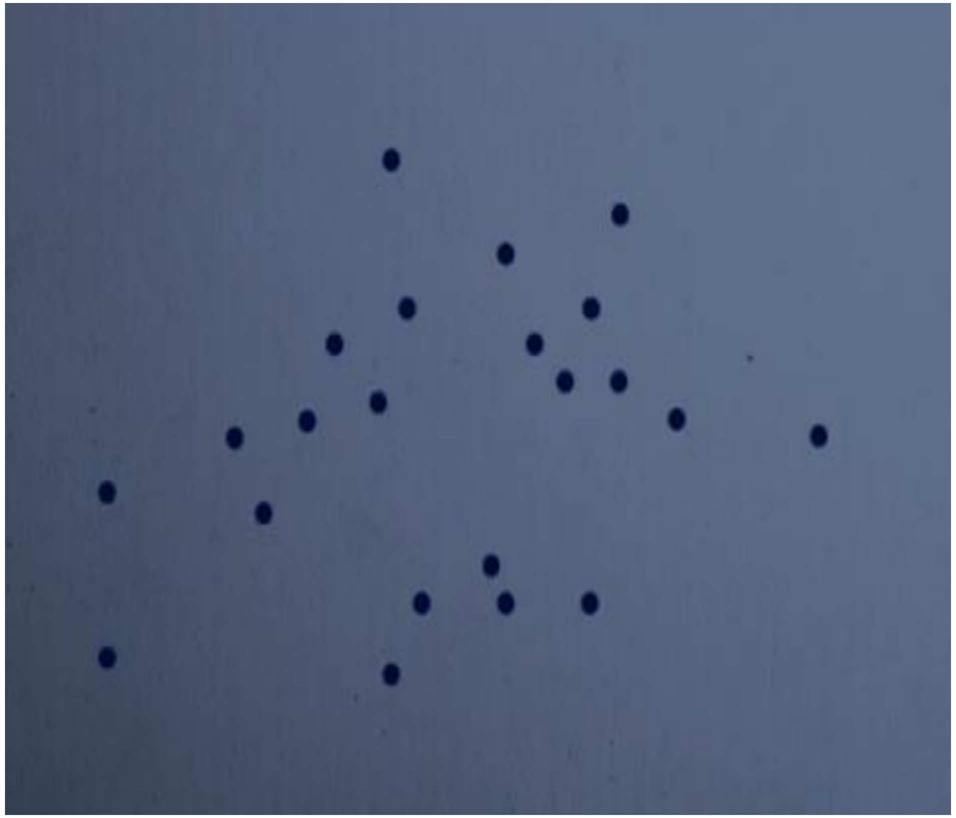
Mixture of Gaussians1



(가정)미리 Gaussian에 분포(개수)를 알고 있다.
각각의 평균과 분산 존재
세개의 Gaussian 중에서 Random하게 하나를 선택한다.
선택할 확률 : p(g1), p(g2), p(g3)
선택한 Gaussian에서 평균과 분산에 부합하도록 데이터를 생성한다.(점들)
데이터를 얻을 확률은 아래와 같다.

