EXPECTATION MAXIMIZATION ALGORITHM

Lecture Notes

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

Some concepts and illustrations in this lecture are adapted from the textbooks,
Pattern Recognition and Machine Learning, C. M. Bishop, *Springer*, 2006.

EM Algorithm

- Say that the probability of the temperature outside your window for each of the 24 hours of a day $\mathbf{z} \in R^{24}$ depends on the season $\lambda \in \{$ summer, fall, winter, spring$\}$, and that you know the seasonal temperature distribution $p(\mathbf{z}|\lambda)$.
- But say you can only measure the average temperature $\mathbf{y} = T(\mathbf{z})$ for the day, and you'd like to guess what season λ it is (for example, is spring here yet?).
- The maximum likelihood estimate of λ maximizes $p(y|\lambda)$, but in some cases this may be hard to find.
- That's when EM is useful – it takes your observed data $\mathbf{y}$, iteratively makes guesses about the complete data $\mathbf{z}$, and iteratively finds the that maximizes $p(\mathbf{z}|\lambda)$ over λ.
- In this way, EM tries to find the maximum likelihood estimate of λ given $\mathbf{y}$.

List of Items for EM Algorithm

- Some observed data $\mathbf{y}$ with a density $p(\mathbf{y}|\lambda)$
- A description of some data $\mathbf{z}$ that you wish you had, with a density $p(\mathbf{z}|\lambda)$
- You do not observe $\mathbf{z}$ directly; instead, you observe $\mathbf{y} = T(\mathbf{z})$ for some function T.
- λ parameter you want to estimate

EM Algorithm

$$\hat{\lambda}_{MLE} = \arg\max_{\lambda \in \Lambda} \log p(\mathbf{y}|\lambda)$$

Step 1 Pick an initial guess $\lambda^k = \lambda_0$ with $k = 0$.

Step 2 Given the observed data $\mathbf{y}$ calculate how likely it is that the complete data is exactly $\mathbf{z}$, that is, the conditional distribution $p(\mathbf{z}|\mathbf{y}, \lambda^k)$.

Step 3 Compute the Q function $\int \log p(\mathbf{z}, \mathbf{y}|\lambda) p(\mathbf{z}|\mathbf{y}, \lambda^k) d\theta$ which is the expected log-likelihood of $p(\mathbf{z}|\lambda)$ with respect to $p(\mathbf{z}|\mathbf{y}, \lambda^k)$.

Step 4 Make a new guess λ^{k+1} for λ that maximizes the Q or (the expected) log-likelihood of $p(\mathbf{z}|\lambda)$.

$$\lambda^{m+1} = \arg\max_{\lambda} Q(\lambda|\lambda^k) = \int \log p(\mathbf{z}, \mathbf{y}|\lambda) p(\mathbf{z}|\mathbf{y}, \lambda^k) d\mathbf{z} = \int \log p(\mathbf{z}, \mathbf{y}|\lambda) \frac{p(\mathbf{z}|\lambda^k)}{p(\mathbf{y}|\lambda^k)} d\mathbf{z}$$

since $p(\mathbf{z}|\mathbf{y}, \lambda) = \frac{p(\mathbf{z},\mathbf{y}|\lambda)}{p(\mathbf{y}|\lambda)} = \frac{p(\mathbf{z}|\lambda)}{p(\mathbf{y}|\lambda)}$ and T is a deterministic function *i.e.* knowing $\mathbf{z}$ means you also know $\mathbf{y}$.

Step 5 Repeat 2 to 4 until convergence.

Ector's Problem

Let the random variable X_1, represent the number of round dark objects, X_2, represent the number of square dark objects, and X_3, represent the number of light objects.
Let $\mathbf{x} = [x_1, x_2, x_3]^T$ be the vector of values the random variables take for some image.

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \left(\frac{n!}{x_1!x_2!x_3!}\right)\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{4}+\frac{p}{4}\right)^{x_2}\left(\frac{1}{2}-\frac{p}{4}\right)^{x_3}$$

where p is an unkown parameter and $n = x_1 + x_2 + x_3$.

