

# Expectation-Maximization (EM) Algorithm

# Contents

- Introduction
- Main Body
- Mixture Model
- EM-Algorithm on GMM
- Appendix – [Missing Data](#)

# EM Algorithm

Introduction

# Introduction

- EM is typically used to compute **maximum likelihood estimates** given **incomplete samples**.
- The EM algorithm estimates the parameters of a model **iteratively**.
  - **Starting from some initial guess, each iteration consists of**
    - an **E** step (Expectation step)
    - an **M** step (Maximization step)

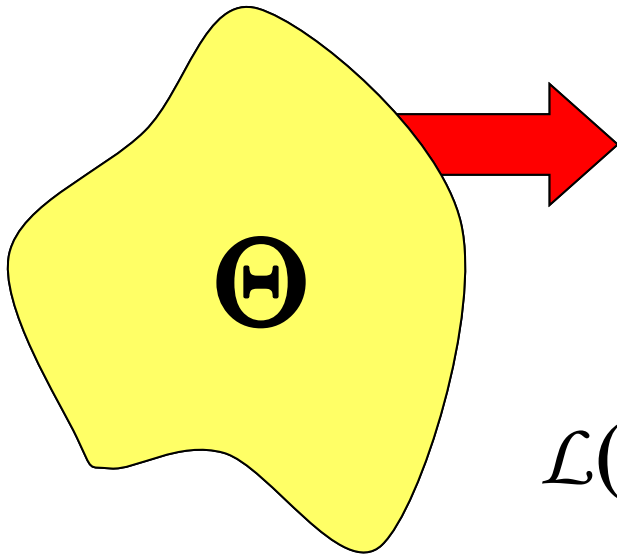
# Applications

- Discovering the value of **latent variables**
- Estimating the parameters of **HMMs**
- Estimating parameters of **finite mixtures**
- Unsupervised learning of **clusters**
- Filling in **missing data** in samples
- ...

# EM Algorithm

Main Body

# Maximum Likelihood

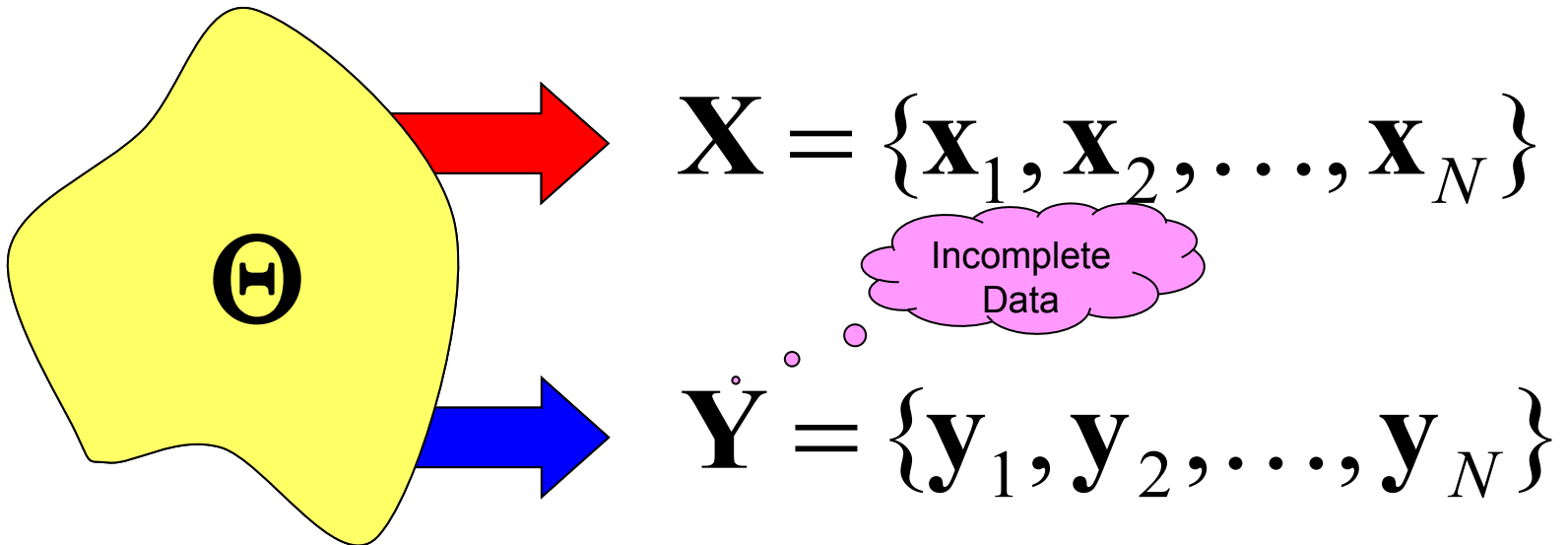


$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$$

$$\mathcal{L}(\Theta | \mathbf{X}) = p(\mathbf{X} | \Theta) = \prod_{i=1}^N p(\mathbf{x}_i | \Theta)$$

