Expectation-Maximization (EM) Algorithm

Contents

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)

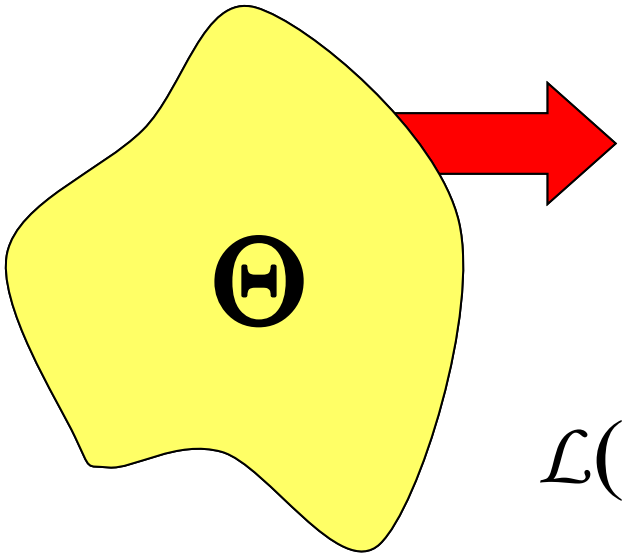
Applications

- 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

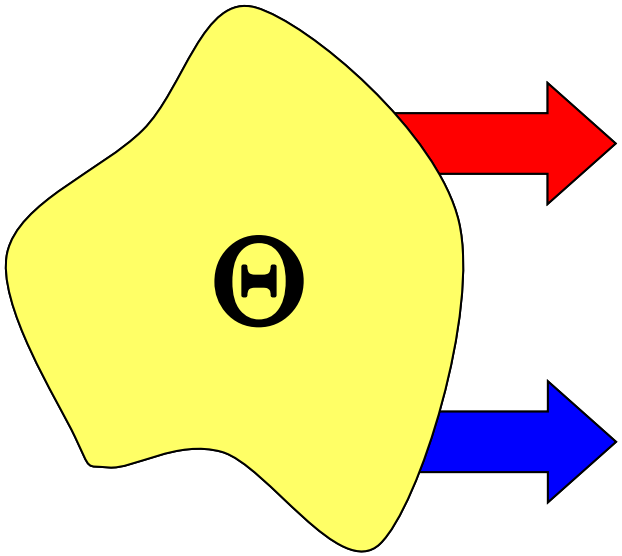


$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

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

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

Latent Variables



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

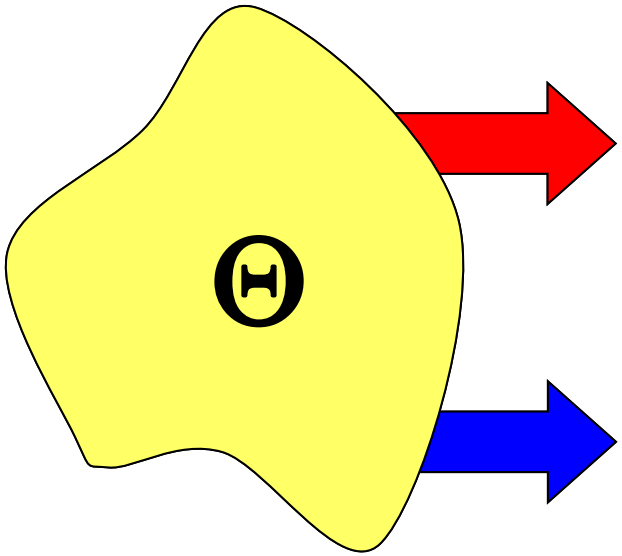
Incomplete Data

$$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$$

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

Complete Data Likelihood

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



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

$$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$$

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

$$= p(\mathbf{Y} \mid \mathbf{X}, \Theta)\, p(\mathbf{X}|\Theta)$$

Complete Data Likelihood

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

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

A function of random variable **Y**.