Ector's Problem

Let $\mathbf{y} = [y_1, y_2]^T$ be the number of dark objects and number of light objects detected, respectively, so that $y_1 = x_1 + x_2$ and $y_2 = x_3$ and let the corresponding random variables be Y_1, and Y_2. The likelihood is

$$P(Y_1 = y_1, Y_2 = y_2|p) = \begin{pmatrix} n \\ y_1 \end{pmatrix} \left(\frac{1}{2} + \frac{p}{4}\right)^{y_1} \left(\frac{1}{2} - \frac{p}{4}\right)^{y_2}$$

Ector's Problem

If (X_1, X_2, X_3) has a multinomial distribution with probabilities (p_1, p_2, p_3) then

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \left(\frac{n!}{x_1!x_2!x_3!}\right) p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

$$\begin{aligned} P(X_1 + X_2 = y_1, X_3 = x_3) &= \sum_{i=0}^{y_1} P(X_1 = i, X_2 = y_1 - i, X_3 = x_3) \\ = \sum_{i=0}^{y_1} \frac{n!}{i!(y_1 - i)!x_3!} p_1^i p_2^{y_1 - i} p_3^{x_3} &= \frac{n!}{y_1!x_3!} p_3^{x_3} \sum_{i=0}^{y_1} \frac{y_1!}{i!(y_1 - i)!} p_1^i p_2^{y_1 - i} \\ = \frac{n!}{y_1!x_3!}(p_1 + p_2)^{y_1} p_3^{x_3} &= P(Y_1 = y_1, Y_2 = y_2 | \mathbf{p^k}) \end{aligned}$$

Expectation Step

We assume the latent variables x_1 and x_2 and compute their conditional expectations;
$x_1^{k+1} = E[x_1|y_1, p^k]$ and $x_1^{k+1} = E[x_2|y_2, p^k]$.
x_3 is directly observed since $x_3 = y_3$.

Ector's Problem

To compute $x_1^{k+1} = E[x_1|y_1, y_2, p^k]$ we first determine

$$
\begin{aligned}
P(X_1 = x_1|Y_1 = y_1, Y_2 = y_2, \mathbf{p}^k) = & \quad \frac{P(X_1{=}x_1, Y_1{=}y_1, Y_2{=}y_2|\mathbf{p}^k)}{P(Y_1{=}y_1, Y_2{=}y_2|\mathbf{p}^k)} \\
= & \quad \frac{P(X_1{=}x_1, X_2{=}y_1{-}x_1, X_3{=}x_3|\mathbf{p}^k)}{P(Y_1{=}y_1, Y_2{=}y_2|\mathbf{p}^k)} \\
= & \quad \frac{y_1!x_3!}{n!(p_1{+}p_2)^{y_1}p_3^{x_3}} \frac{n!p_1^{x1}p_2^{y_1-x_1}p_3^{x_3}}{x_1!(y_1{-}x_1)!x_3!} \\
= & \quad \frac{y_1!}{x_1!(y_1{-}x_1)!}p_1^{x_1}p_2^{y_1-x_1}\frac{1}{(p_1{+}p_2)^{y_1}}
\end{aligned}
$$

Ector's Problem

Therefore,

$$x_1^{k+1} = E[x_1|y_1, y_2, p^k] = \sum_{x_1=0}^{y_1} x_1 \, P(X_1 = x_1|Y_1 = y_1, Y_2 = y_2, \mathbf{p}^k)$$

$$= \sum_{x_1=0}^{y_1} x_1 \frac{y_1!}{x_1!(y_1 - x_1)!} p_1^{x_1} p_2^{y_1 - x_1} \frac{1}{(p_1 + p_2)^{y_1}}$$