$$\Theta^* = \arg \max_{\Theta} \mathcal{L}(\Theta | \mathbf{X})$$

# Latent Variables

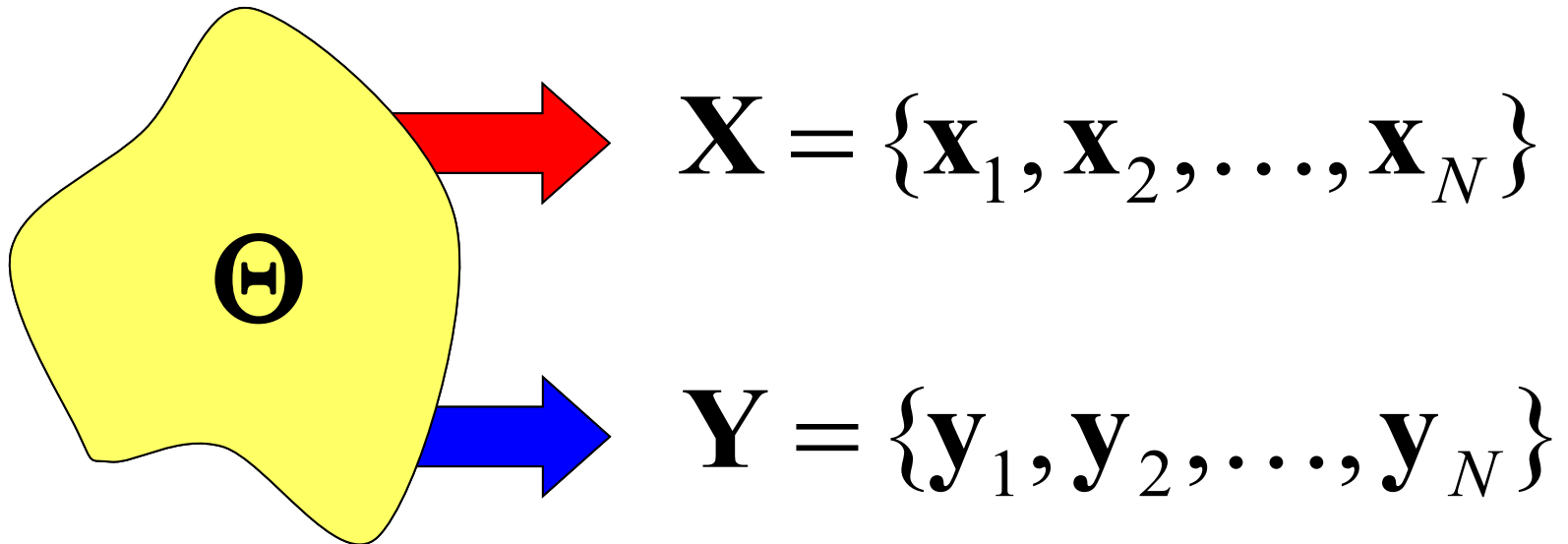


Complete Data  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$



# Complete Data Likelihood

Complete Data  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$

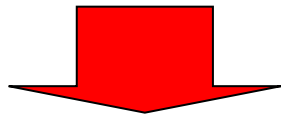


$$\begin{aligned}\mathcal{L}(\Theta | \mathbf{Z}) &= p(\mathbf{Z} | \Theta) = p(\mathbf{X}, \mathbf{Y} | \Theta) \\ &= p(\mathbf{Y} | \mathbf{X}, \Theta) p(\mathbf{X} | \Theta)\end{aligned}$$

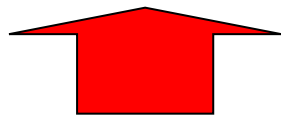
# Complete Data Likelihood

Complete Data  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$

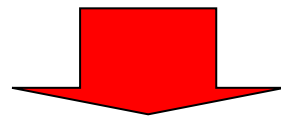
$$\mathcal{L}(\Theta | \mathbf{Z}) = \underbrace{p(\mathbf{Y} | \mathbf{X}, \Theta)}_{\text{A function of latent variable } \mathbf{Y} \text{ and parameter } \Theta} \underbrace{p(\mathbf{X} | \Theta)}_{\text{A function of parameter } \Theta}$$



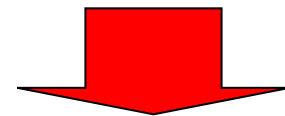
A function of  
random  
variable  $\mathbf{Y}$ .



If we are given  $\Theta$ ,



The result is in  
term of random  
variable  $\mathbf{Y}$ .



Computable

# Expectation

Expectation:

$$E(X) = \int xp(X = x)dx = \int xp(x)dx$$

$$E(f(X)) = \int f(x)p(x)dx$$

Conditional Expectation:

$$E(X | Y = y) = \int xp(x | y)dx$$

$$E(f(X) | Y = y) = \int f(x)p(x | y)dx$$

$$\mathcal{L}(\Theta | \mathbf{Z}) = p(\mathbf{X}, \mathbf{Y} | \Theta)$$

# Expectation Step

Let  $\Theta^{(i-1)}$  be the parameter vector obtained at the  $(i-1)^{th}$  step.

Define (Conditional Expectation of log likelihood of complete data)

$$Q(\Theta, \Theta^{(i-1)}) = E[\log \mathcal{L}(\Theta | \mathbf{Z}) | \mathbf{X}, \Theta^{(i-1)}]$$
$$= \begin{cases} \int_{\mathbf{y} \in \mathbf{Y}} \log p(\mathbf{X}, \mathbf{y} | \Theta) \cdot p(\mathbf{y} | \mathbf{X}, \Theta^{(i-1)}) d\mathbf{y} & \text{continuous} \\ \sum_{\mathbf{y} \in \mathbf{Y}} \log p(\mathbf{X}, \mathbf{y} | \Theta) \cdot p(\mathbf{y} | \mathbf{X}, \Theta^{(i-1)}) & \text{discrete} \end{cases}$$

$$\mathcal{L}(\Theta | \mathbf{Z}) = p(\mathbf{X}, \mathbf{Y} | \Theta)$$

# Maximization Step

$$\Theta^{(i)} = \arg \max_{\Theta} Q(\Theta, \Theta^{(i-1)})$$

Define

$$Q(\Theta, \Theta^{(i-1)}) = E[\log \mathcal{L}(\Theta | \mathbf{Z}) | \mathbf{X}, \Theta^{(i-1)}]$$

$$= \begin{cases} \int_{\mathbf{y} \in \mathbf{Y}} \log p(\mathbf{X}, \mathbf{y} | \Theta) \cdot p(\mathbf{y} | \mathbf{X}, \Theta^{(i-1)}) d\mathbf{y} & \text{continuous} \\ \sum_{\mathbf{y} \in \mathbf{Y}} \log p(\mathbf{X}, \mathbf{y} | \Theta) \cdot p(\mathbf{y} | \mathbf{X}, \Theta^{(i-1)}) & \text{discrete} \end{cases}$$