A function of latent variable **Y** and parameter **Θ**

A function of parameter **Θ**

If we are given **Θ**,

The result is in term of random variable **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:

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

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

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

Expectation Step

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

Define (Conditional Expectation of log likelihood of complete data)

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

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

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

Maximization Step

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

Define

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

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

EM Algorithm

Mixture Model

Mixture Models

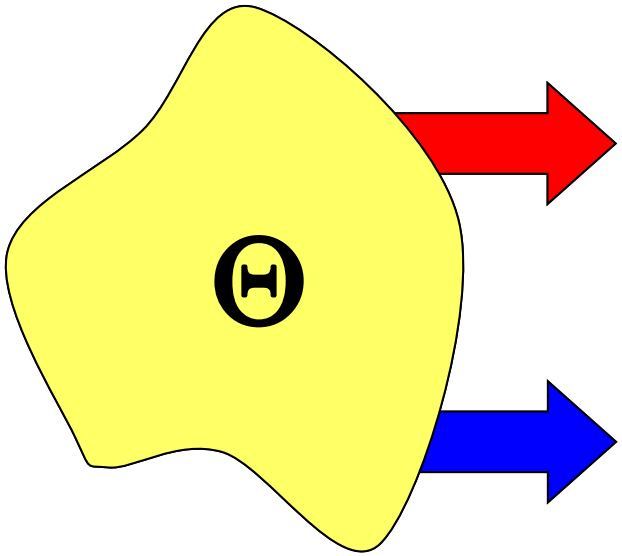
- If there is a reason to believe that a data set is comprised of several distinct populations, a mixture model can be used.
- It has the following form:

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

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

Mixture M

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



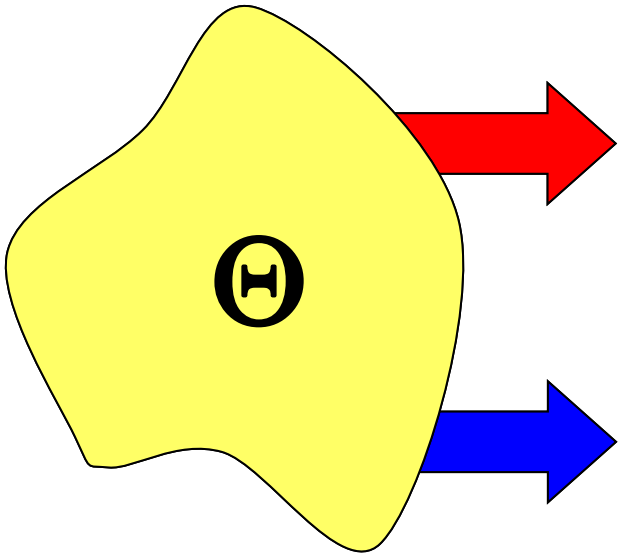
$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