$$E[x_1|y_1, y_2, p^k] = y_1 \frac{p_1}{p_1 + p_2}$$

Self-Study Question: Show that

$$x_2^{k+1} = E[x_2|y_1, y_2, p^k] = y_1 \frac{p_2}{p_1 + p_2}$$

Ector's Problem

Maximization Step

We maximize the log-likelihood with respect to the unknown parameter p,

$$\frac{d}{dp} \log P(X_1 = x1, X_2 = x_2, X_3 = x_3) =$$

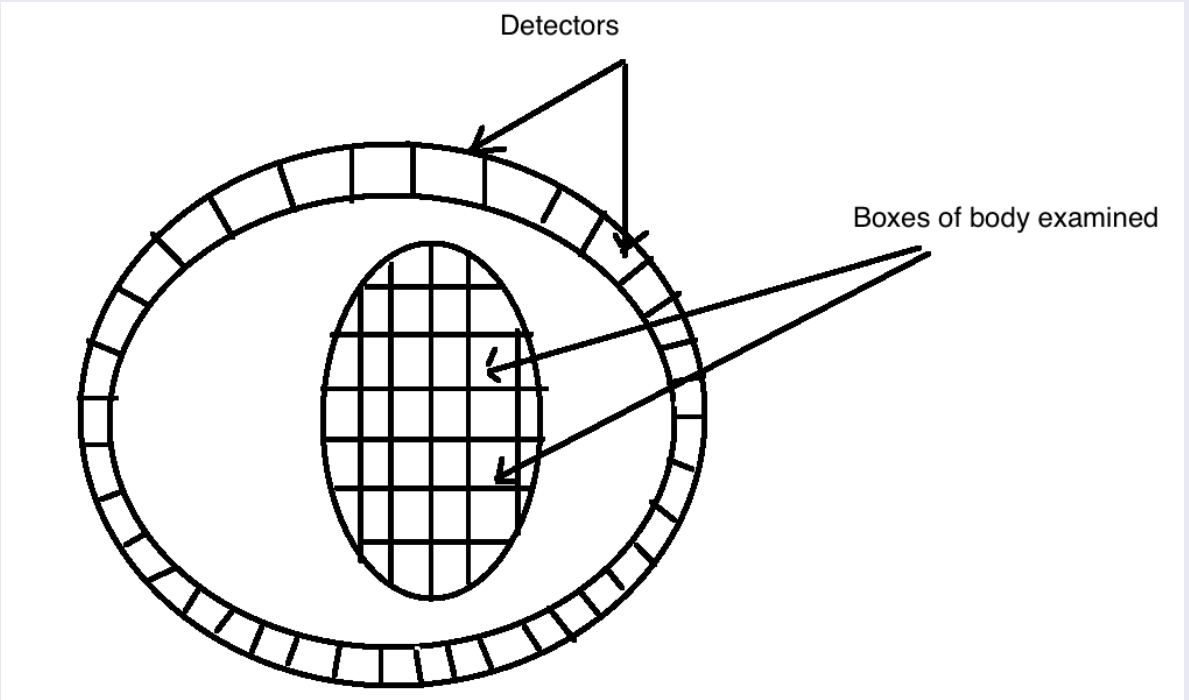
$$\frac{d}{dp} \log \left(\frac{n!}{x_1! x_2! x_3!} \right) \left(\frac{1}{4} \right)^{x_1} \left(\frac{1}{4} + \frac{p}{4} \right)^{x_2} \left(\frac{1}{2} - \frac{p}{4} \right)^{x_3} = 0$$

which yields

$$p^{(k+1)} = \frac{2x_2^k - x_3}{x_2^k + x_3}$$

Emission Tomography (PET SPECT)

The probability of detecting an event originated from box j in detector tube i:
P(event detected in tube i —event occurred in box j) $= H_{ij}$,
where $\mathbf{H}$ is called the system matrix.



Emission Tomography (PET SPECT)

Sum of Poisson Processes is another Poisson :

Given two Poisson distributed random variables y_1 and y_2 and their corresponding distribution parameters λ_1 and λ_2. If $y = x_1 + x_2$, then the probability density function of y is

$$P(y = n) = P(x_1 + x_2 = n) = \sum_{k=0}^{n} P(x_1 = k)P(x_2 = n - k)$$

$$= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^k}{(n-k)!} = \frac{1}{n!} e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$$

$P(y = n) = \frac{1}{n!} e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^n$ which is another Poisson with parameter $\sum_i \lambda_i$.