# EM Algorithm

Mixture Model

# Mixture Models

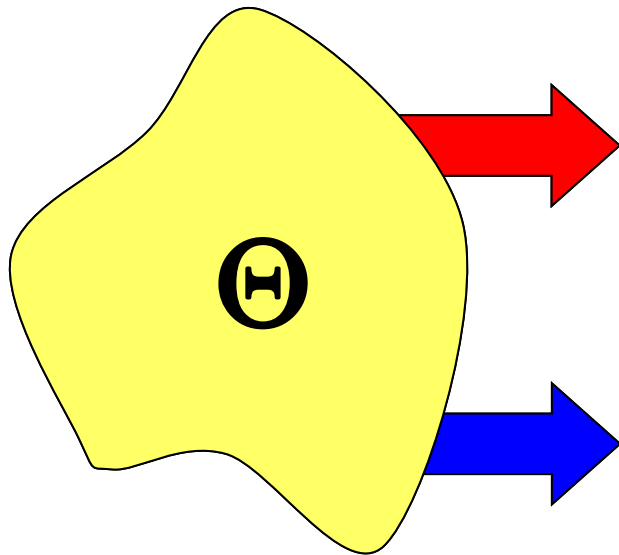
- If there is a reason to believe that a data set is comprised of **several distinct populations**, a mixture model can be used.
- It has the following form:

$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j) \quad \text{with} \quad \sum_{j=1}^M \alpha_j = 1$$

$$\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$$

# Mixture Model

$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)$$



$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$$

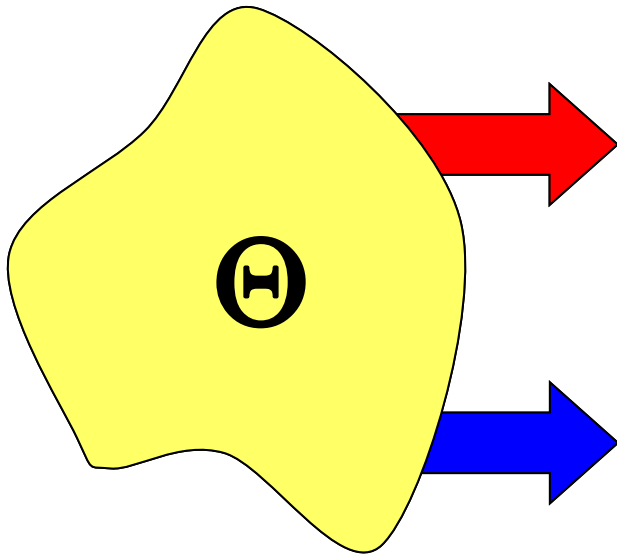
$$\mathbf{Y} = \{y_1, y_2, \dots, y_N\}$$

Let  $y_i \in \{1, \dots, M\}$  represents the source that generates the data.



# Mixture Models

$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)$$



$\mathbf{x}$

$$p(\mathbf{x} | y = j, \Theta) = p_j(\mathbf{x} | \theta_j)$$

$y = j$

$$p(y = j | \Theta) = \alpha_j$$

Let  $y_i \in \{1, \dots, M\}$  represents the source that generates the data.

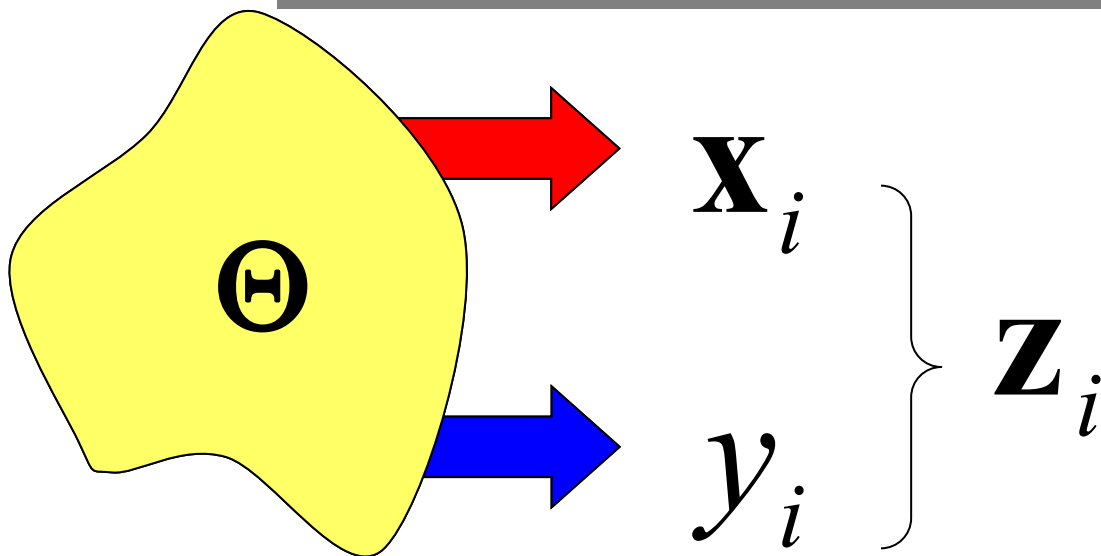
# Mixture Models

$$p(\mathbf{x} | y = j, \Theta) = p_j(\mathbf{x} | \theta_j)$$

$$p(y = j | \Theta) = \alpha_j$$

$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)$$

$$p(\mathbf{z}_i | \Theta) = p(\mathbf{x}_i, y_i | \Theta) = p(y_i | \mathbf{x}_i, \Theta) p(\mathbf{x}_i | \Theta)$$



