The maximization procedure involves alternating between two steps

t: step $\quad t = 0, 1, 2 \ldots$

E-step

I $\quad Q^{(t+1)} = \underset{Q}{\operatorname{argmax}}\; G(Q, \theta^{(t)})$ $\quad$ closed form

equivalent steps

Compute $Q^{(t+1)}$ with elements

$$Q^{(t+1)}_{i,y} = P_{\theta^{(t)}}[Y = y_i \mid X = x_i]$$

M-step

II $\quad \theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\; G(Q^{(t+1)}, \theta)$

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\; F(Q^{t+1}, \theta)$$

EM algorithm $\quad \theta_0, \theta_1, \theta_2 \ldots$

$Y \to X$

We know the property of G function that $G(Q, \theta) \le L(\theta)$

that means $G(Q, \theta^{(t)}) \le L(\theta^{(t)})$

Now, if we substitute a Q matrix with elements

$$\boxed{Q_{i,y} = P_\theta[Y = y_i \mid X = x_i]}$$

in $G(Q, \theta)$

$$G(Q, \theta) = \sum_{i=1}^{m} \sum_{y_i=1}^{k} \Big(Q_{i,y} \log \underbrace{P_\theta[X = x_i, Y = y_i]}_{A} - Q_{i,y} \log \underbrace{Q_{i,y}}_{B} \Big)$$

$y_i \equiv y$

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} Q_{i,y} \log \frac{P_\theta[X = x_i, Y = y_i]}{Q_{i,y}} \qquad \frac{A}{B}$$

$$P(X, Y) = P(Y \mid X)\, P(X)$$

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} P_\theta[Y = y_i \mid X = x_i] \log \frac{P_\theta[X = x_i, Y = y_i]}{P_\theta[Y = y_i \mid X = x_i]}$$

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} P_\theta[Y = y_i \mid X = x_i] \log \underbrace{P_\theta[X = x_i]}_{\text{does not depend on } y}$$

$$= \sum_{i=1}^{m} \log P_\theta[X = x_i] \underbrace{\sum_{y_i=1}^{k} P_\theta[\underbrace{Y = y_i}_{\text{label}} \mid X = x_i]}_{=1} = L(\theta).$$

$$= \sum_{i=1}^{m} \log P_\theta[X = x_i]$$

Thus, substituting $Q_{i,y} = P_{\underline{\theta}}[Y = y_i \mid X = x_i]$ achieves the maximum value of the $G(Q, \underline{\theta})$ function when $\underline{\theta}$ is fixed.

$\therefore$ we write $$Q^{(t+1)} = \arg\max_{Q} G(Q, \underline{\theta}^{(t)})$$

has elements $$Q_{i,y}^{(t+1)} = P_{\underline{\theta}^{(t)}}[Y = y_i \mid X = x_i]$$

The iterative procedure to maximize the G function is called Expectation Maximization.

The first step is the E step and the second step is the M step.

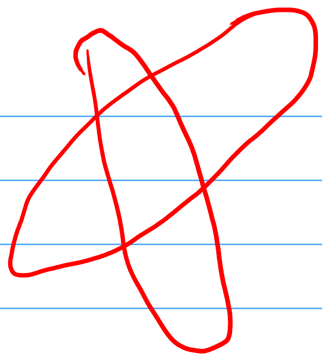
The EM procedure ensures that the observed data log likelihood never decreases.

$$L(\underline{\theta}^{(t+1)}) \geq L(\underline{\theta}^{(t)}).$$

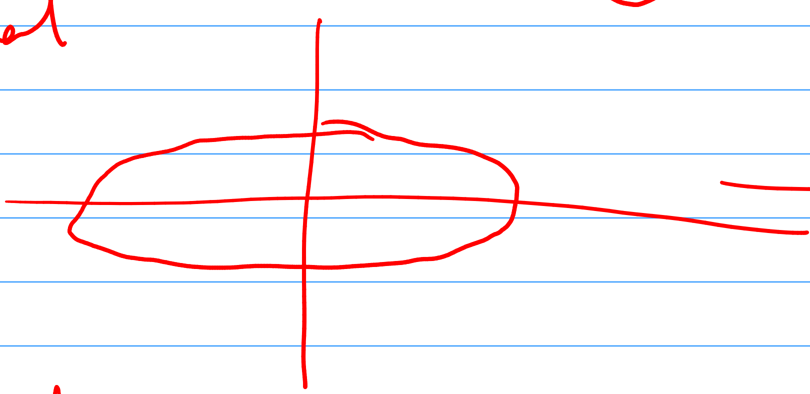
$$L(\underline{\theta}^{t+1}) = G(Q^{(t+2)}, \underline{\theta}^{(t+1)}) \geq G(Q^{(t+1)}, \underline{\theta}^{(t)}) = L(\underline{\theta}^{(t)})$$

