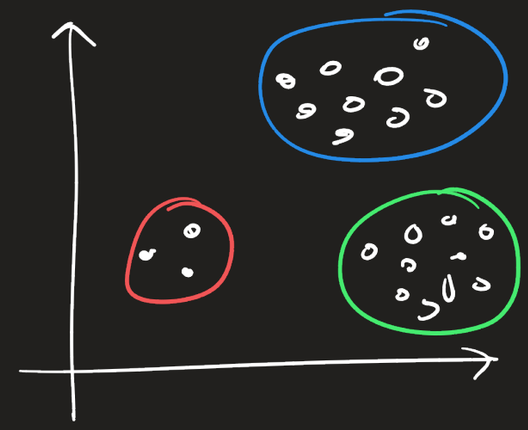
Gaussian Mixture Model (GMM)



For every sample x in the training set,

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

π_k ← weight of each component OR responsability of k^{th} Gaussian component in representing the entire mixture.

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

$K \equiv$ # Gaussian components (hyperparameter)

such that $\sum_{k=1}^{K} \pi_k = 1$

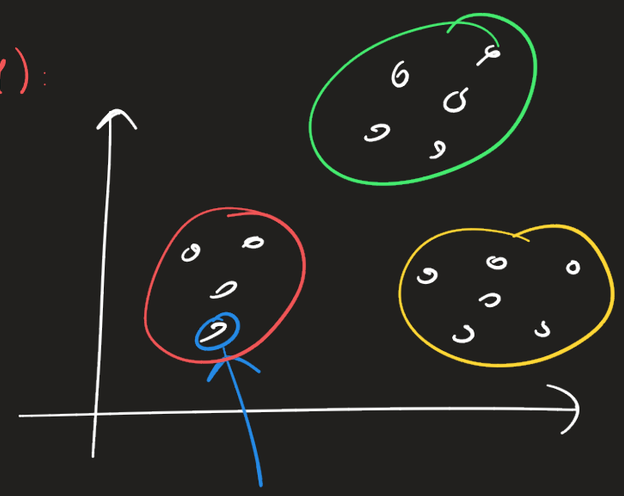
and $0 \leq \pi_k \leq 1$

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

OBSERVED DATA LIKELIHOOD (as a mixture model):

$$\mathcal{L}^{\circ} = \prod_{i=1}^{N} P(x_i|\theta)$$

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



set of parameters:

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

optimization problem:

$$\arg\max_{\theta} \mathcal{L}^{\circ}$$

② Applying log to $\mathcal{L}^\circ$:

Log-likelihood:

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

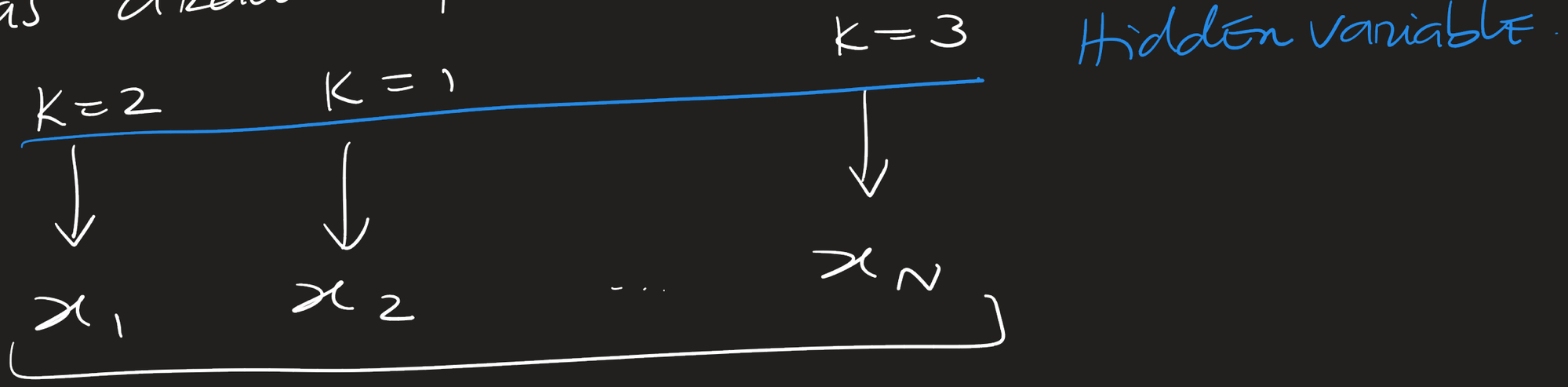
$$= \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \cdot G(x_i | \mu_k, \Sigma_k) \right)$$

intractable density function

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = 0 \quad \dots \text{ no analytical solution } \dots$$

Expectation - Maximization (EM) Algorithm

It introduces a **hidden latent variable**, z, that encodes a label from which Gaussian component each sample x_i was drawn from.