$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)$$

$$p(\mathbf{x} | y = j, \Theta) = p_j(\mathbf{x} | \theta_j)$$

$$p(y = j | \Theta) = \alpha_j$$

$$p(\mathbf{z}_i | \Theta) = p(\mathbf{x}_i, y_i | \Theta) = p(y_i | \mathbf{x}_i, \Theta) p(\mathbf{x}_i | \Theta)$$

$$\begin{aligned} p(y_i | \mathbf{x}_i, \Theta) &= \frac{p(\mathbf{x}_i, y_i, \Theta)}{p(\mathbf{x}_i, \Theta)} = \frac{p(\mathbf{x}_i | y_i, \Theta) p(y_i, \Theta)}{p(\mathbf{x}_i, \Theta)} \\ &= \frac{p(\mathbf{x}_i | y_i, \Theta) p(y_i | \Theta) p(\Theta)}{p(\mathbf{x}_i | \Theta) p(\Theta)} = \frac{p(\mathbf{x}_i | y_i, \Theta) p(y_i | \Theta)}{p(\mathbf{x}_i | \Theta)} \\ &= \frac{p_{y_i}(\mathbf{x}_i | \theta_{y_i}) \alpha_{y_i}}{\sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)} \end{aligned}$$

$$p(\mathbf{x} | y = j, \Theta) = p_j(\mathbf{x} | \theta_j)$$

$$p(y = j | \Theta) = \alpha_j$$

$$p(\mathbf{x} | \Theta) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)$$

$$p(\mathbf{z}_i | \Theta) = p(\mathbf{x}_i, y_i | \Theta) = \text{[redacted]} p(\mathbf{x}_i | \Theta)$$

$$p(y_i | \mathbf{x}_i, \Theta) = \frac{p(\mathbf{x}_i, y_i, \Theta)}{p(\mathbf{x}_i, \Theta)} = \frac{p(\mathbf{x}_i | y_i, \Theta) p(y_i, \Theta)}{p(\mathbf{x}_i, \Theta)}$$

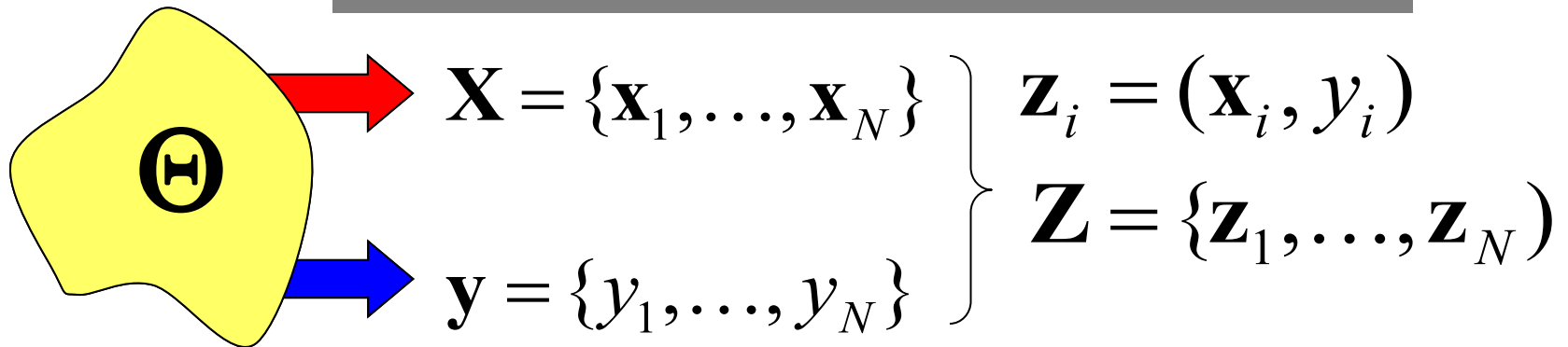
$$= \frac{p(\mathbf{x}_i | y_i, \Theta)}{p(\mathbf{x}_i, \Theta)}$$

$$= \frac{p_{y_i}(\mathbf{x}_i | \theta_{y_i}) \alpha_{y_i}}{\sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \theta_j)}$$

Given  $\mathbf{x}$  and  $\Theta$ , the conditional density of  $y$  can be computed.

# Complete-Data Likelihood Function

$$\mathcal{L}(\Theta | \mathbf{Z}) = p(\mathbf{X}, \mathbf{y} | \Theta) = \prod_{i=1}^N \alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})$$



$$\begin{aligned} \mathcal{L}(\Theta | \mathbf{Z}) &= p(\mathbf{Z} | \Theta) = p(\mathbf{X}, \mathbf{y} | \Theta) = p(\mathbf{X} | \mathbf{y}, \Theta) p(\mathbf{y} | \Theta) \\ &= \prod_{i=1}^N p(\mathbf{x}_i | \underbrace{y_i, \Theta}_{\theta_{y_i}}) p(y_i | \underbrace{\Theta}_{\alpha_{y_i}}) = \prod_{i=1}^N \alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i}) \end{aligned}$$

# Expectation

$$\mathcal{L}(\Theta | \mathbf{Z}) = p(\mathbf{X}, \mathbf{y} | \Theta) = \prod_{i=1}^N \alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})$$

$$\log \mathcal{L}(\Theta | \mathbf{Z}) = \sum_{i=1}^N \log [\alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})]$$

$$Q(\Theta, \Theta^g) = E[\log \mathcal{L}(\Theta | \mathbf{Z}) | \mathbf{X}, \Theta^g] \quad \Theta^g: \text{Guess}$$

$$= \sum_{\mathbf{y} \in \mathcal{Y}} \log \mathcal{L}(\Theta | \mathbf{Z}) p(\mathbf{y} | \mathbf{X}, \Theta^g)$$

$$= \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^N \log [\alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})] \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g)$$