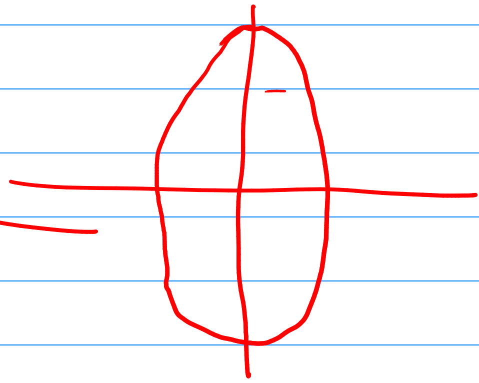
$$\because G(Q^{(t+2)}, \underline{\theta}^{(t+1)}) \geq G(Q^{(t+1)}, \underline{\theta}^{(t+1)})$$
$$\geq G(Q^{(t+1)}, \underline{\theta}^{(t)})$$

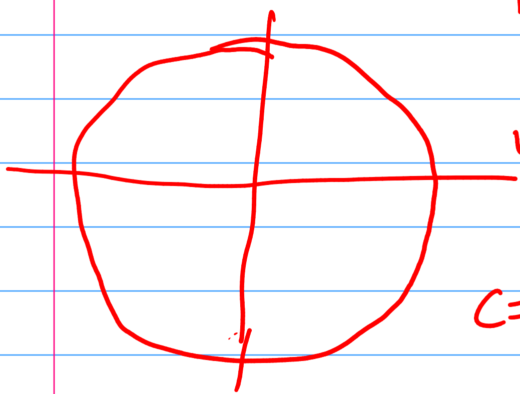
Full Covariance



Diagonal

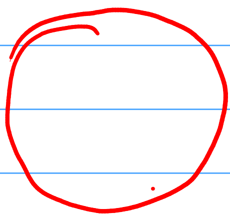






variance values are same

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



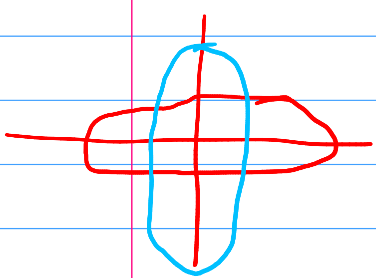
$\mu_1\ \Sigma_1 \qquad \mu_2\ \Sigma_2 \qquad \cdots \qquad \mu_K\ \Sigma_K$

Σ : full / Diag / Identity

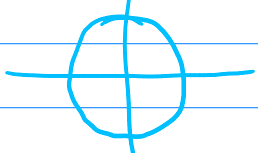
$$\begin{bmatrix} d_1 & & & & 0 \\ & d_2 & & & \\ & & d_3 & & \\ 0 & & & \ddots & \\ & & & & d_K \end{bmatrix}$$

Axis aligned Gaussians

hyperellipses



$$d_1 I \begin{bmatrix} d_1 & & 0 \\ & d_1 & \\ & & d_1 \\ 0 & & & \ddots & d_1 \end{bmatrix}$$



hyperspheres

$$C = \sum_{\underbrace{i=1}}^{m} \underbrace{x_i x_i^T}$$

sum of outer product of vectors.

$\underline{x_i}$ $d \times 1$

$x_i^T x_i$

$1 \times d \times d \times 1$

$= 1 \times 1$

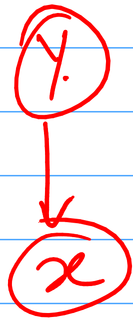
scalar

(dot prod)

Outer product $\underline{x_i}\ x_i^T$

$d \times 1 \times 1 \times d$

$= d \times d$ matrix



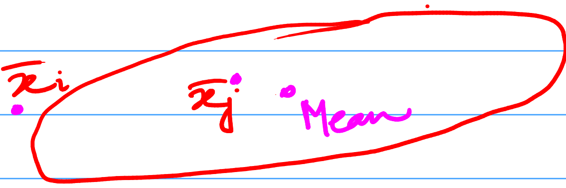
GMM structure

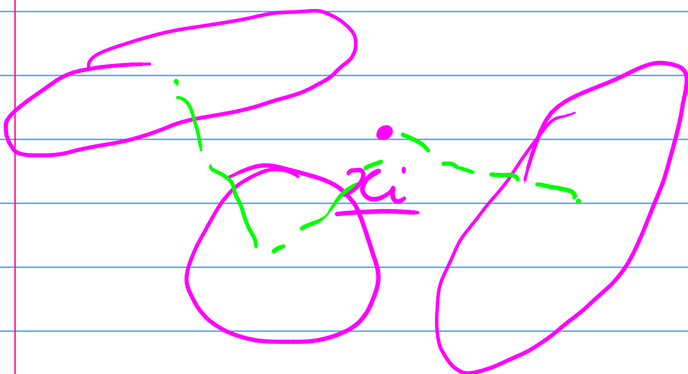
$P(X = x \mid Y = y)$ Gaussian

$\phi,\ \mu,\ \Sigma,$

Organize the data into Gaussian (hyper ellipse) cluster

$$\hat{y}_i = \underset{y}{\operatorname{argmax}}\ \underbrace{P(Y = y \mid \underline{X} = \underline{x_i})}$$





