we have data $X = \{x_1, \dots, x_n\}$

if we had cluster membership

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

$$\sigma_i = \sqrt{\frac{1}{|C_i|} \sum_{x \in C_i} (x - \mu_i)^2}$$

EM

$$w_j = 1/K$$

$$\left.\begin{matrix} \mu_j \\ \sigma_j \end{matrix}\right\} \sim N(0, 1)$$

$\longrightarrow$

E step calculate for each x

prob. that it belongs to each cluster

$$p(\text{dist } j \mid x, \theta_j)$$

M step

using prob from e-step

calculate new estimates for θ

stop when θ doesn't change too much

E step $K = 2$

$$p(\text{dist}_j \mid x_i, \theta) = \frac{w_j \, p(x_i \mid \theta_j)}{w_1 \, p(x_i \mid \theta_1) + w_2 \, p(x_i \mid \theta_2)}$$

M step

$$w_j = \frac{1}{N} \sum_{i=1}^{N} p(\text{dist}\, j \mid x_i, \theta_j)$$

$$\mu_j = \frac{\sum_{i=1}^{N} p(\text{dist}\, j \mid x_i, \theta_j)\, x_i}{\sum_{i=1}^{N} p(\text{dist}\, j \mid x_i, \theta_j)}$$

σ_j^2

$$\sigma_j = \frac{\sum_{i=1}^{N} P(\text{dist } j \mid x_i, \theta_j)(x_i - \mu_j)^2}{\sum_{i=1}^{N} P(\text{dist } j \mid x_i, \theta_j)}$$