Mixture Model

DATASET : $\{x_i\}_{i=1}^{N}$

DATA Likelihood :

$$P(x|\theta) = \sum_{k=1}^{K} \pi_k \cdot P(x|\theta_k)$$

$K \equiv$ # components (hyperparameter)

where $0 \leq \pi_k \leq 1 \quad \forall k$

and $\sum_{k=1}^{K} \pi_k = 1$

$\theta \equiv$ set of all parameters for all K components

$\theta = \{\theta_k\}_{k=1}^{K}$

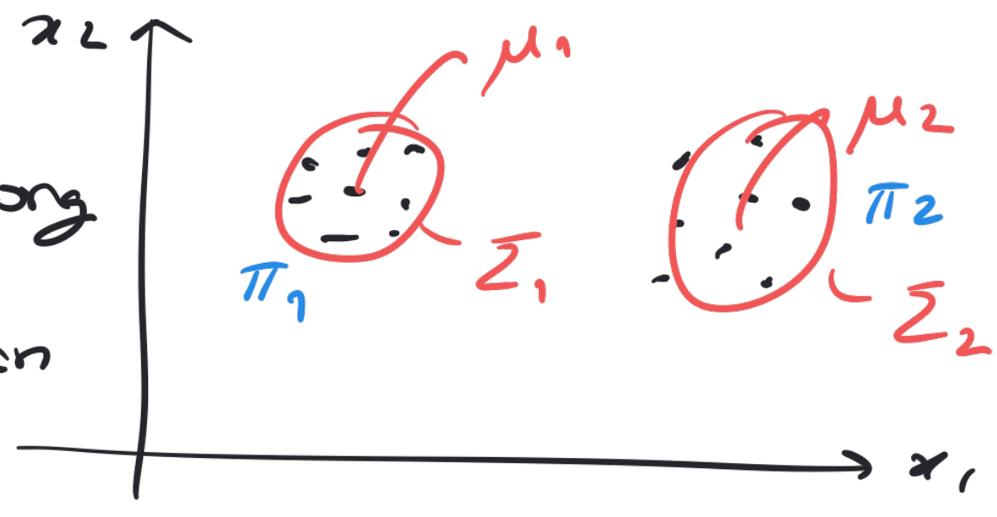
Gaussian Mixture Models (GMMs)

$$P(x|\theta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x | \mu_k, \Sigma_k)$$

where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

PARAMETERS: $\theta = \{ \pi_k, \mu_k, \Sigma_k \}_{k=1}^{K}$

Each x_i can only belong to a single Gaussian component.



DATASET: $\{x_i\}_{i=1}^{N}$

OBSERVED DATA LIKELIHOOD (i.i.d.): $\mathcal{L}^o = \prod_{i=1}^{N} P(x_i|\theta)$

$$\mathcal{L}^o = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

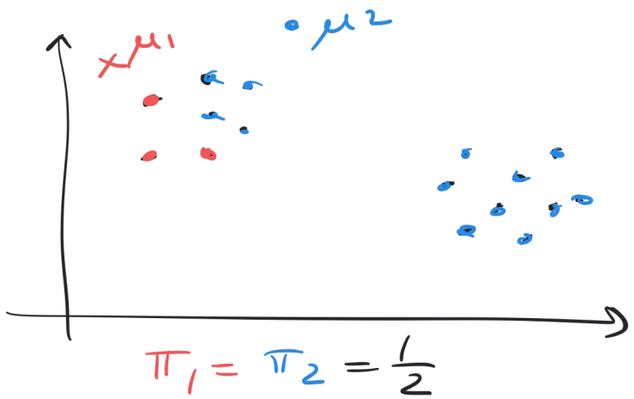
$$\frac{1}{\sqrt{2\pi|\Sigma_k|}^{1/2}} \cdot \underbrace{\mathcal{N}(x|\mu_k, \Sigma_k)}_{\text{Gaussian component}}$$