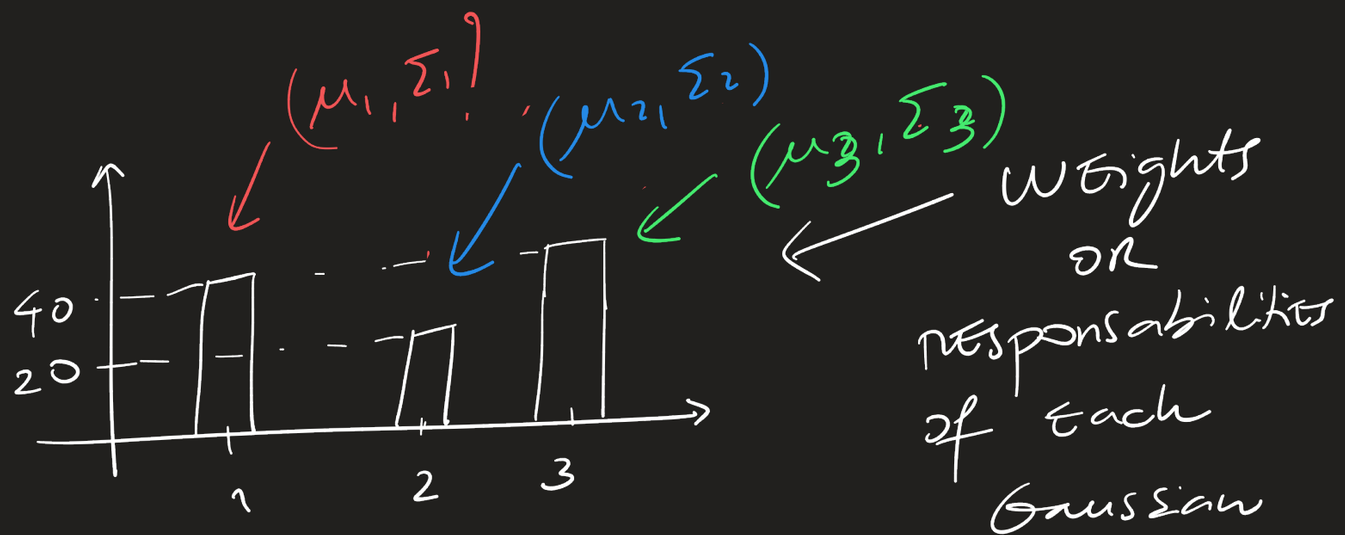


x_i will have a label z_i which corresponds to the Gaussian component it was drawn from.

