# Exponential Family and Maximum Likelihood, Gaussian Mixture Models and the EM Algorithm

by Korbinian Schwinger

# Overview

- Exponential Family
- Maximum Likelihood
- The EM Algorithm
- Gaussian Mixture Models

# Exponential Family

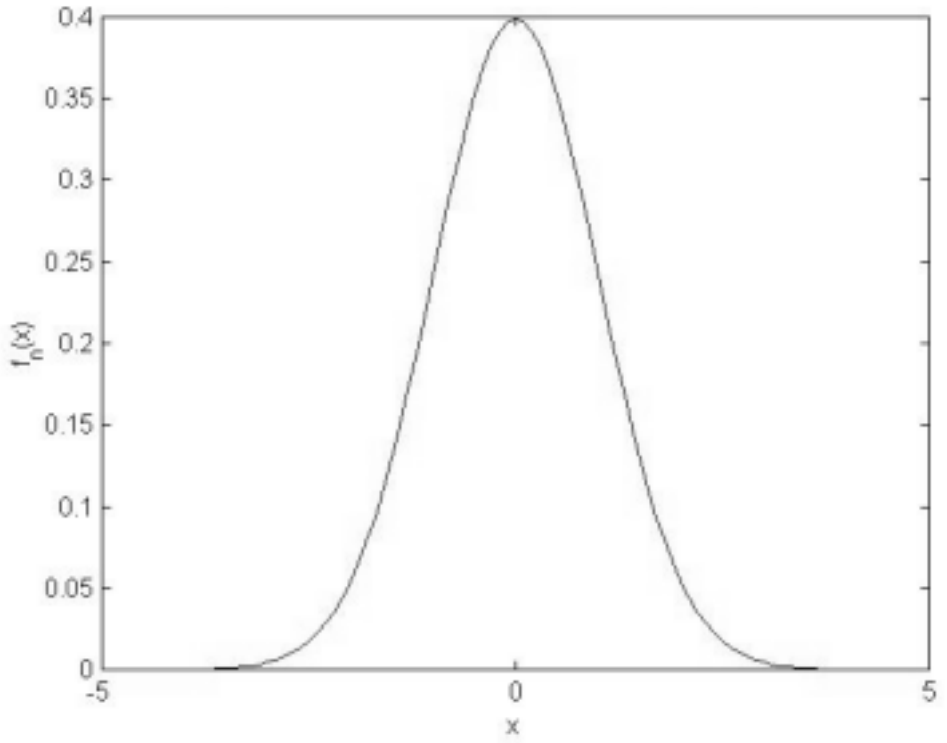
- A density function of a random variable x is member of the exponential family if the distribution function can be written as :

$$f(x|\theta) = h(x)\exp\left(b(\theta) + \sum_{i=1}^{n} w_i(\theta)\, t_i(x)\right)$$