$P(X = x_i \mid Y = y)$

likelihood of generating x_i from Gaussian y

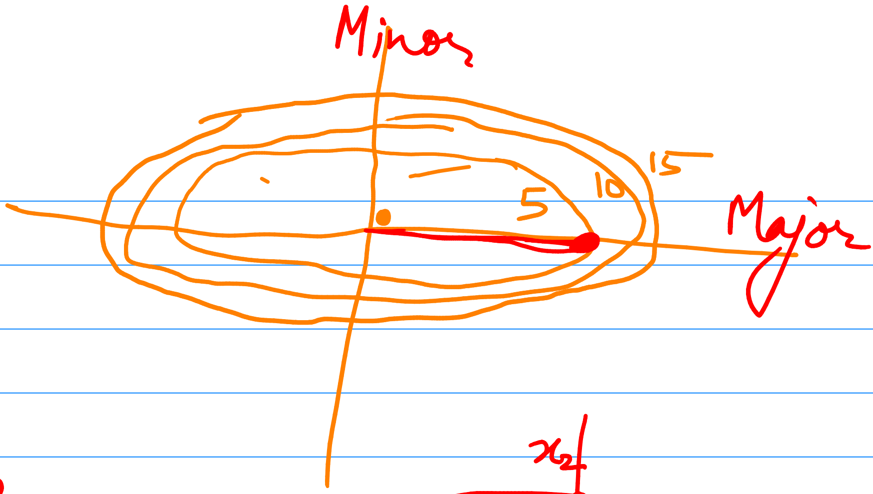
Mahalanobis distance

$$(\underline{x} - \underline{\mu})^T (\underline{x} - \underline{\mu})$$

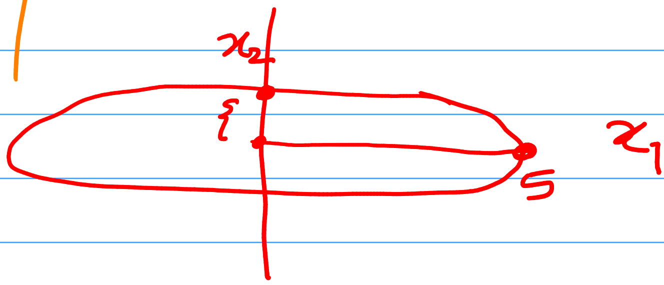
$$(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$$

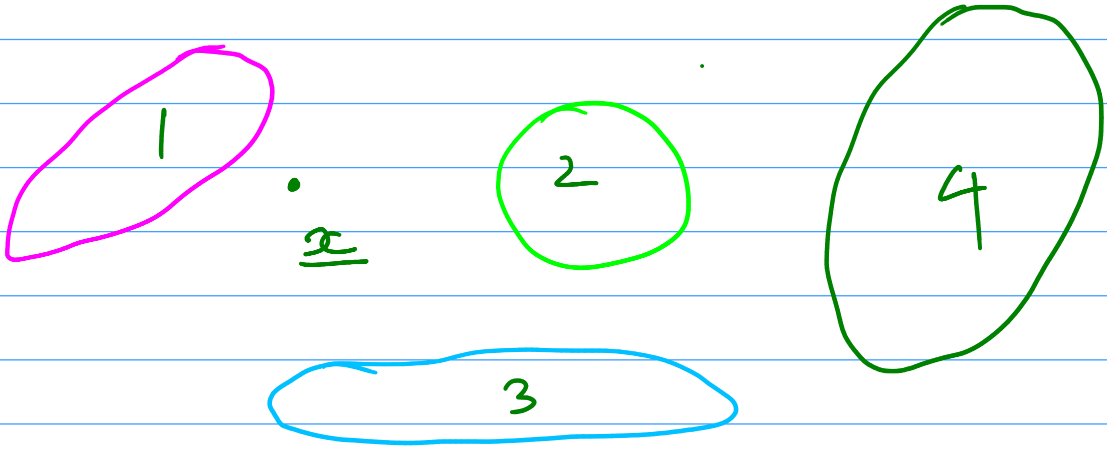
Euclidean distance.

$$\|\underline{x} - \underline{\mu}\|^2$$



$$(x_1-\mu_1)^2 + (x_2-\mu_2)^2$$





$$\underbrace{P(X=\underline{x})} = \sum_k P(X=\underline{x} \mid Y=y_k)\,P(Y=y$$