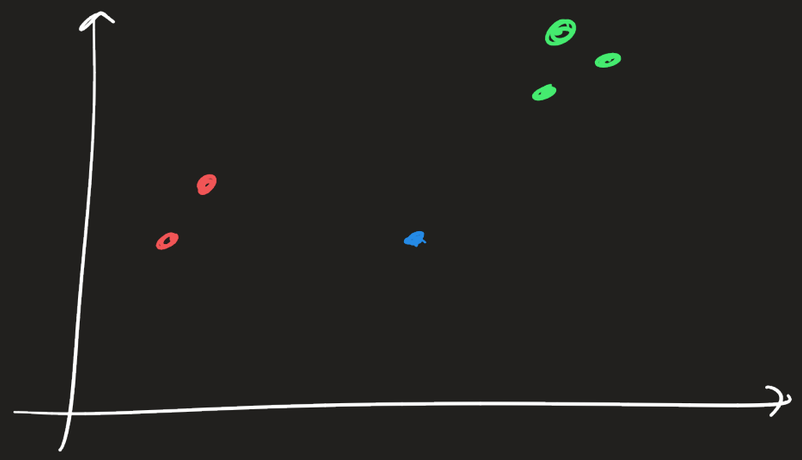
$$z_i \in \{1, 2, \ldots, K\}$$

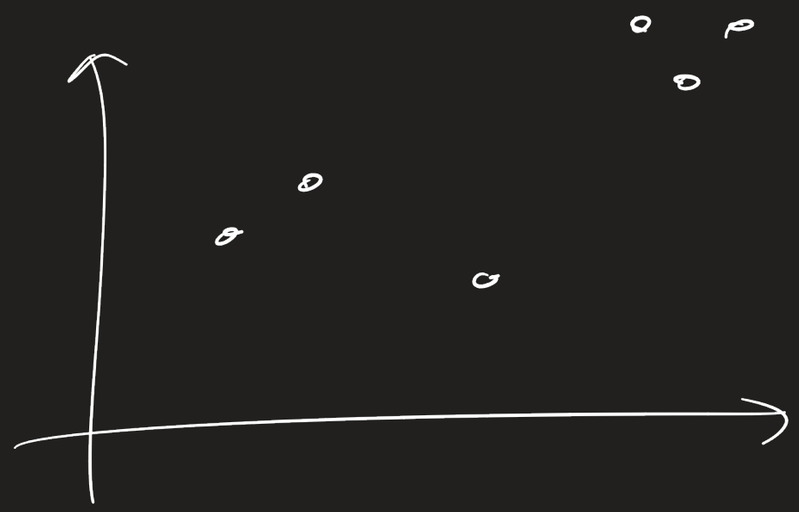
↑

Gaussian component label.



Weights
OR
responsabilities
of each
Gaussian





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

Hidden latent variable: $z_i \equiv$ TRUE UNOBSERVABLE value for the Gaussian component that x_i was drawn from

$$z_i \in \{1, 2, \dots, K\}$$

COMPLETE DATA LIKELIHOOD:

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

such that $\sum_{k=1}^{K} \pi_k = 1$

and $0 \leq \pi_k \leq 1$

EM Algorithm

1. Initialize iteration $t = 0$.
2. Initialize the parameter values
$$\theta^{(t)} = \left\{ \pi_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)} \right\}_{k=1}^{K}$$
3. Expectation (E) - step.
Fix $\theta^{(t)}$, and find the labels $\left\{ z_i^{(t)} \right\}_{i=1}^{N}$.
4. Maximization (M) - step.
Fix the labels $\left\{ z_i^{(t)} \right\}_{i=1}^{N}$. Update the parameters $\theta^{(t+1)}$.
5. go back to step (3) until convergence criteria is met.

The optimization (or objective) function for EM algorithm is:

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

Expected value over R.V. z.

$\theta \equiv$ learnable parameters (variables)

$\theta^{(t)} \equiv$ numerical values for θ.

$X \equiv$ training data

Reminder

① X is a discrete R.V., — P.M.F.

$$\mathbb{E}_X[X] = \sum_i x_i \cdot p_X(x_i)$$

② X is a continuous R.V., — P.D.F.

$$\mathbb{E}_X[X] = \int_{-\infty}^{+\infty} x_i \cdot f_X(x_i)\, dx_i$$

③ X is discrete R.V.,

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

— // — LOTUS (Law of the Unconscious Statistician)

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

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

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

↓ optimize this term in the M-step. (Finds the $\{\mu_k, \Sigma_k, \pi_k\}_{k=1}^{K}$)

↓ optimizing for in E-STEP (Finds the z_i's)

$$= \sum_{i=1}^{N} \sum_{z_i=1}^{K} \underline{\left[\ln(\pi_{z_i}) - \frac{d}{2}\ln(2\pi) - \frac{d}{2} \cdot \ln(|\Sigma_{z_i}|) - \frac{1}{2}(x_i - \mu_{z_i})^T \cdot \Sigma_{z_i}^{-1} \cdot (x_i - \mu_{z_i}) \right]} \cdot \underline{P(z_i | X, \theta^{(t)})}$$

$$= \sum_{i=1}^{N} \sum_{K=1}^{K} \underline{\left[\ln(\pi_z) - \frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln(|\Sigma_K|) - \frac{1}{2}(x_i - \mu_K)^T \cdot \Sigma_K^{-1} \cdot (x_i - \mu_K) \right]} \cdot \underline{P(z_i = K | X, \theta^{(t)})}$$

E – STEP:

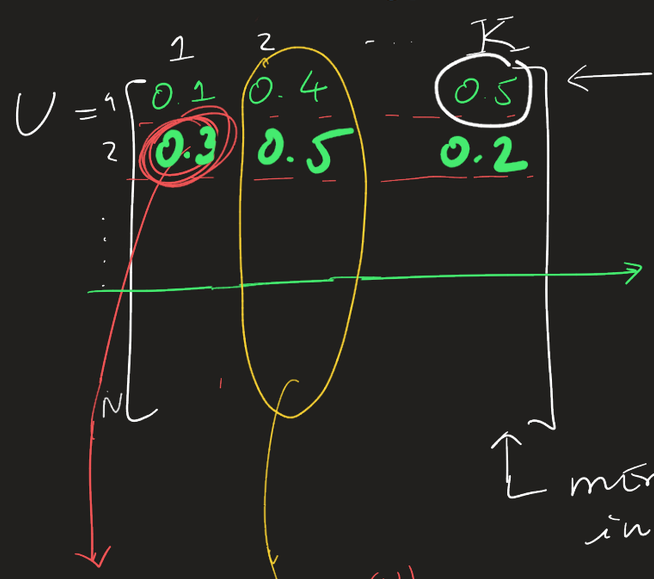
Fix the current value for parameters, $\theta^{(t)}$

Estimate the following:

$$P(z_i = K \mid x_i, \theta^{(t)}) = \frac{\overbrace{P(x_i \mid z_i = K, \theta^{(t)})}^{G(x_i \mid \mu_K, \Sigma_K)} \cdot \overbrace{P(z_i \mid \theta^{(t)})}^{\pi_K}}{P(x_i \mid \theta^{(t)})}$$

$$= \frac{G(x_i \mid \mu_K^{(t)}, \Sigma_K^{(t)}) \cdot \pi_K^{(t)}}{\sum_{j=1}^{K} G(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)}) \cdot \pi_j^{(t)}} = U_{iK} \equiv \text{membership of sample } x_i \text{ in component } K.$$

In computation, this can be written as a matrix, U, the membership matrix. $N \times K$



← membership values of sample x_1 in each gaussian component

$$\sum_{j=1}^{K} U_{ij} = 1, \forall_i$$

↳ membership of all points in gaussian comp. K.

$$P(z_i = 1 \mid x_2, \theta^{(t)}) = U_{12}$$

$$\sum_{i=1}^{N} U_{i2}$$

Find the column/label with largest probability

$$\rightarrow \begin{bmatrix} x_1 \in C_K \\ x_2 \in C_2 \\ \vdots \end{bmatrix} = Z = \begin{bmatrix} K \\ 2 \\ \vdots \\ \vdots \end{bmatrix}_{N \times 1}$$

Component labels for all samples

M-STEP

Fix the z's from E-STEP and optimize for $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$

We can use MLE or MAP for parameter estimation.

There illustrations will use MLE.

I'm going to further assume that

$$\Sigma_k = \sigma_k^2 \cdot I \quad \text{(isotropic covariance)}$$

$$|\Sigma_k| = (\sigma_k^2)^d \quad , \quad \Sigma_k^{-1} = (\sigma_k^2)^{-1} \cdot I$$

↑ assuming d features

$$Q(\theta, \theta^{(t)}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left[\ln(\pi_k) - \frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln(\sigma_k^2) - \frac{1}{2\sigma_k^2}(x_i - \mu_k)^T(x_i - \mu_k) \right] \cdot U_{ik}$$

where $U_{ik} = P(z_i = k \mid x_i, \theta^{(t)})$

(which is fixed during M-Step).

M-Step:

1) Starting w/ μ_k:

$$\frac{\partial Q}{\partial \mu_k} = 0 \iff \sum_{i=1}^{N} \frac{1}{\sigma_k^2}(x_i - \mu_k) \cdot U_{ik} = 0$$

$$\iff \sum_{i=1}^{N} x_i \cdot U_{ik} = \sum_{i=1}^{N} \mu_k \cdot U_{ik}$$

$$\iff \boxed{\mu_k = \frac{\sum_{i=1}^{N} x_i \cdot U_{ik}}{\sum_{i=1}^{N} U_{ik}}}$$

Weighted average.