# Gauss in 1 dimension

- Density function with the Parameters $\mu=0$ and $\sigma^2=1$

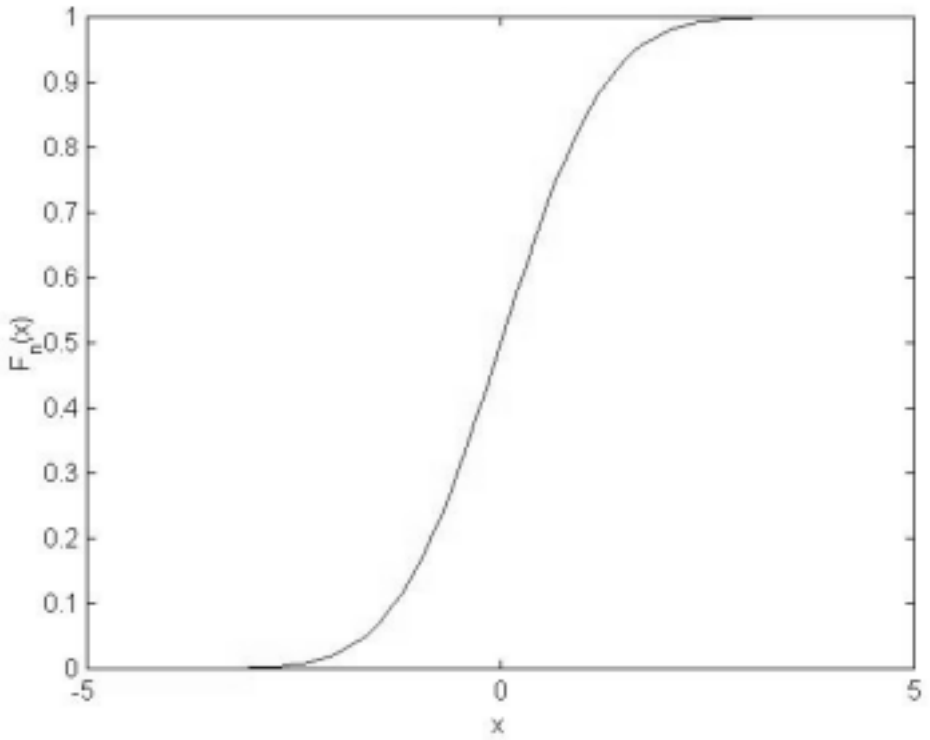
$$f_x(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$



# Gauss in 1 dimension

- Distribution function with the Parameters $\mu=0$ and $\sigma^2=1$

$$F_x(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



# Gauss in 2 dimensions

- μ becomes a vector, the variance becomes the covariance-matrix, which shows the relationship between the 2 Coordinates :

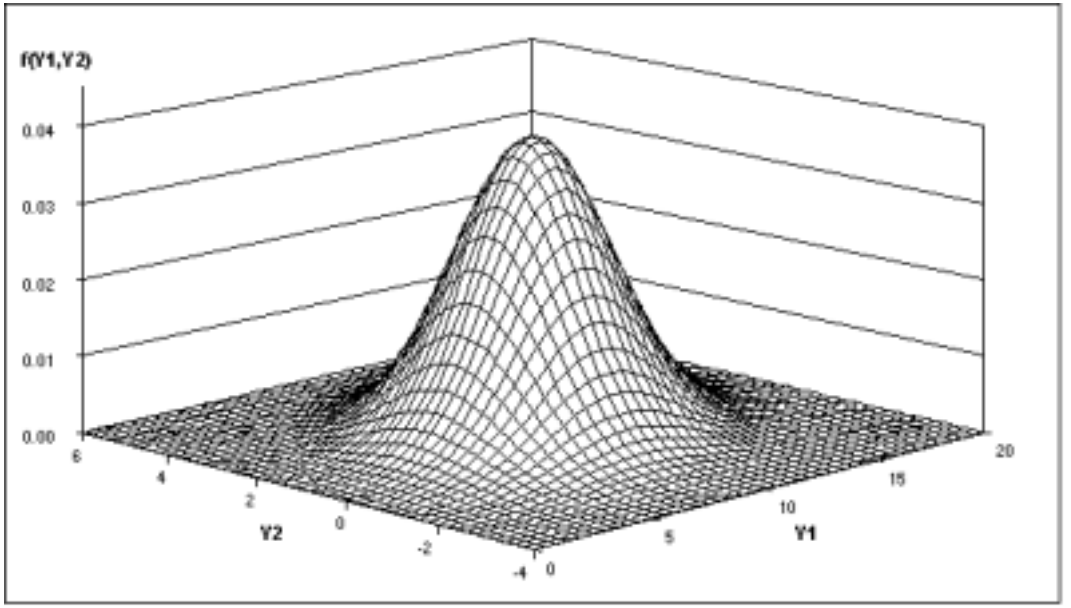
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$$

# Gauss in 2 dimensions

- Density with the Parameters :

$$\mu = \begin{pmatrix} 10 \\ 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 9 & 1 \\ 1 & 2 \end{pmatrix}$$



# Gauss in n-dimensions

- The parameters become :