Emission Tomography (PET SPECT)

Conditional Probability of a Poisson Process is Binomial :

$$P(x_1 = x | x_1 + x_2 = y) = \frac{P(x_1 = x, x_1 + x_2 = y)}{P(x_1 + x_2 = y)}$$

$$P(x_1 = x, x_1 + x_2 = y) = P(x_1 = x)P(x_2 = y - x)$$

$$= e^{-\lambda_1} \frac{\lambda_1^x}{x!} e^{-\lambda_2} \frac{\lambda_2^{y-x}}{(y-x)!}$$

$$P(x_1 + x_2 = y) = e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^y}{y!}$$

$$E[x_1 = x | x_1 + x_2 = y] = \frac{y!}{x!(y-x)!} \frac{e^{-\lambda_1} e^{-\lambda_2}}{e^{-(\lambda_1+\lambda_2)}} \frac{\lambda_1^x \lambda_2^{y-x}}{(\lambda_1 + \lambda_2)^y}$$

which shows that it is Binomially distributed (n, p) with parameters $(y = x_1 + x_2, \frac{\lambda_1}{\lambda_1+\lambda_2}) \longrightarrow Binom(\sum_k x_k, \frac{\lambda_k}{\sum_k \lambda_k})$

Emission Tomography (PET SPECT)

The complete data will be denoted as a matrix $\mathbf{G}$, where G_{ij} represents the number of detected events in tube i originated from box j.
If we could observe $\mathbf{G}$ directly, the solution to the whole problem would be simple: $\hat{f}_j = \sum_i G_{ij}$
$E[g_i]$ is the average/expected number of events detected in tube i

$$E[g_i] = E(\sum_j G_{ij}) = \sum_j H_{ij} f_j$$

In matrix-vector notation:

$$E[\mathbf{g}] = \mathbf{Hf}$$

Image reconstruction is to invert this equation to solve for image $\mathbf{f}$ based on what we observe $\mathbf{g}$ as cumulative events from all boxes.

Emission Tomography (PET SPECT)

G_{ij} : Poisson distributed ($\sim e^{-\lambda}\frac{\lambda^k}{k!}$) complete data (observed + hidden)

The likelihood function is with $\lambda \longrightarrow E(G_{ij})$

$$L(\mathbf{f}) = P(\mathbf{G}|\mathbf{f}) = \prod_i \prod_j P(G_{ij}|\mathbf{f}) = \prod_i \prod_j e^{-E[G_{ij}]} \frac{E[G_{ij}]^{G_{ij}}}{G_{ij}!}$$

G_{ij} represents the number of detected events in tube i originated from box j.

$E[G_{ij}]$ is the expected number of emissions from box j measured in tube i:

$$E[G_{ij}] = H_{ij} f_j$$

Emission Tomography (PET SPECT)

The log-likelihood function is

$$\log L(\mathbf{f}) = \sum_i \sum_j -f_j H_{ij} + G_{ij} \ln f_j H_{ij} - \ln G_{ij}!$$

Expectation :

$$E_{G_{ij}|\mathbf{g},\hat{\mathbf{f}}^k}\left[\sum_i \sum_j -f_j H_{ij} + G_{ij} \ln f_j H_{ij} - \ln G_{ij}!|\mathbf{g}, \hat{\mathbf{f}}^{\mathbf{k}}\right]$$

$$= \sum_i \sum_j \Big(- f_j H_{ij} + E[G_{ij}|\mathbf{g}, \hat{\mathbf{f}}^{\mathbf{k}}] \ln(f_j H_{ij}) - E[\ln G_{ij}!|\mathbf{g}, \hat{\mathbf{f}}^{\mathbf{k}}]\Big)$$

Emission Tomography (PET SPECT)

Expectation :

Considering the constraint that $g_i = \sum_j G_{ij}$ (like $y = x_1 + x_2$)

and G_{ij} are independent Poisson random variables, with mean $\lambda \longrightarrow E[G_{ij}]$

the conditional probability of G^{ij} upon g_i

$$P[G_{ij}|\mathbf{g}, \hat{\mathbf{f}}^{\mathbf{k}}]$$

is Binomially distributed with parameters $(\sum_j G_{ij}, \frac{E[G_{ij}]}{\sum_j E[G_{ij}]})$.