STEP 1: Provide a value for # components

e.g. $K=2$

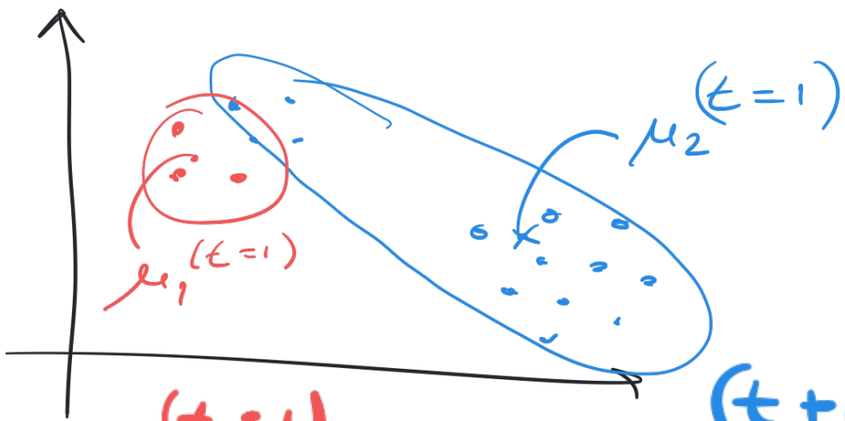
STEP 2: Initialize the parameters $\theta^{(t)} = \{\pi_k, \mu_k, \Sigma_k\}_k$

$t \equiv$ iteration



STEP 3: Assign a membership value for every point.

STEP 4: Fix membership value and update the parameters θ.



STEP 5: Go back to Step 3 and continue until convergence

Applications of GMM :

① DATA Likelihood Estimation.

② clustering.

Partition the data into clusters based on density similarity.

In MAP, we can include a prior on any of the parameters $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ but it will not have any conjugate prior relationship.

OBSERVED DATA LIKELIHOOD:

$$\mathcal{L}^{\circ} = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

$$\mathcal{L} = \ln(\mathcal{L}^{\circ})$$

$$= \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right)$$

this is a "difficult" problem.

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = 0$$

Expectation - Maximization (EM) Algorithm

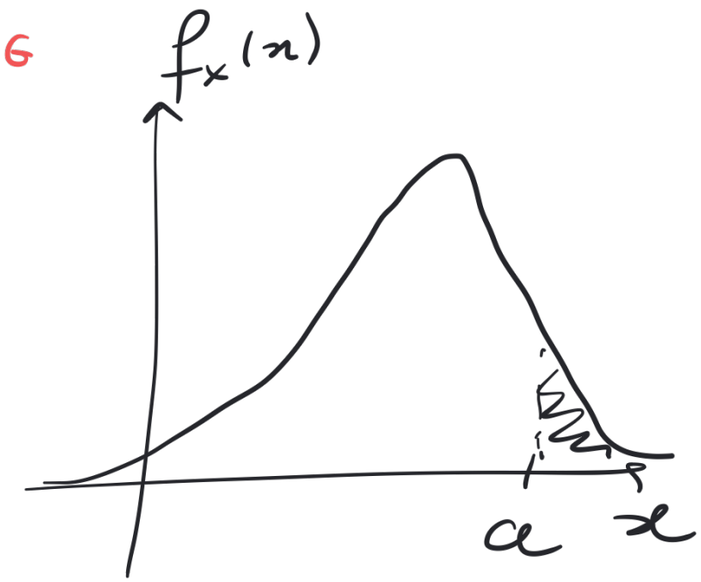
1. Introduce Hidden Latent Variables that will simplify the problem.

2. Iterative greedy algorithm
   - ↳ Solution will depend on the initialization
   - ↳ it may converge to a local optima.