# Expectation

$$\mathcal{L}(\Theta | \mathbf{Z}) = p(\mathbf{X}, \mathbf{y} | \Theta) = \prod_{i=1}^N \alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})$$

$$\log \mathcal{L}(\Theta | \mathbf{Z}) = \sum_{i=1}^N \log [\alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})]$$

$$Q(\Theta, \Theta^g) = E[\log \mathcal{L}(\Theta | \mathbf{Z}) | \mathbf{X}, \Theta^g] \quad \Theta^g: \text{Guess}$$

$$\begin{aligned} &= \sum_{\mathbf{y} \in \mathcal{Y}} \log \mathcal{L}(\Theta | \mathbf{Z}) p(\mathbf{y} | \mathbf{X}, \Theta^g) \\ &= \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^N \log [\alpha_{y_i} p_{y_i}(\mathbf{x}_i | \theta_{y_i})] \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \\ &= \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^N \sum_{l=1}^M \delta_{y_i, l} \log [\alpha_l p_l(\mathbf{x}_i | \theta_l)] \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \end{aligned}$$

# Expectation

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g)$$
$$= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g)$$

Zero when  $y_i \neq l$

$$Q(\Theta, \Theta^g) = \sum_{i=1}^N \sum_{l=1}^M \delta_{y_i, l} \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g)$$



# Expectation

$$\begin{aligned}
 Q(\Theta, \Theta^g) &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \sum_{\mathbf{y} \in \mathcal{Y}} \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \\
 &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \sum_{y_1=1}^M \cdots \sum_{y_i=1}^M \cdots \sum_{y_N=1}^M \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \\
 &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \left( \sum_{y_1=1}^M \cdots \sum_{\substack{y_{i-1}=1 \\ y_{i+1}=1}}^M \cdots \sum_{\substack{y_N=1 \\ j=1 \\ j \neq i}}^M p(y_j | \mathbf{X}, \Theta^g) \right) p(l | \mathbf{x}_i, \Theta^g)
 \end{aligned}$$

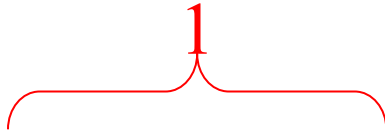
# Expectation

$$\begin{aligned}
 Q(\Theta, \Theta^g) &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \sum_{y \in \mathcal{Y}} \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \\
 &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \sum_{y_1=1}^M \cdots \sum_{y_i=1}^M \cdots \sum_{y_N=1}^M \delta_{y_i, l} \prod_{j=1}^N p(y_j | \mathbf{x}_j, \Theta^g) \\
 &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \left[ \sum_{y_1=1}^M \cdots \sum_{\substack{y_{i-1}=1 \\ y_{i+1}=1}}^M \cdots \sum_{\substack{y_N=1 \\ j=1 \\ j \neq i}}^M p(y_j | \mathbf{X}, \Theta^g) \right] p(l | \mathbf{x}_i, \Theta^g) \\
 Q(\Theta, \Theta^g) &= \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \left[ \prod_{\substack{j=1 \\ j \neq i}}^N \left( \sum_{y_j=1}^M p(y_j | \mathbf{X}, \Theta^g) \right) \right] p(l | \mathbf{x}_i, \Theta^g)
 \end{aligned}$$

# Expectation

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log[\alpha_l p_l(\mathbf{x}_i | \theta_l)] \left[ \prod_{\substack{j=1 \\ j \neq i}}^N \left( \sum_{y_j=1}^M p(y_j | \mathbf{X}, \Theta^g) \right) \right] p(l | \mathbf{x}_i, \Theta^g)$$


# Maximization

$$\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$$

Given the initial guess  $\Theta^g$ ,

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

We want to find  $\Theta$ , to maximize the above expectation.

In fact, iteratively.

# EM Algorithm

EM-Algorithm on  
GMM

# The GMM (Gaussian Mixture Model)

Gaussian model of a  $d$ -dimensional source, say  $j$  :

$$p_j(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right]$$

$$\theta_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

GMM with  $M$  sources:

$$p_j(\mathbf{x} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) = \sum_{j=1}^M \alpha_j p_j(\mathbf{x} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$\begin{aligned} \alpha_j &\geq 0 \\ \sum \alpha_j &= 1 \end{aligned}$$

# Goal

Mixture Model

$$p(\mathbf{x} | \Theta) = \sum_{l=1}^M \alpha_l p_l(\mathbf{x} | \theta_l)$$

$$\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$$

subject to 
$$\sum_{l=1}^M \alpha_l = 1$$

To maximize:

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

# Goal

Mixture Model

$$p(\mathbf{x} | \Theta) = \sum_{l=1}^M \alpha_l p_l(\mathbf{x} | \theta_l)$$

Correlated  
with  $\alpha_l$  only.

$$\Theta = (\alpha_1, \dots)$$

Correlated  
with  $\theta_l$  only.

subject to

$$\sum_{l=1}^M \alpha_l = 1$$

To maximize:

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$



# Finding $\alpha_l$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

Due to the constraint on  $\alpha_l$ 's, we introduce *Lagrange Multiplier*  $\lambda$ , and solve the following equation.

$$\frac{\partial}{\partial \alpha_l} \left[ \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \lambda \left( \sum_{l=1}^M \alpha_l - 1 \right) \right] = 0, \quad l = 1, \dots, M$$



$$\sum_{i=1}^N \frac{1}{\alpha_l} p(l | \mathbf{x}_i, \Theta^g) + \lambda = 0, \quad l = 1, \dots, M$$



$$\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) + \alpha_l \lambda = 0, \quad l = 1, \dots, M$$

# Finding $\alpha_l$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

$$\lambda = -N$$



$$\sum_{l=1}^M \sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) + \lambda \sum_{l=1}^M \alpha_l = 0$$



$$\sum_{i=1}^N \sum_{l=1}^M p(l | \mathbf{x}_i, \Theta^g) + \lambda \sum_{l=1}^M \alpha_l = 0$$

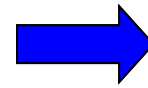
} 1                      } 1  
} N



$$\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) + \alpha_l \lambda = 0, \quad l = 1, \dots, M$$