2) w.r.t. σ_k^2 :

$$\frac{\partial Q}{\partial \sigma_k^2} = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} \left[-\frac{d}{2} \cdot \frac{1}{\sigma_k^2} \cdot \frac{2}{(2\sigma_k^2)^2} (x_i - \mu_k)^T (x_i - \mu_k) \right] \cdot v_{ik} = 0$$

$$\Leftrightarrow \sigma_k^2 = \frac{\sum_{i=1}^{N} (x_i - \mu_k)^T (x_i - \mu_k) \cdot v_{ik}}{d \sum_{i=1}^{N} v_{ik}} \longrightarrow$$

$\equiv$ weighted variance

$$\Sigma_k = \sigma_k^2 \cdot I$$

3) w.r.t. π_k:

Must obey the constraints on π_k

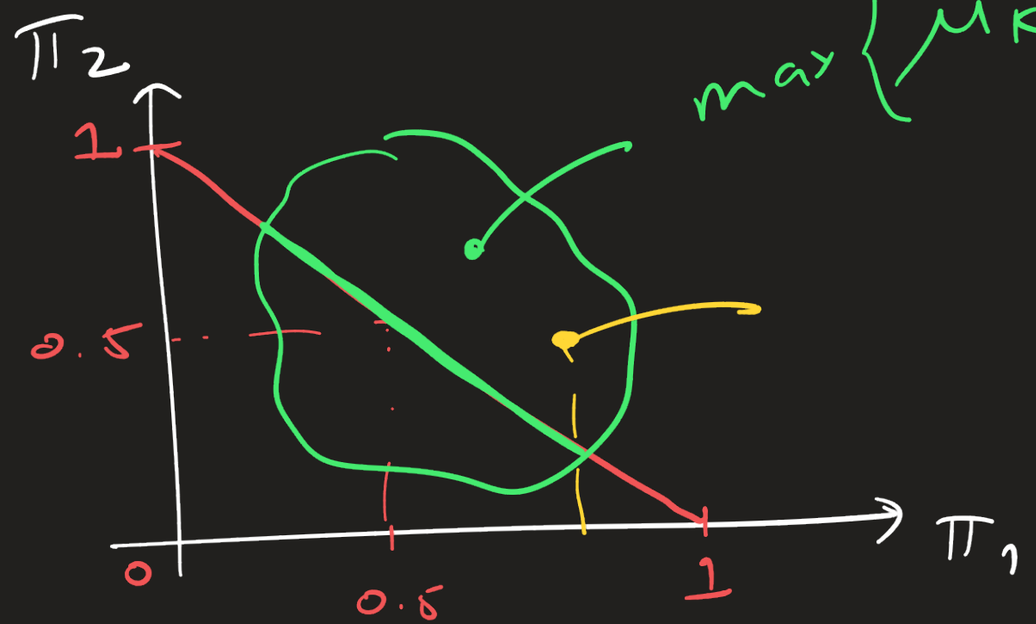
1) $\sum_{k=1}^{K} \pi_k = 1$ ✗ $\Leftrightarrow 1 - \sum_{k=1}^{K} \pi_k = 0$

2) $0 \leq \pi_k \leq 1$, $\forall k$ ✓

To satisfy (1), we add it to the optimization function — Lagrangian optimization problem.

$$Q_\pi(\theta, \theta^{(t)}) = Q(\theta, \theta^{(t)}) + \lambda \left(1 - \sum_{k=1}^{K} \pi_k\right)$$

Lagrangian multiplier



$\max\{\mu_k, \Sigma_k\}$

Appendix E
on the
Bishop
textbook

$$\frac{\partial Q_\pi}{\partial \pi_k} = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{1}{\pi_k} \cdot u_{ik} - \lambda = 0$$

$$\Leftrightarrow \pi_k = \frac{\sum_{i=1}^{N} u_{ik}}{\lambda}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} u_{ik}$$

average Responsability value for component k.

Sum over k, $\sum_{k=1}^{K} \pi_k = 1$:

$$\sum_{k=1}^{K} \left(\sum_{i=1}^{N} u_{ik} \cdot \frac{1}{\lambda} \right) = 1$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{1}{\lambda} \sum_{k=1}^{K} u_{ik} \cdot = 1$$

$$= 1$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{1}{\lambda} = 1 \quad \Leftrightarrow \frac{1}{\lambda} \cdot N = 1 \quad \Leftrightarrow \lambda = N$$