EXAMPLE : CENSORED data

$\underbrace{x_1, x_2, \dots, x_m}_{\equiv \text{MEASURE THE TRUE VALUE}}, \underbrace{x_{m+1}, \dots, x_N}_{= a \equiv \text{CENSORED at a fixed value}}$

TRAINING DATA



PROB. MASS fct : $p_x(x) = \begin{cases} x, & x < a \\ a, & x \geq a \end{cases}$

OBSERVED DATA LIKELIHOOD : $\mathcal{L} = \prod_{i=1}^{m} P(x_i|\theta) \cdot \prod_{j=m+1}^{N} \left(\sum_{x_j = a}^{\infty} P(x_j|\theta) \right)$

EM introduces the hidden latent variables Z that, if given, would simplify the problem.

In the censored data example:

$z_j \equiv$ true (unknown) values for censored samples, $j = m+1, \ldots, N$

With this z variable:

$$\mathcal{L}^{\circ} = \prod_{i=1}^{m} P(x_i|\theta) \cdot \prod_{j=m+1}^{N} P(z_j|\theta)$$

$$\mathcal{L} = \ln \mathcal{L}^{\circ}$$

To find z's, the EM algorithm will find the expectation value (E-STEP) for data likelihood given some initialization for parameters θ.

Optimization function to search for

z's given initial values for

parameters $\theta^{(t)}$ ($t \rightarrow$ iteration):

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_z \left[\ln(\mathcal{L}^\circ) \mid X, \theta^{(t)} \right]$$

expectation

SET of unknown parameters

initializations (numerical values)

$X \equiv$ training data or feature matrix

X is a discrete R.V.

$\mathbb{E}_X[X]$ ≡ mean of X

$$\mathbb{E}_X[X] = \sum_x x \cdot p_X(x)$$

$$\mathbb{E}_X[f(x)] = \sum_x f(x) \cdot p_X(x)$$

$$Q(\theta, \theta^t) = E_z[\ln \mathcal{L}^\circ \mid X, \theta^t]$$

$$Q(\theta, \theta^t) = \sum_{z_i} \ln(\mathcal{L}^\circ) \cdot P(z_i \mid X, \theta^t)$$

EM

$$\arg\max_{\{\theta, z\}} Q(\theta, \theta^t)$$

PSEUDO-CODE for EM algorithm

① Initialization $t=0$

② Initialize parameters θ, $\theta^{(t)} = \{\mu_k^{(t)}, \Sigma_k^{(t)}, \pi_k^{(t)}\}$

③ E-STEP: Hold $\theta^{(t)}$ fixed and find z's that maximize $Q(\theta, \theta^{(t)})$

④ M-STEP (MAXIMIZATION STEP): Hold z's for step ③ and find new values for θ.

⑤ Iterate between steps ③ and ④ until convergence.

Convergence criteria:

⑤.1 MAXIMUM # iterations

⑤.2 threshold for the difference in parameter values between 2 consecutive iterations.