Remembering that the expected value of $Binom(N, p)$ is Np and $E[G_{ij}] = f_j H_{ij}$

$$E[G_{ij}|\mathbf{g}, \hat{\mathbf{f}}^{\mathbf{k}}] = \sum_j G_{ij} \frac{\hat{f}_j^k H_{ij}}{\sum_m \hat{f}_m^k H_{im}} = g_i \frac{\hat{f}_j^k H_{ij}}{\sum_m \hat{f}_m^k H_{im}}$$

Emission Tomography (PET SPECT)

Maximization :

$$\log L(\mathbf{f}) = \sum_i \sum_j -f_j H_{ij} + G_{ij} \ln f_j H_{ij} - \ln G_{ij}!$$

$$\frac{\partial}{\partial f_l} E[\log L(\mathbf{f})|\mathbf{g}, \hat{\mathbf{f}}^k] = 0 = -\sum_i H_{il} + \sum_i E[G_{il}|\mathbf{g}, \hat{\mathbf{f}}^k] \frac{1}{f_l}$$

$$0 = -\sum_i H_{il} + \sum_i g_i \frac{\hat{f}_l^k H_{il}}{\sum_m \hat{f}_m^k H_{im}} \frac{1}{f_l}$$

$$f_l = \hat{f}^{k+1} = \frac{\hat{f}_l^k}{\sum_i H_{il}} \sum_i \frac{H_{il} g_i}{\sum_m \hat{f}_m^k H_{im}}$$

Generation of PET Data

```
% Poisson model Data
N_Points = 1000; % Number of data points detected
D = 80; % Number of detectors
R = 50 % Radius of circular gantry in cm
Nx = 40; % x resolution
Ny =  40; % y resolution
B = Nx * Ny; % number of boxes

Pet_Image = zeros(Nx,Ny);
S= randi(100,10,10);
Pet_Image(15+[1:size(S,1)], 15 + [1:size(S,2)]) = S ;
%Pet_Image = 100*imfilter(Pet_Image,fspecial('gaussian',8,2)); % tissue density f
Index = find(Pet_Image>0) ;

H = rand(D,length(Index)); % Transfer function for
% (event detected  in tube i |  event occurred in box j)
M = sum(H); % normalized to give 1 for the probability of detecting an event over all Detectors
H = H./(ones(D,1)*M);
lambda = Pet_Image(Index);
g = H*lambda;     % total number of events detected from all boxes in detector i
```

Self-Study Question

n maids choose a spice out of four choices with distribution $\mathbf{y} = [y_1 \cdots y_4]^T$ where y_i is the number of maids that chose spice i.
$\mathbf{y}$ is distributed according to a multinomial distribution

$$P(\mathbf{y}|\lambda) = \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}$$

with $\mathbf{p} \in (0,1)^4$ and $p_1 + p_2 + p_3 + p_4 = 1$.

$$p_\lambda = [\tfrac{1}{2} + \tfrac{1}{4}\lambda \quad \tfrac{1}{4}(1-\lambda) \quad \tfrac{1}{4}(1-\lambda) \quad \tfrac{1}{4}\lambda],\ \lambda \in (0,1).$$

The complete data $\mathbf{x} = [x_1 \cdots x_5]^T$ has a multinomial distribution with number of trials n and the probability of

$$q_\lambda = [\tfrac{1}{2} \quad \tfrac{1}{4}\lambda \quad \tfrac{1}{4}(1-\lambda) \quad \tfrac{1}{4}(1-\lambda) \quad \tfrac{1}{4}\lambda],\ \lambda \in (0,1).$$

$$\mathbf{y} = T(\mathbf{x}) = [x_1 + x_2 \quad x_3 \quad x_4 \quad x_5]^T$$

Derive the EM iteration equations to determine the λ.