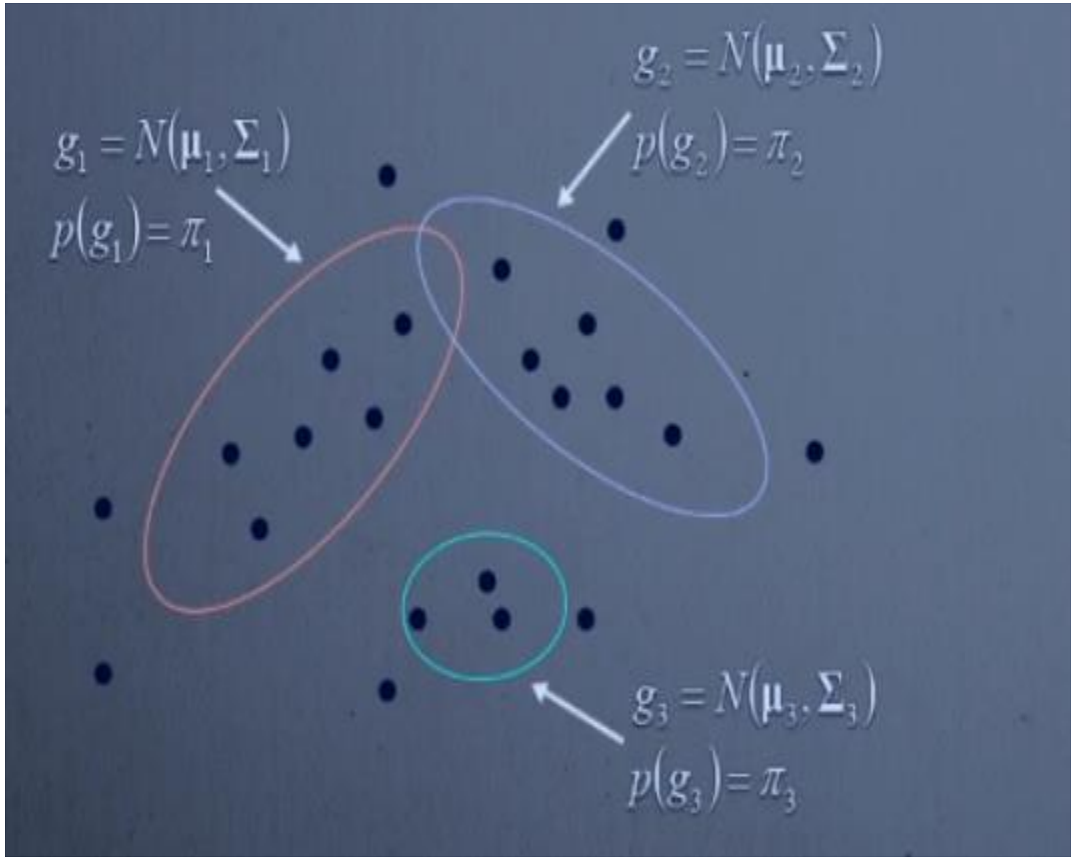


Mixture of Gaussians2



(가정)미리 Gaussian에 분포(개수)를 알고 있다.
각각의 평균과 분산을 모르는 상태
단지 Gaussian 개수만 알고, 평균과 분산 그리고 Gaussian이 선택될 확률을 모른다면 적용할 수 없다.
어느 Data가 어느 Gaussian에서 만들어졌는지 알 수 없다.

Mixture of Gaussians3



(가정)미리 Gaussian에 분포(개수)를 알고 있다.
각각의 평균과 분산을 알고, Gaussian이 선택될 확률을 알고 있다.
어느 Data가 어느 Gaussian에서 만들어졌는지
찾는 방법
Zki : k라는 데이터를 봤을 때, k가 i번째 Gaussian에 의해 생성됐을 확률

$$z_{ki} = p(g_i \mid \mathbf{x}_k) = \frac{p(\mathbf{x}_k \mid g_i)p(g_i)}{p(\mathbf{x}_k)} = \frac{p(\mathbf{x}_k \mid g_i)p(g_i)}{\sum_{j=1}^{3} p(\mathbf{x}_k \mid g_j)p(g_j)} = \frac{p(\mathbf{x}_k \mid g_i)\pi_i}{\sum_{j=1}^{3} p(\mathbf{x}_k \mid g_j)\pi_j}$$

Mixture of Gaussians4

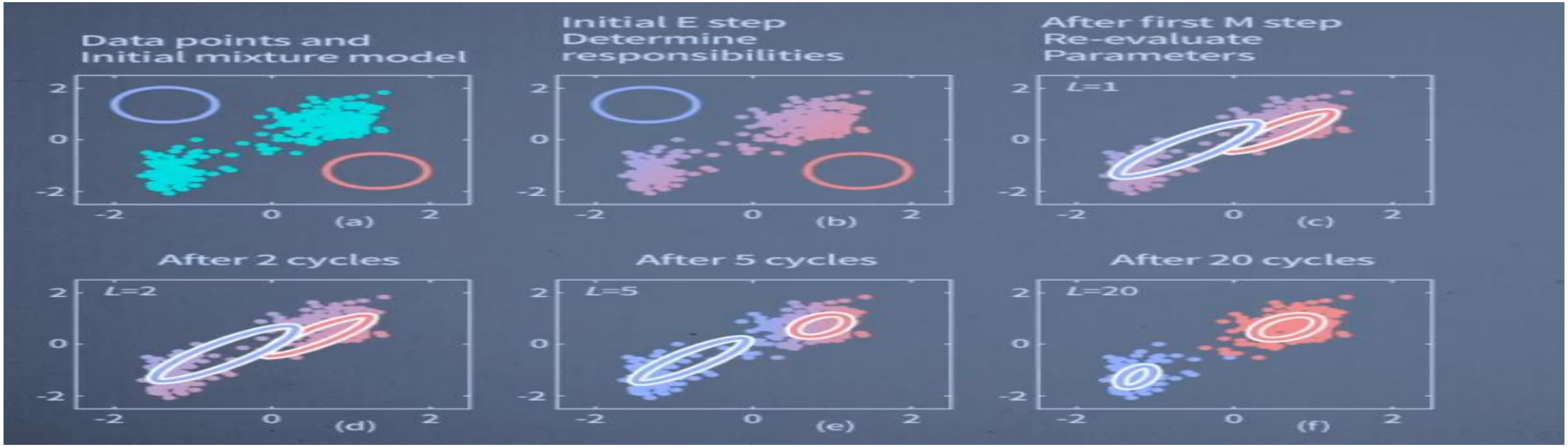
(가정)미리 Gaussian에 분포(개수)를 알고 있다.
Gaussian이 선택될 확률을 모르고, 평균 분산을 모르는 상태이다. 대신 어느 Data가 어느 Gaussian에서 만들어졌는지 알고 있다.
평균과 분산 Gaussian이 선택될 확률 찾기
평균과 분산 : 같은 점 색 점들끼리 구하면 됨
Gaussian 선택될 확률 : 같은색점 개수/전체 점들 개수