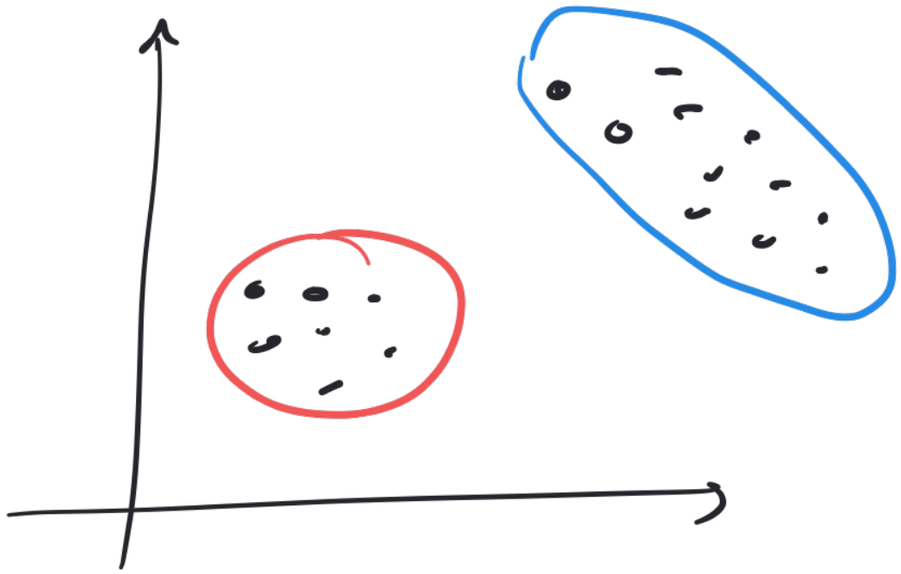
EM algorithm is an

Alternating optimization

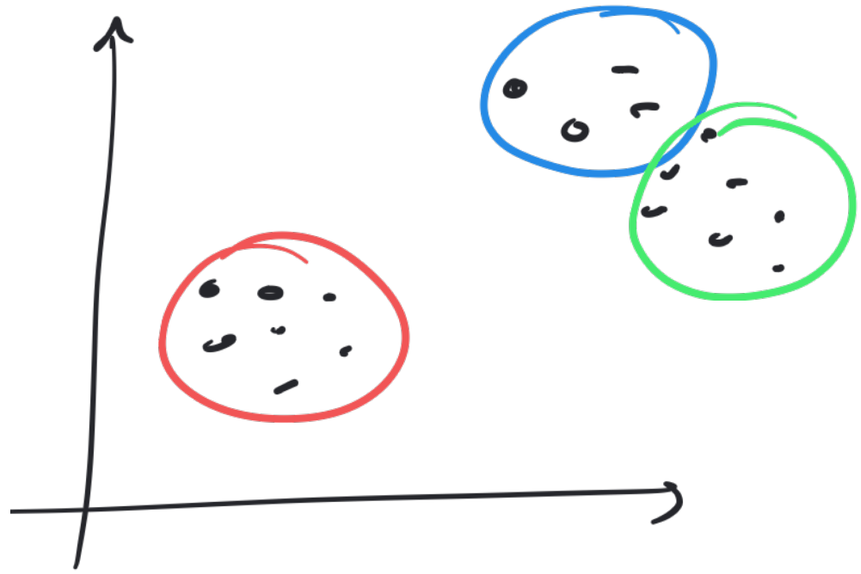
algorithm.

① sensitive to the initialization of parameters θ.

② Convergence is not guaranteed.



K = 2



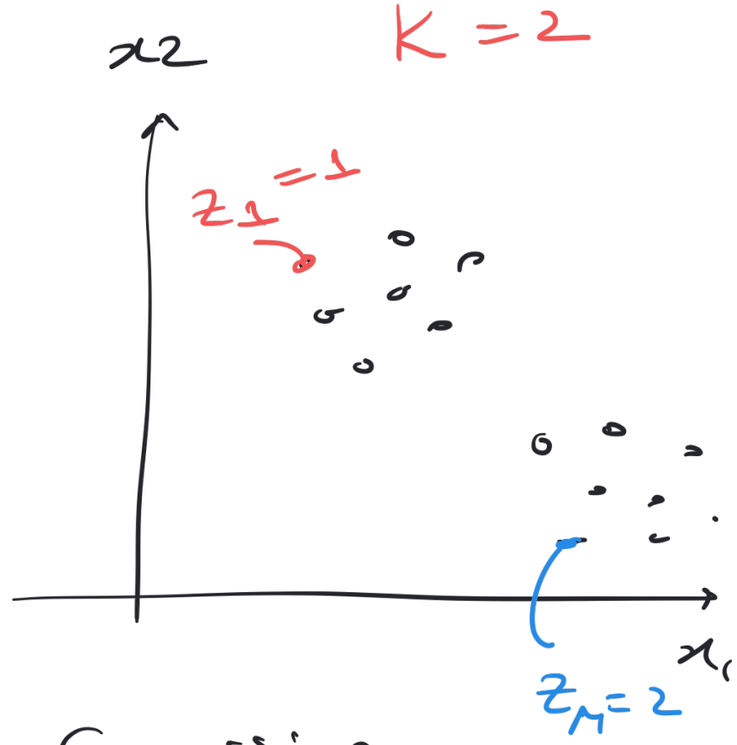
K = 3

GMM

$$\mathcal{L}^{\circ} = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i | \mu_k, \Sigma_k)$$