Gaussian Mixture Model (GMM)

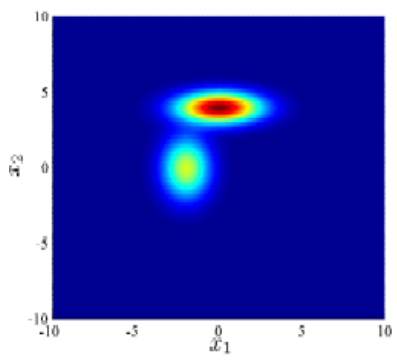
If we make a number of observations that are assumed to be generated by K Gaussians and if we want to find the means and covariances of the Gaussians, and the probability (weight) that a point comes from each of the Gaussians, we can use the GMM.

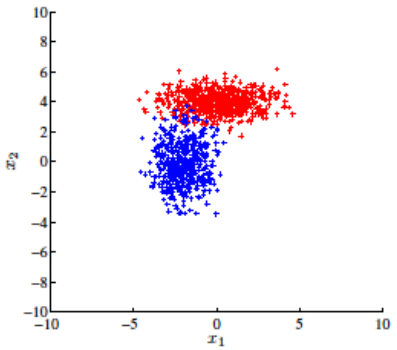
Given a sample $\mathbf{y} \in \mathcal{R}^D$ from a GMM with K components, we estimate its parameter set $\lambda = \{(\pi_j, \mu_j, \Sigma_j)\}_{j=1}^{K}$.

$$\mathcal{N}(\mathbf{y}|\mu, \Sigma) \triangleq \frac{exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)\right)}{(2\pi)^{D/2}|\Sigma|^{1/2}}$$

$$p(\mathbf{y}|\lambda) = \sum_{i=1}^{K} \pi_i \phi(\mathbf{y}|\mu_i, \Sigma_i)$$

$$\sum_{i=1}^{K} \pi_i = 1 \text{ and } \pi_i \geq 0$$





EM clustering by a Gaussian Mixture Model

$$p(\mathbf{y}|\lambda) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mu_i, \Sigma_i) = \sum_{i=1}^{K} \pi_i \frac{\exp(-\frac{1}{2}(\mathbf{y}-\mu_i)^T \Sigma_j^{-1}(\mathbf{y}-\mu_i)}{(2\pi)^{D/2}|\Sigma|^{1/2}}$$

Let us introduce a $K-$dimensional binary random variable, $\mathbf{z}$ whose all elements but one is zero, and it defines from which Gaussian component the data $\mathbf{y}$ came.

$P(z_k = 1) = \pi_k,$

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k},$$

$$p(\mathbf{y}|\mu_k, \Sigma_k) \ , \ k = 1, \ldots, K.$$

Conditional distribution of $\mathbf{y}$ given $\mathbf{z}$ is

$$p(\mathbf{y}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k)^{z_k}$$

Conditional probability of $\mathbf{z}$ given $\mathbf{y}$ is

$$p(z_k = 1|\mathbf{y}) = \gamma(z_k) = \frac{p(\mathbf{y}|z_k{=}1)p(z_k{=}1)}{\sum_{j=1}^{K} p(\mathbf{y}|z_j{=}1)p(z_j{=}1)} = \frac{\pi_k \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}|\mu_j, \Sigma_j)}$$

Multiple Observations

For a set of observations $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ and hidden variables $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^T$

The log-likelihood function for the discrete $\mathbf{Z}$ becomes

$$\log(\mathbf{Y}|\lambda) = \log \left\{ \sum_{\mathbf{Z}} p(\mathbf{Y}, \mathbf{Z}|\lambda) \right\}$$