# Finding $\alpha_l$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$



$$\lambda = -N$$



$$\alpha_l = \frac{1}{N} \sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)$$

$$p(l | \mathbf{x}_i, \Theta^g) = \frac{\alpha_l^g p_l(\mathbf{x}_i | \theta_l^g)}{\sum_{j=1}^M \alpha_j^g p_j(\mathbf{x} | \theta_j^g)}$$



$$\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) + \alpha_l \lambda = 0, \quad l = 1, \dots, M$$

# Finding $\theta_l$

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[p_l(\mathbf{x}_i | \theta_l)] p(l | \mathbf{x}_i, \Theta^g)$$

Consider GMM

Only need to maximize  
this term

$$p_l(\mathbf{x} | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_l|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right]$$

$$\theta_l = (\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$$

$$\log[p_l(\mathbf{x} | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)] = \underbrace{-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_l|^{1/2}}_{\text{unrelated}} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)$$

unrelated

$$Q(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l | \mathbf{x}_i, \Theta^g) + \sum_{l=1}^M \sum_{i=1}^N \log[\text{redacted}] p(l | \mathbf{x}_i, \Theta^g)$$

Only need to maximize  
this term

Therefore, we want to maximize:

$$Q'(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \left( -\frac{1}{2} \log |\Sigma_l|^{1/2} - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_l) \right) p(l | \mathbf{x}_i, \Theta^g)$$

**How?** knowledge on matrix algebra is needed.

$$\log[ p_l(\mathbf{x} | \boldsymbol{\mu}_l, \Sigma_l) ] = \underbrace{-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_l|^{1/2}}_{\text{unrelated}} - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)$$

$$p(l | \mathbf{x}_i, \Theta^g) = \frac{\alpha_l^g p_l(\mathbf{x}_i | \theta_l^g)}{\sum_{j=1}^M \alpha_j^g p_j(\mathbf{x} | \theta_j^g)}$$

Therefore, we want to maximize:

$$Q'(\Theta, \Theta^g) = \sum_{l=1}^M \sum_{i=1}^N \left( -\frac{1}{2} \log |\Sigma_l|^{1/2} - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_l) \right) p(l | \mathbf{x}_i, \Theta^g)$$

$$\boldsymbol{\mu}_l = \frac{\sum_{i=1}^N \mathbf{x}_i p(l | \mathbf{x}_i, \Theta^g)}{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)}$$

$$\Sigma_l = \frac{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) (\mathbf{x}_i - \boldsymbol{\mu}_l) (\mathbf{x}_i - \boldsymbol{\mu}_l)^T}{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)}$$

# Summary

$$p(l | \mathbf{x}_i, \Theta^g) = \frac{\alpha_l^g p_l(\mathbf{x}_i | \theta_l^g)}{\sum_{j=1}^M \alpha_j^g p_j(\mathbf{x} | \theta_j^g)}$$

## EM algorithm for GMM

Given an initial guess  $\Theta^g$ , find  $\Theta^{new}$  as follows

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)$$

$$\boldsymbol{\mu}_l^{new} = \frac{\sum_{i=1}^N \mathbf{x}_i p(l | \mathbf{x}_i, \Theta^g)}{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)}$$

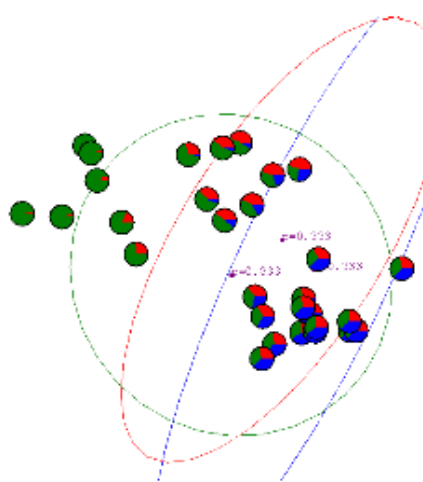
$$\boldsymbol{\Sigma}_l^{new} = \frac{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g) (\mathbf{x}_i - \boldsymbol{\mu}_l^{new})(\mathbf{x}_i - \boldsymbol{\mu}_l^{new})^T}{\sum_{i=1}^N p(l | \mathbf{x}_i, \Theta^g)}$$

$$\Theta^g \leftarrow \Theta^{new}$$

Not converge

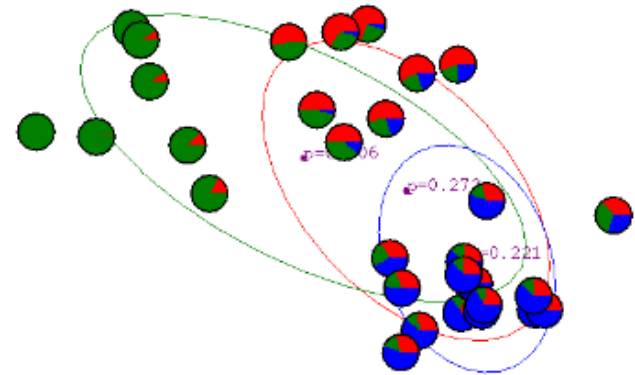


### Example: EM for GMM



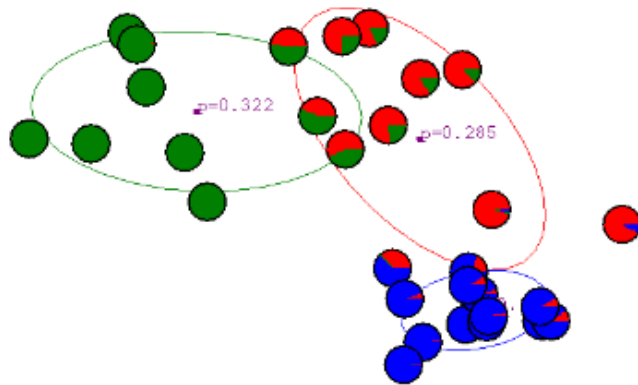
Initial model parameters.