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{nn} \end{pmatrix}$$

$$\sigma_{ij} = E\left[(x_i - \mu_i)(x_j - \mu_j)\right]$$

# Maximum Likelihood

- We have a density function p(x|Θ) which depends on parameters in Θ
- We want to find out, which parameters in Θ are the most probable
- First, we need a set of data X of size N
- With this set we can create a new density function

$$p(X|\Theta)=\prod_{i=1}^{N} p(x_i|\Theta)=L(\Theta|X)$$

# Maximum Likelihood

- Best set of parameters maximizes the likelihood function

$$\Theta^{max} = \underset{\Theta}{argmax}\, L(\Theta|X)$$

- Often log(L(Θ|X)) is maximized
- If distribution is Gaussian the maximum can be found with the derivative

# Maximum Likelihood - example

- We have got a random variable with a gaussian distribution with the Parameters μ and σ²
- The likelihood function is

$$L(\bar{x};\mu,\sigma^2)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\prod_{i=1}^{n}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

# Maximum Likelihood - example

- Now we have to calculate the logaritm of the likelihood function

$$\ln\left(L\left(\bar{x}\,;\mu\,,\sigma^2\right)\right)=-n\left(\ln\sqrt{2\pi}+\ln\sigma\right)+\sum_{i=1}^{n}\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)$$

# Maximum Likelihood - example

- Next step is to create the derivative of the log-likelihood

$$\frac{\delta \ln(L)}{\delta \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \stackrel{!}{=} 0$$

$$\frac{\delta \ln(L)}{\delta \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} \stackrel{!}{=} 0$$

# Maximum Likelihood - example

- At the end we have got :

$$\mu = \bar{x}$$

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

# Basic EM

- Used to find the maximum-likelihood estimate when features are missing
  - Limitations of the observation process
  - calculation can be simplified
- First, we assume the observed data X are generated by some distribution
- Second, we assume, that a complete set of data Z=X+Y exists

# Basic EM

- Then we can create a density function :

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta)$$

- With this density function, the complete-data likelihood, p(X,Y|Θ), can be formed
- Then the Algorithm calculates the expected value of the complete-data log-likelihood

$$Q(\Theta, \Theta^{(i-1)}) = E[\log(p(X, Y|\Theta))|X, \Theta^{(i-1)}]$$

$$= \int_{y \in Y} \log(p(X, y|\Theta)\, f(y|X, \Theta^{i-1}))\, dy$$

# Basic EM

- The evaluation of the expectation is called the E-step
- The M-step is to maximize the expectation, we computed in the first step :

$$\Theta^{(i)} = \underset{\Theta}{argmax}\, Q(\Theta, \Theta^{(i-1)})$$

- These steps are repeated as necessary – each iteration is guaranteed to increase the likelihood

# Basic EM - example

- We have got 4 Points in 2D Coordinates :

$$\left\{\begin{pmatrix}0\\2\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}?\\4\end{pmatrix}\right\}$$

- We assume, the model is a Gaussian :

$$\theta=\begin{pmatrix}\mu_1\\\mu_2\\\sigma_1^2\\\sigma_2^2\end{pmatrix}$$

# Basic EM - example

- The first estimation, we get with the assumption of an Gaussian model :

$$\theta^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

# Basic EM - example

- Now we calculate the expected value :

$$Q(\theta,\theta^0)=E[\ln(p(x_g,x_b;\theta|\theta^0;D_g))]$$

$$=\int_{-\infty}^{\infty}[\sum_{k=1}^{3}\ln(p(x_k|\theta))+\ln(p(x_4|\theta))]\,p(x_{41}|\theta^0;x_{42}=4)\,dx_{41}$$

$$=\sum_{k=1}^{3}[\ln(p(x_k|\theta))]+\int_{-\infty}^{\infty}\ln\left(p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}|\theta\right)\right)\underbrace{\frac{p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}|\theta^0\right)}{\left(\int_{-\infty}^{\infty}p\left(\begin{pmatrix}x'_{41}\\4\end{pmatrix}|\theta^0\right)dx'_{41}\right)}}_{=K}dx_{41}$$