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)\cdot \mathbf{x}_j}{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)}$$

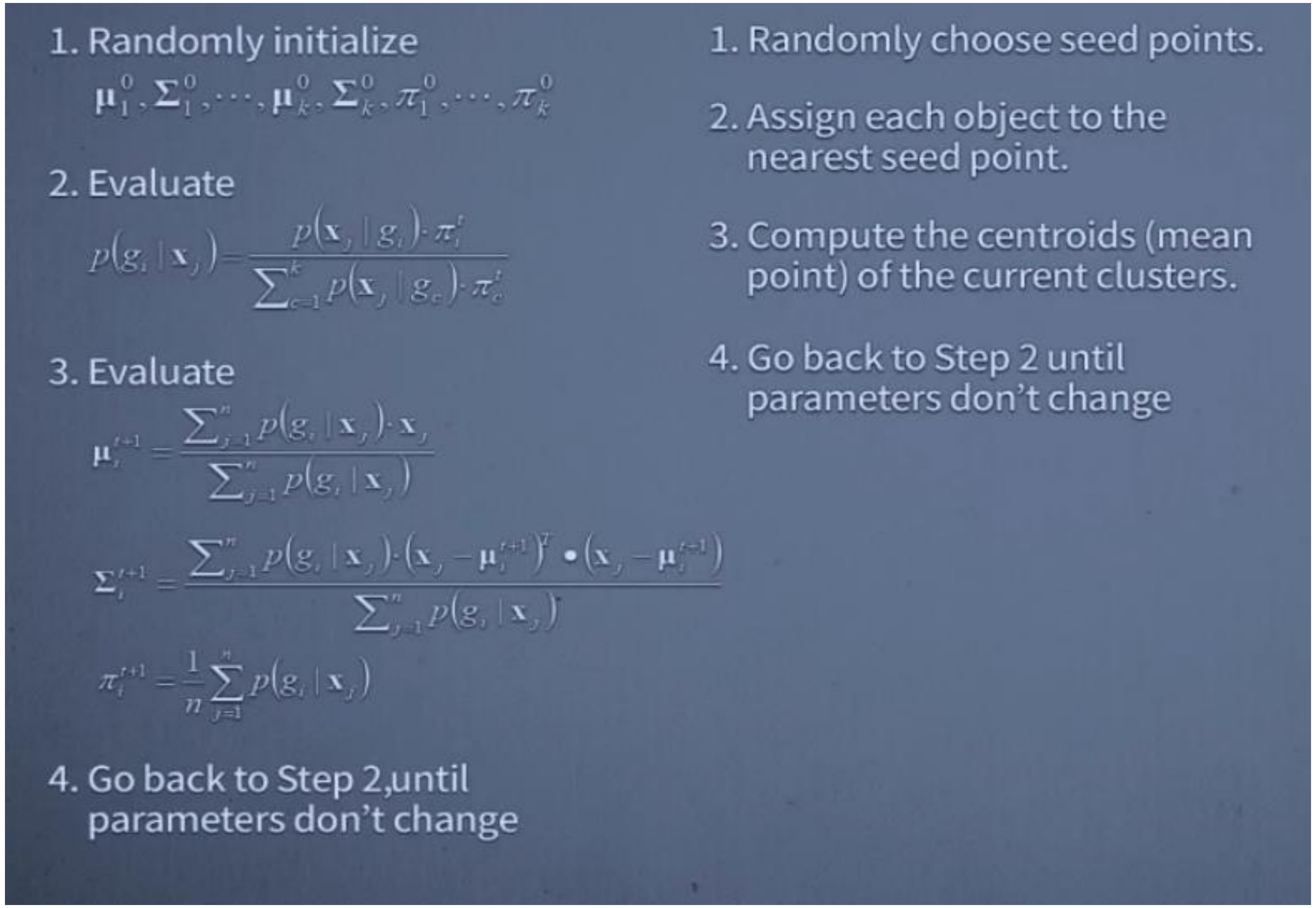
$$\pi_i = \frac{\sum_{j=1}^{n} p(g_i \mid \mathbf{x}_j)}{n}$$

Gaussian Mixture Model

- 어떤 데이터가 어떤 Gaussian에서 나왔는지(Zki) 그리고,
- Parameter(Gaussian이 선택될 확률)과 평균 그리고 분산을 알면 GMM문제를 해결할 수 있다.



GMM vs K-means



1. Randomly initialize(평균, 분산, 확률)
1. 시작점을 랜덤으로 설정
2. 각 데이터가 속할 확률을 계산한다.
2. 각 Object들을 가까운 point에 할당한다.
3. 평균, 분산, 확률 계산
3. 평균을 계산한다.
4. 2번 3번 반복(동일)
-> 값이 거의 변하지 않을 때가지 반복