where $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$

and $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$



Hidden Latent variables : $z_i \equiv$ label of Gaussian component from which point x_i was drawn from

$$z_i = \{1, 2, \ldots, K\}$$

Complete DATA likelihood:
(this includes z's

$$\mathcal{L}^c = \prod_{i=1}^{N} \pi_{z_i} \cdot \mathcal{N}(x_i | \mu_{z_i}, \Sigma_{z_i})$$

EM OPTIMIZATION:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_z [\ln \mathcal{L}^c | X, \theta^t]$$

$$Q(\theta, \theta^t) = E_z[\ln \mathcal{L}^c | X, \theta^t]$$

$$= \sum_{z_i=1}^{K} \ln(\mathcal{L}^c) \cdot P(z_i | X, \theta^t)$$

$$= \sum_{z_i=1}^{K} \ln\left(\prod_{i=1}^{N} \pi_{z_i} \cdot N(x_i | \mu_{z_i}, \Sigma_{z_i})\right) \cdot P(z_i | X, \theta^t)$$

E-STEP: Fix the parameters so this term is constent

M-STEP: this term will be constant.

$$\mathcal{L}^{C} = \prod_{i=1}^{N} \pi_{z_i} \cdot \mathcal{N}(x_i | \mu_{z_i}, \Sigma_{z_i})$$

where

$$\mathcal{N}(x_i | \mu_{z_i}, \Sigma_{z_i}) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_{z_i}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x_i - \mu_{z_i})^{T} \cdot \Sigma_{z_i}^{-1} \cdot (x_i - \mu_{z_i})\right)$$

$d \equiv$ dimension of feature space.

$$Q(\theta, \theta^t) = \sum_{z_i=1}^{K} \ln\left(\prod_{i=1}^{N} \pi_{z_i} \cdot \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i})\right) \cdot P(z_i|X, \theta^t)$$

$$= \sum_{k=1}^{K} \ln\left(\prod_{i=1}^{N} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)\right) \cdot P(z_i = k|X, \theta^t)$$

$$= \sum_{k=1}^{K}\left[\sum_{i=1}^{N}\left(\ln(\pi_k) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)\right)\right] \cdot$$

$$\cdot \; P(z_i = k|X, \theta^t)$$

Let's assume $\Sigma_k = \sigma_k^2 \cdot I$

$$|\Sigma_k| = (\sigma_k^2)^d$$

$$\Sigma_k^{-1} = \frac{1}{\sigma_k^2} \cdot I$$

Plugging in:

$$Q(\theta, \theta^t) = \sum_{k=1}^{K} \left[\sum_{i=1}^{n} \left[\ln(\pi_k) - \frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln(\sigma_k^2) - \frac{1}{2\sigma_k^2} \|x_i - \mu_k\|^2 \right] \right] \cdot P(z_i = k \mid x, \theta^t)$$