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

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

Mixture Models

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



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

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

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

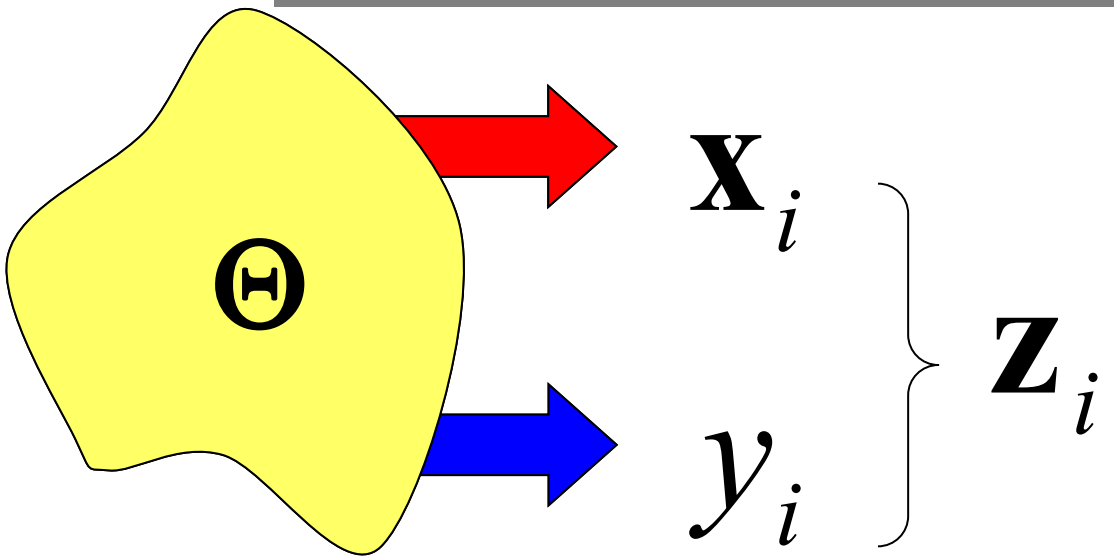
Mixture Models

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

$$p(y_i \mid \mathbf{x}_i, \Theta) = \frac{p(\mathbf{x}_i, y_i, \Theta)}{p(\mathbf{x}_i, \Theta)} = \frac{p(\mathbf{x}_i \mid y_i, \Theta) p(y_i, \Theta)}{[illegible]}$$

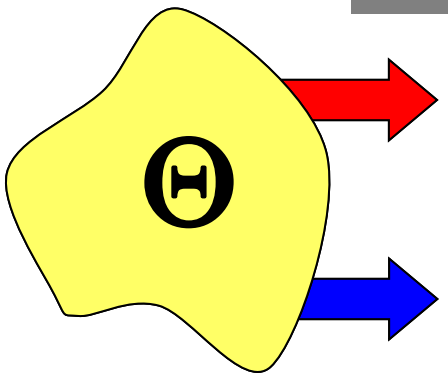
$$= \frac{p(\mathbf{x}_i \mid y_i, \Theta)}{p(\mathbf{x}_i \mid}$$

Given **x** and **Θ**, the conditional density of *y* can be computed.

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

Complete-Data Likelihood Function

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



$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \qquad \mathbf{y} = \{y_1, \ldots, y_N\}$$

$$\mathbf{z}_i = (\mathbf{x}_i, y_i) \qquad \mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$$

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

$$= \prod_{i=1}^{N} p(\mathbf{x}_i \mid \underbrace{y_i}_{\theta_{y_i}}, \Theta) \underbrace{p(y_i|\Theta)}_{\alpha_{y_i}} = \prod_{i=1}^{N} \alpha_{y_i} p_{y_i}(\mathbf{x}_i \mid \theta_{y_i})$$

Expectation

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

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

$$Q(\Theta, \Theta^g) = E[\log \mathcal{L}(\Theta \mid \mathbf{Z}) \mid \mathbf{X}, \Theta^g]$$

Θ^g: Guess

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

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

Expectation

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

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

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^g) = E[\log \mathcal{L}(\boldsymbol{\Theta} \mid \mathbf{Z}) \mid \mathbf{X}, \boldsymbol{\Theta}^g]$$

$\boldsymbol{\Theta}^g$: Guess

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

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

$$= \sum_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{y_i, l} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \prod_{j=1}^{N} p(y_j \mid \mathbf{x}_j, \boldsymbol{\Theta}^g)$$

Expectation

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

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

Zero when $y_i \neq l$

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

Expectation

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \sum_{\mathbf{y} \in \mathcal{Y}} \delta_{y_i, l} \prod_{j=1}^{N} p(y_j \mid \mathbf{x}_j, \boldsymbol{\Theta}^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \sum_{y_1=1}^{M} \cdots \sum_{y_i=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{y_i, l} \prod_{j=1}^{N} p(y_j \mid \mathbf{x}_j, \boldsymbol{\Theta}^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \left(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{\substack{j=1 \\ j \neq i}}^{N} p(y_j \mid \mathbf{X}, \boldsymbol{\Theta}^g) \right) p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)$$

Expectation

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \sum_{\mathbf{y} \in \mathcal{Y}} \delta_{y_i, l} \prod_{j=1}^{N} p(y_j \mid \mathbf{x}_j, \boldsymbol{\Theta}^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \sum_{y_1=1}^{M} \cdots \sum_{y_i=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{y_i, l} \prod_{j=1}^{N} p(y_j \mid \mathbf{x}_j, \boldsymbol{\Theta}^g)$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(\mathbf{x}_i \mid \theta_l)\right] \left(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{\substack{j=1 \\ j \neq i}}^{N} p(y_j \mid \mathbf{X}, \boldsymbol{\Theta}^g) \right) p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)$$

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

Expectation

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

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

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

$$\overbrace{\qquad\qquad}^{1}$$

Maximization

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

Given the initial guess Θ^g,

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

We want to find Θ, to maximize the above expectation.