EM clustering by a Gaussian Mixture Model

If the complete data is the observed data Y plus some missing (also called latent or hidden) data $\{\mathbf{Y}, \mathbf{Z}\}$,
The $Q-$function over the domain of $\mathbf{Z}$ because the only random part the complete data is Z

$$\begin{aligned} Q(\lambda|\lambda_m) &= E_{\mathbf{Z}|\mathbf{Y},\lambda^m}[\log p(\mathbf{Y},\mathbf{Z}|\lambda)] \\ &= \sum_{\mathbf{Z}} \log p(\mathbf{Y},\mathbf{Z}|\lambda) p(\mathbf{Z}|\mathbf{Y},\lambda^m) d\mathbf{Z} \end{aligned}$$

Since $\gamma^m(z_{nk})$ as our guess at the m^{th} iteration of the probability that the n^{th} sample belongs to the k^{th} Gaussian component,

$$\gamma^m(z_{nk}) \triangleq P(z_{nk}=1|\mathbf{y}_n,\lambda^m) = \frac{\pi_k^m \mathcal{N}(\mathbf{y}_n|\mu_k^m,\Sigma_k^m)}{\sum_{j=1}^{K} \pi_j^m \mathcal{N}(\mathbf{y}_n|\mu_j^m,\Sigma_j^m)}$$

which satisfies $\sum_{j=1}^{K} \gamma^m(z_{nj}) = 1.$

The E-step:

$$Q(\lambda|\lambda_m) = \mathrm{E}_{\mathbf{Z}|\mathbf{Y},\lambda^m}[\log p(\mathbf{Y},\mathbf{Z}|\lambda)] = \mathrm{E}_{\mathbf{Z}|\mathbf{Y},\lambda^m}\left[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \pi^{z_{nk}} \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)^{z_{nk}}\right]$$

$$= \mathrm{E}_{\mathbf{Z}|\mathbf{Y},\lambda^m}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\log \pi_k + \log \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)\right\}\right]$$

The posterior distribution of $\mathbf{Z}$ is

$$p(\mathbf{Z}|\mathbf{Y},\lambda) = p(\mathbf{Y}|\mathbf{Z},\lambda)p(\mathbf{Z}|\lambda)/p(\mathbf{Y})$$

$$p(\mathbf{Z}|\mathbf{Y},\lambda) \propto \prod_{k=1}^{K} \pi^{z_{nk}} \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)^{z_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)]^{z_{nk}}$$

The expected value of the indicator variable z_{nk} is

$$\mathrm{E}_{\mathbf{Z}|\mathbf{Y},\lambda^m}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk}[\pi_k \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)^{z_{nk}}}{\sum_{z_{nk}}[\pi_j \mathcal{N}(\mathbf{y}_n|\mu_j,\Sigma_j)^{z_{nj}}} = \frac{\pi_k \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}_n|\mu_j,\Sigma_j)} = \gamma(z_{nk})$$

$$Q(\lambda|\lambda_m) = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\{\log \pi_k + \log \mathcal{N}(\mathbf{y}_n|\mu_k,\Sigma_k)\}$$

The M-Step

$$Q(\lambda|\lambda^m) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big(\log\pi_k - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{1}{2}(\mathbf{y}_n-\mu_k)^T\Sigma_k^{-1}(\mathbf{y}_n-\mu_k) - \tfrac{D}{2}\log 2\pi\Big)$$

M-Step is to $\underset{\lambda}{\textit{maximize}}\ Q(\lambda|\lambda^m)$ subject to $\sum_{k=1}^{K}\pi_k = 1,\ \pi_k \geq 0.$

We form the Lagrangian,

$$\mathcal{L}(\lambda,\eta) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big(\log\pi_k - \tfrac{1}{2}\log|\Sigma_k| - \tfrac{1}{2}(\mathbf{y}_n-\mu_k)^T\Sigma_k^{-1}(\mathbf{y}_n-\mu_k) - \tfrac{D}{2}\log 2\pi\Big) + \eta\Big(\sum_{k=1}^{K}\pi_k - 1\Big)$$

and optimize it with respect to λ.