① Initialize the parameters

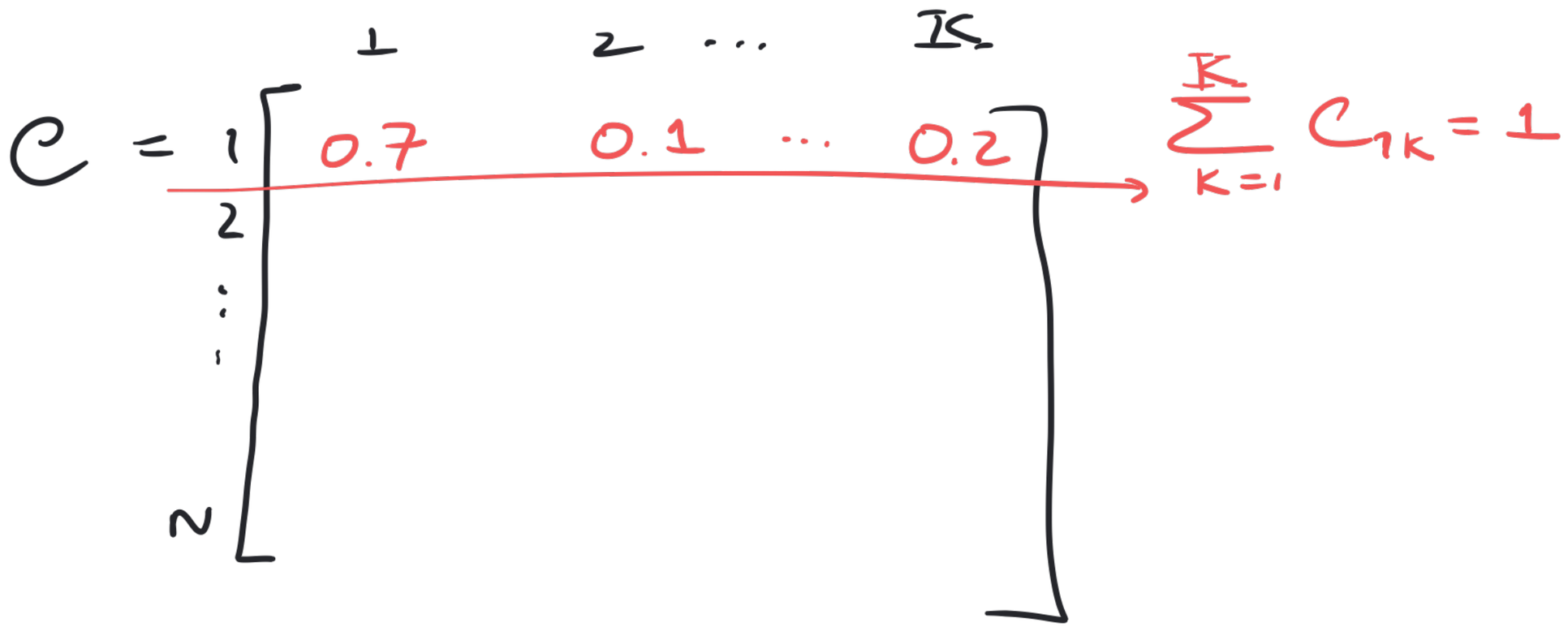
$$\theta^{(t)} = \left\{ \pi_k^{(t)}, \mu_k^{(t)}, \sigma_k^{2\,(t)} \right\}_{k=1}^{K}$$

② E-STEP

Hold $\theta^{(t)}$ fixed and solve for

$$P(z_i = k \mid x, \theta^t) = c_{ik}$$

$\equiv$ membership of sample x_i in component k



Soft membership assignment.

E-STEP

$$P(z_i \mid x_i, \theta^{(t)}) = \frac{P(x_i \mid z_i, \theta^{(t)}) \cdot P(z_i \mid \theta^{(t)})}{P(x_i \mid \theta^{(t)})}$$

$$= \frac{P(x_i \mid z_i, \theta^{(t)}) \cdot P(z_i \mid \theta^{(t)})}{\sum_{k=1}^{K} P(x_i \mid z_i = k, \theta^{(t)}) \cdot P(z_i = k \mid \theta^{(t)})}$$

$$= \frac{\mathcal{N}(x_i \mid \mu_{z_i}^{(t)}, \Sigma_{z_i}^{(t)}) \quad \pi_{z_i}^{(t)}}{\sum_{k=1}^{K} \mathcal{N}(x_i \mid \mu_{z_k}^{(t)}, \Sigma_k^{(t)}) \cdot \pi_k^{(t)}} \equiv c_{ik}$$