In fact, iteratively.

EM Algorithm

EM-Algorithm on GMM

The GMM (Guassian Mixture Model)

Guassian model of a d-dimensional source, say j :

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

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

GMM with M sources:

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

$$\alpha_j \geq 0$$

$$\sum \alpha_j = 1$$

Goal

Mixture Model

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

$$\boldsymbol{\Theta} = (\alpha_1, \ldots, \alpha_M, \theta_1, , \ldots, \theta_M)$$

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

To maximize:

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

Goal

Mixture Model

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

Correlated with α_l only.

Correlated with θ_l only.

$\Theta=(\alpha_1,\cdots$

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

To maximize:

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

Finding α_l

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

Due to the constraint on α_l's, we introduce *Lagrange Multiplier* λ, and solve the following equation.

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

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

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

Finding α_l

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

$$\lambda = -N$$

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

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

$$\underbrace{\underbrace{\sum_{l=1}^{M} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)}_{1}}_{N} \qquad \underbrace{\lambda \sum_{l=1}^{M} \alpha_l}_{1}$$

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

Finding α_l

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

$$\lambda=-N$$

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

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

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

Finding θ_l

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

Consider GMM

Only need to maximize this term

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

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

$$\log[\,p_l(\mathbf{x}\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)]=\underbrace{-\tfrac{d}{2}\log 2\pi}-\tfrac{1}{2}\log\mid\boldsymbol{\Sigma}_l\mid^{1/2}-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)$$

unrelated

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

Only need to maximize this term

Therefore, we want to maximize:

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

How? knowledge on matrix algebra is needed.

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

unrelated

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

Therefore, we want to maximize:

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

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

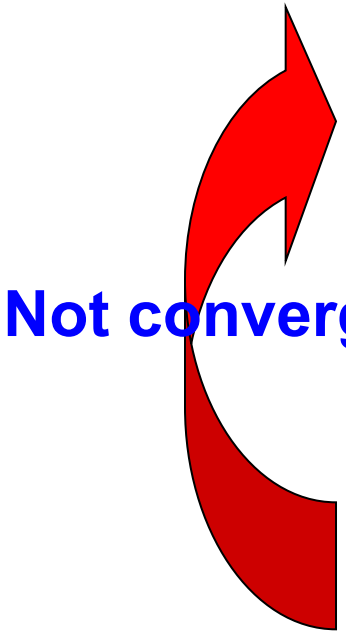
$$\Sigma_l = \frac{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)(\mathbf{x}_i - \boldsymbol{\mu}_l)(\mathbf{x}_i - \boldsymbol{\mu}_l)^T}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \boldsymbol{\Theta}^g)}$$

Summary

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

EM algorithm for GMM

Given an initial guess Θ^g, find Θ^{new} as follows



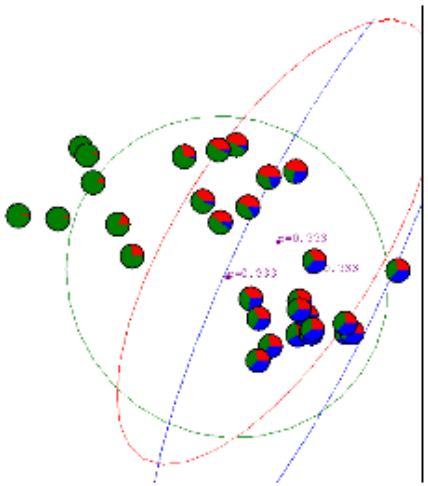
$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \Theta^g)$$

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

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \Theta^g)(\mathbf{x}_i - \mu_l^{new})(\mathbf{x}_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l \mid \mathbf{x}_i, \Theta^g)}$$