### Example: EM for GMM



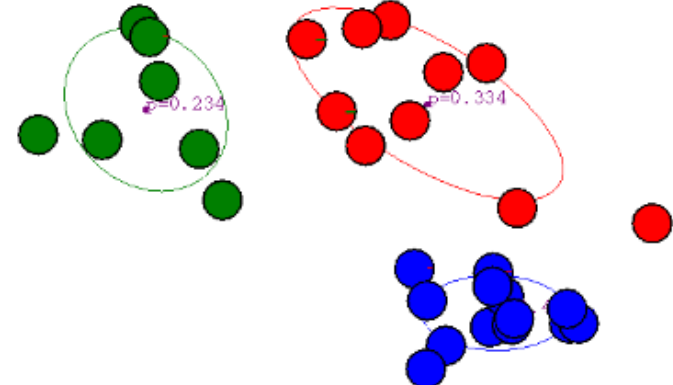
After first iteration

### Example: EM for GMM



After fifth iteration

### Example: EM for GMM



After convergence



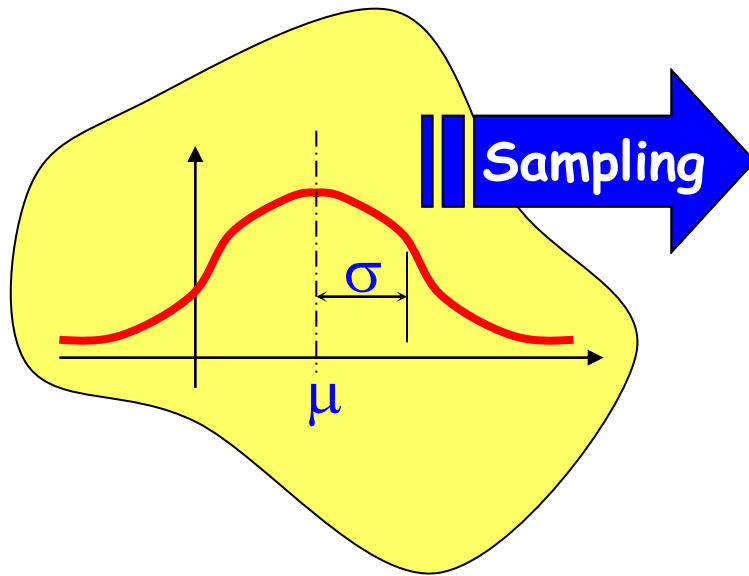
# APPENDIX

# EM Algorithm

Example:  
Missing Data

# Univariate Normal Sample

$$X \sim N(\mu, \sigma^2)$$



$$\mathbf{X} = (x_1, x_2, \dots)$$

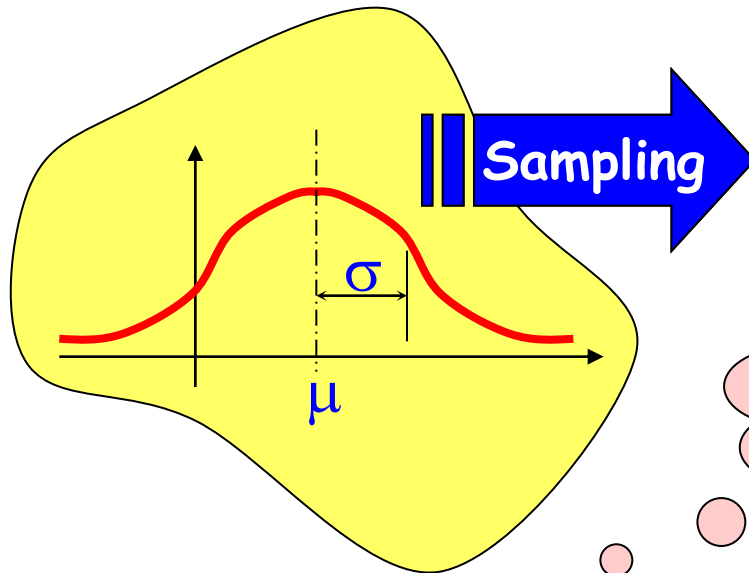
$$\hat{\mu} = ?$$

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]$$

$$\hat{\sigma}^2 = ?$$

# Maximum Likelihood

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x - \mu)}{2\sigma^2}\right]$$



$$\mathbf{X} = (x_1, x_2, \dots)$$

We want to maximize it.

$$\mathcal{L}(\mu, \sigma^2 | \mathbf{X}) = f(\mathbf{x} | \mu, \sigma^2) = f(x_1 | \mu, \sigma^2) \dots f(x_n | \mu, \sigma^2)$$

Given  $\mathbf{x}$ , it is a function of  $\mu$  and  $\sigma^2$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$


# Log-Likelihood Function

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2 \mid \mathbf{x}) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[ -\sum_{i=1}^n \frac{(x_i - \boldsymbol{\mu})^2}{2\sigma^2} \right]$$

$$\ell(\boldsymbol{\mu}, \sigma^2 \mid \mathbf{x}) = \log \mathcal{L}(\boldsymbol{\mu}, \sigma^2 \mid \mathbf{x})$$

$$= \frac{n}{2} \log \frac{1}{2\pi\sigma^2} - \sum_{i=1}^n \frac{(x_i - \boldsymbol{\mu})^2}{2\sigma^2}$$

$$= -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\boldsymbol{\mu}}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\boldsymbol{\mu}^2}{2\sigma^2}$$



Maximize  
this instead

By setting

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \sigma^2 \mid \mathbf{x}) = 0 \quad \text{and} \quad \frac{\partial}{\partial \sigma^2} \ell(\boldsymbol{\mu}, \sigma^2 \mid \mathbf{x}) = 0$$

# Max. the Log-Likelihood Function

$$\ell(\mu, \sigma^2 \mid \mathbf{X}) = -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2 \mid \mathbf{X}) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu}{\sigma^2} = 0 \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