# Basic EM - example

- K is constant and can be brought out of the integral :

$$Q(\theta;\theta^0)=\sum_{k=1}^{3}[\ln(p(x_k|\theta))]+\frac{1}{K}\int_{-\infty}^{\infty}\ln\left(p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}|\theta\right)\right)*$$

$$*\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}\exp([-\frac{1}{2}(x_{41}^2+4^2)])\,dx_{41}$$

# Basic EM - example

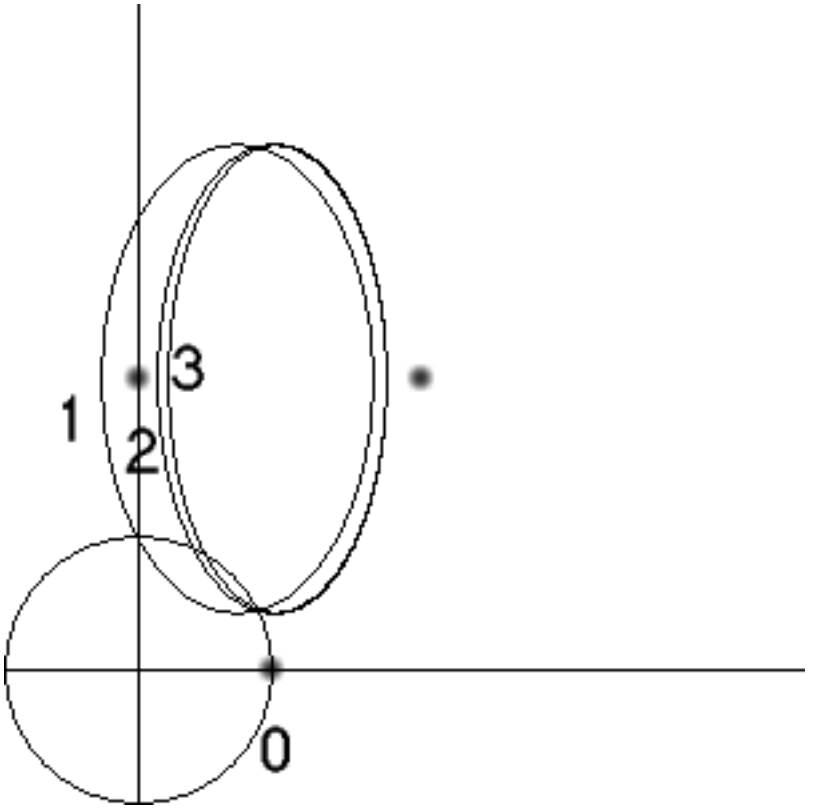
- At the end we get (E step completed)

$$Q(\theta;\theta^0)=\sum_{k=1}^{3}[\ln(p(x_k|\theta))]-\frac{1+\mu_1^2}{2\sigma_1^2}-\frac{(4-\mu_2)^2}{2\sigma_2^2}-\ln(2\pi\sigma_1\sigma_2)$$

- Then, we can calculate the parameters, which maximize Q(.|.)

$$\theta^1=\begin{pmatrix}0.75\\2.0\\0.938\\2.0\end{pmatrix}$$

# Basic EM - example



# Gaussian Mixture Models

- We assume the following probilistic model:

$$p(x|\Theta)=\sum_{i=1}^{M}\alpha_i\, p_i(x|\theta_i)$$

- The parameters are $\Theta=(\alpha_1,\ldots,\alpha_M,\theta_1,\ldots,\theta_M)$ and $\sum_{i=1}^{M}\alpha_i=1$
- With a lot of calculations we get the function

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

# Gaussian Mixture Models

- After some more pages of calculations you get a new estimation of the parameters :

$$\alpha_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i, \Theta^g) \qquad \mu_l^{new} = \frac{\sum_{i=1}^{N} x_i\, p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

$$\sum\nolimits_l^{new} = \frac{\sum_{i=1}^{N} p(l|x_i, \Theta^g)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

- These values can be used to repeat the iteration as often as needed

# Gaussian Mixture Models - example

Java Demo Applet