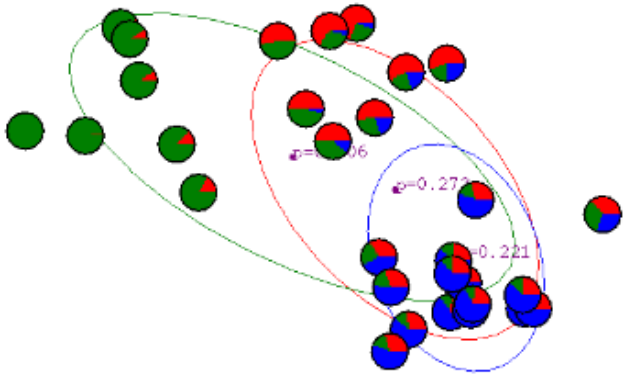
$$\Theta^g \leftarrow \Theta^{new}$$

Example: EM for GMM



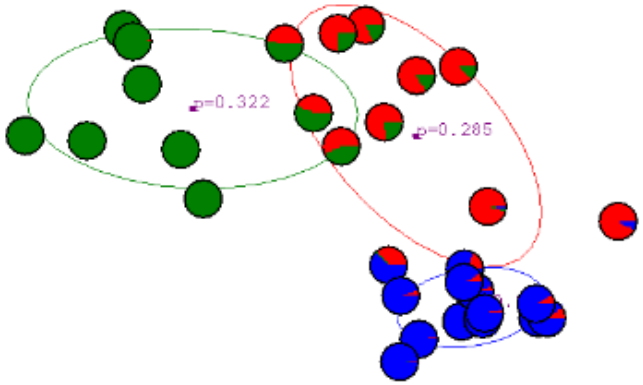
Initial model parameters.