# Max. the Log-Likelihood Function

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \hat{\mu}^2$$

$$\ell(\mu, \sigma^2 \mid \mathbf{X}) = -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}$$

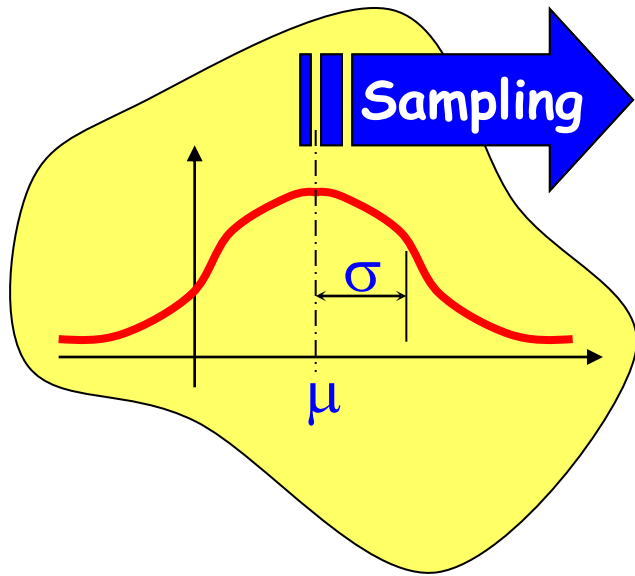
$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2 \mid \mathbf{X}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n x_i^2 - \frac{\mu}{\sigma^4} \sum_{i=1}^n x_i + \frac{n\mu^2}{2\sigma^4} = 0$$

$$\begin{aligned} n\sigma^2 &= \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{2}{n} \left( \sum_{i=1}^n x_i \right)^2 + \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \end{aligned}$$

# Miss Data

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \hat{\mu}^2$$



$$\mathbf{X} = (x_1, \dots, \underbrace{\dots, 1, \dots}_{\text{Missing data}})$$

$$\hat{\mu} = \frac{1}{n} \left( \sum_{i=1}^m x_i + \sum_{j=m+1}^n x_j \right)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^m x_i^2 + \sum_{j=m+1}^n x_j^2 \right) - \hat{\mu}^2$$



# E-Step

$$\hat{\mu} = \frac{1}{n} \left( \sum_{i=1}^m x_i + \sum_{j=m+1}^n x_j \right)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^m x_i^2 + \sum_{j=m+1}^n x_j^2 \right) - \hat{\mu}^2$$

Let  $\mu^{(t)}$   
 $\sigma^{2(t)}$  be the estimated parameters at the initial of the  $t^{\text{th}}$  iterations

$$E_{\hat{\mu}^{(t)}, \sigma^{2(t)}} \left[ \sum_{j=m+1}^n x_j \mid \mathbf{X} \right] = (n - m) \hat{\mu}^{(t)}$$

$$E_{\hat{\mu}^{(t)}, \sigma^{2(t)}} \left[ \sum_{j=m+1}^n x_j^2 \mid \mathbf{X} \right] = (n - m) \left( \hat{\mu}^{(t)2} + \sigma^{2(t)} \right)$$

$$\hat{\mu} = \frac{1}{n} \left( \sum_{i=1}^m x_i + \sum_{j=m+1}^n x_j \right)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^m x_i^2 + \sum_{j=m+1}^n x_j^2 \right) - \hat{\mu}^2$$

Let  $\mu^{(t)}$  be the estimated parameters at the initial of the  $t^{\text{th}}$  iterations

$$E_{\hat{\mu}^{(t)}, \hat{\sigma}^{2(t)}} \left[ \sum_{j=m+1}^n x_j \mid \mathbf{X} \right] = (n - m) \hat{\mu}^{(t)}$$

$$s_1^{(t)} = \sum_{i=1}^m x_i + (n - m) \hat{\mu}^{(t)}$$

$$E_{\hat{\mu}^{(t)}, \hat{\sigma}^{2(t)}} \left[ \sum_{j=m+1}^n x_j^2 \mid \mathbf{X} \right] = (n - m) \left( \hat{\mu}^{(t)2} + \hat{\sigma}^{2(t)} \right)$$

$$s_2^{(t)} = \sum_{i=1}^m x_i^2 + (n - m) \left( \hat{\mu}^{(t)2} + \hat{\sigma}^{2(t)} \right)$$

# M-Step

$$\hat{\mu} = \frac{1}{n} \left( \sum_{i=1}^m x_i + \sum_{j=m+1}^n x_j \right) \quad \hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^m x_i^2 + \sum_{j=m+1}^n x_j^2 \right) - \hat{\mu}^2$$

Let  $\mu^{(t)}$  and  $\sigma^2^{(t)}$  be the estimated parameters at the initial of the  $t^{\text{th}}$  iterations

$$\mu^{(t+1)} = \frac{S_1^{(t)}}{n}$$

$$S_1^{(t)} = \sum_{i=1}^m x_i + (n-m)\hat{\mu}^{(t)}$$

$$\sigma^2^{(t+1)} = \frac{S_2^{(t)}}{n} - \hat{\mu}^{(t+1)2}$$

$$S_2^{(t)} = \sum_{i=1}^m x_i^2 + (n-m) \left( \hat{\mu}^{(t)2} + \sigma^2^{(t)} \right)$$

# Exercise

$$X \sim N(\mu, \sigma^2)$$

$n = 40$  (10 data missing)

Estimate  $\mu, \sigma^2$  using different initial conditions.

375.081556	243.548664	454.981077
362.275902	382.789939	479.685107
332.612068	374.419161	336.634962
351.383048	337.289831	407.030453
304.823174	418.928822	297.821512
386.438672	364.086502	311.267105
430.079689	343.854855	528.267783
395.317406	371.279406	419.841982
369.029845	439.241736	392.684770
365.343938	338.281616	301.910093