Example: EM for GMM



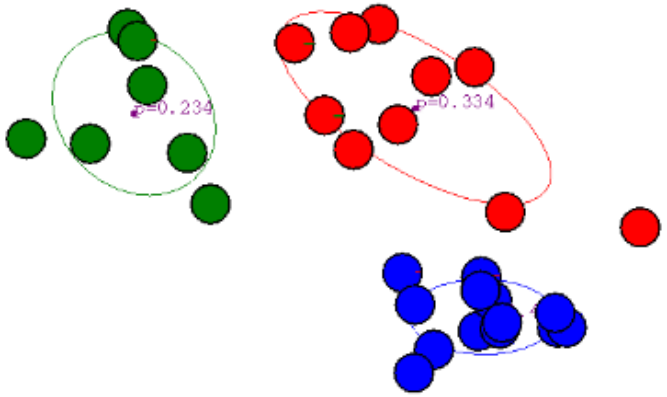
After first iteration

Example: EM for GMM



After fifth iteration

Example: EM for GMM



After convergence

Susanna Ricco

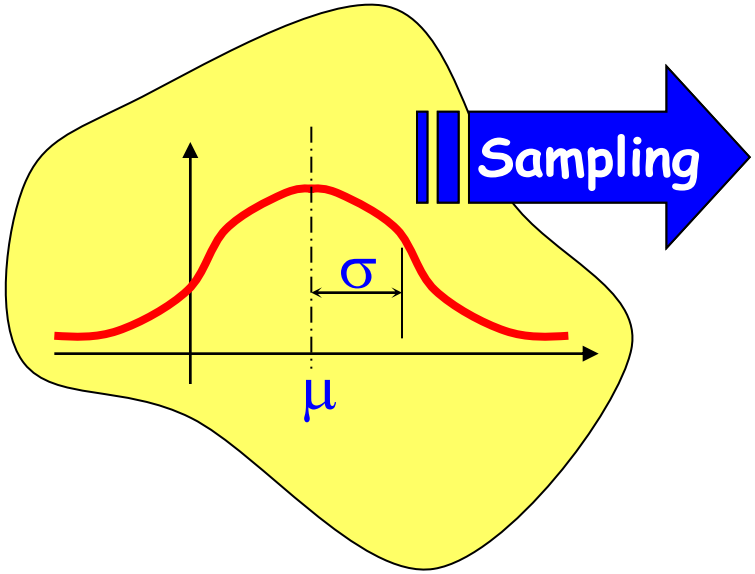
APPENDIX

EM Algorithm

Example: Missing Data

Univariate Normal Sample

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



$$\mathbf{x} = (x_1, x_2, \ldots$$

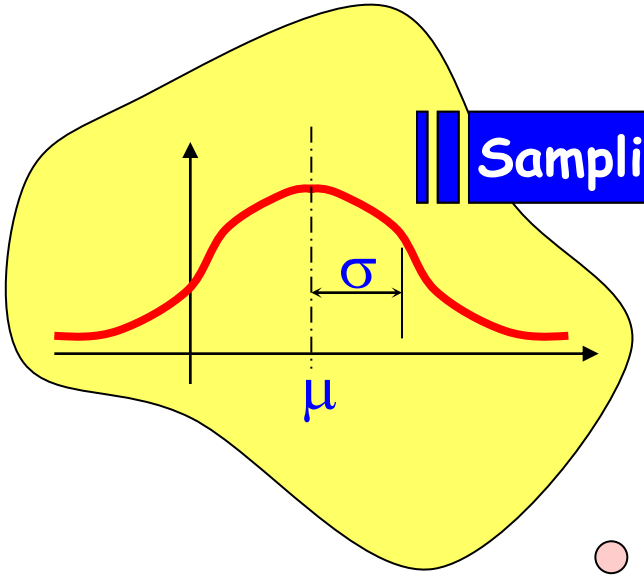
$$\hat{\mu} = ?$$

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

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

Maximum Likelihood

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



$$\mathbf{x} = (x_1, x_2, \ldots$$

We want to maximize it.

$$\mathcal{L}(\mu, \sigma^2 \mid \mathbf{x}) = f(\mathbf{x} \mid \mu, \sigma^2) = f(x_1 \mid \mu, \sigma^2) \cdots \quad \iota, \sigma^2)$$

Given **x**, it is a function of μ and σ^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

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

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

Maximize this instead

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

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

By setting

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

Max. the Log-Likelihood Function

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

$$\frac{\partial}{\partial\mu} l(\mu, \sigma^2 \mid \mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu}{\sigma^2} = 0$$

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

Max. the Log-Likelihood Function

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

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

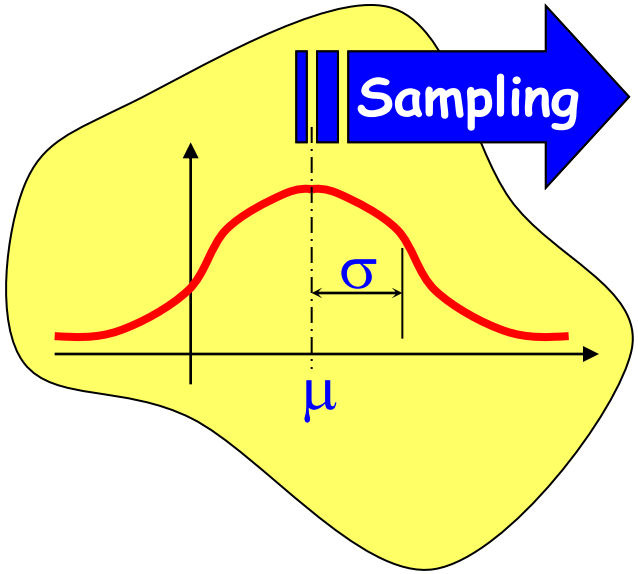
$$\frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2 \mid \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} x_i^2 - \frac{\mu}{\sigma^4}\sum_{i=1}^{n} x_i + \frac{n\mu^2}{2\sigma^4} = 0$$

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

$$= \sum_{i=1}^{n} x_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

Miss Data

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



$$\mathbf{x} = (x_1, \ldots \underbrace{\ldots, \ldots}_{\text{Missing data}} \ldots)$$

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

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

E-Step

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

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

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

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

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

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

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

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

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

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

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

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

M-Step

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

Let $\mu^{(t)}$, $\sigma^{2(t)}$ be the estimated parameters at the initial of the t^{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 μ, σ^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