The M-Step

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}\Big(\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \log \pi_k + \eta\Big(\sum_{j=1}^{K} \pi_j - 1\Big)\Big) = 0, \;\; k = 1, \ldots, K$$

yields

$$\sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{\pi_k} + \eta = 0 \rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) = -\eta \pi_k$$

$$\frac{\partial \mathcal{L}}{\partial \eta} = \sum_{j=1}^{K} \pi_j - 1 = 0 \rightarrow \sum_{j=1}^{K} \pi_j = 1$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) = -\eta \sum_{k=1}^{K} \pi_k = -\eta$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})} = \frac{N_k}{\sum_{j=1}^{K} N_j} = \frac{N_k}{N}, \quad k = 1, \ldots, K.$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$ *i.e.* the the effective number of points assigned to cluster k.

The M-Step

Similarly, we get

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k^{-1} \Big(\mathbf{y}_n - \mu_k \Big) = 0, \quad k = 1, \ldots, K.$$

$$\mu_k^{m+1} = \frac{1}{N_k^m} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{y}_n, \quad k = 1, \ldots, K.$$

and

$$\frac{\partial \mathcal{L}}{\partial \Sigma_k^{-1}} = \frac{1}{2} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{\partial}{\partial \Sigma_k^{-1}} \log |\Sigma_k^{-1}| - \frac{\partial}{\partial \Sigma_j^{-1}} (\mathbf{y}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_n - \mu_k) \right) = 0$$

$$\Sigma_k^{m+1} = \frac{1}{N_k^m} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{y}_n - \mu_k^{m+1}) (\mathbf{y}_n - \mu_k^{m+1})^T, \quad k = 1, \ldots, K.$$

EM Algorithm for GMM

- Initialize $\lambda^0 = \{\pi_k^0, \mu_k^0, \Sigma_k^0\}$, $k = 1, \ldots, K$
- E-Step: Compute

$$\gamma^m(z_{nk}) = \frac{\pi_k^m \phi(\mathbf{y}_n|\mu_k^m, \Sigma_k^m)}{\sum_{j=1}^{K} \pi_j^m \mathcal{N}(\mathbf{y}_n|\mu_j^m, \Sigma_j^m)} \quad \text{and} \quad N_k^m = \sum_{n=1}^{N} \gamma(z_{nk})^m$$

  for $n = 1, \ldots, N$ and $k = 1, \ldots, K$.

- M-Step: Compute the new estimates

$$\pi_k^{m+1} = \frac{N_k^m}{\sum_{j=1}^{K} N_j^m} = \frac{N_k^m}{N}, \quad k = 1, \ldots, K.$$

$$\mu_k^{m+1} = \frac{1}{N_k^m} \sum_{n=1}^{N} \gamma(z_{nk})^m \mathbf{y}_n, \quad k = 1, \ldots, K.$$

$$\Sigma_k^{m+1} = \frac{1}{N_k^m} \sum_{n=1}^{N} \gamma(z_{nk})^m (\mathbf{y}_n - \mu_k^{m+1})(\mathbf{y}_n - \mu_k^{m+1})^T, \quad k = 1, \ldots, K.$$

- Convergence Check:
  and iterate until $|\lambda^{m+1} - \lambda^m| < \delta$.

A GMM Fitting Example

$$\mu_1 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \mu_2 = \begin{bmatrix} -2 \\ 0 \end{bmatrix},$$

$$\Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

$\pi_1 = 0.6$, $\pi_2 = 0.4$

Initial values are

$$\mu_1^0 = \begin{bmatrix} 0.0823 \\ 3.9189 \end{bmatrix}, \mu_2^0 = \begin{bmatrix} -2.0706 \\ -2.2327 \end{bmatrix},$$

$\Sigma_1^0 = \Sigma_2^0 = I_2$,

$\pi_1^0 = \pi_2^0 = 0.5$ and $\delta = 10^{-3}$.

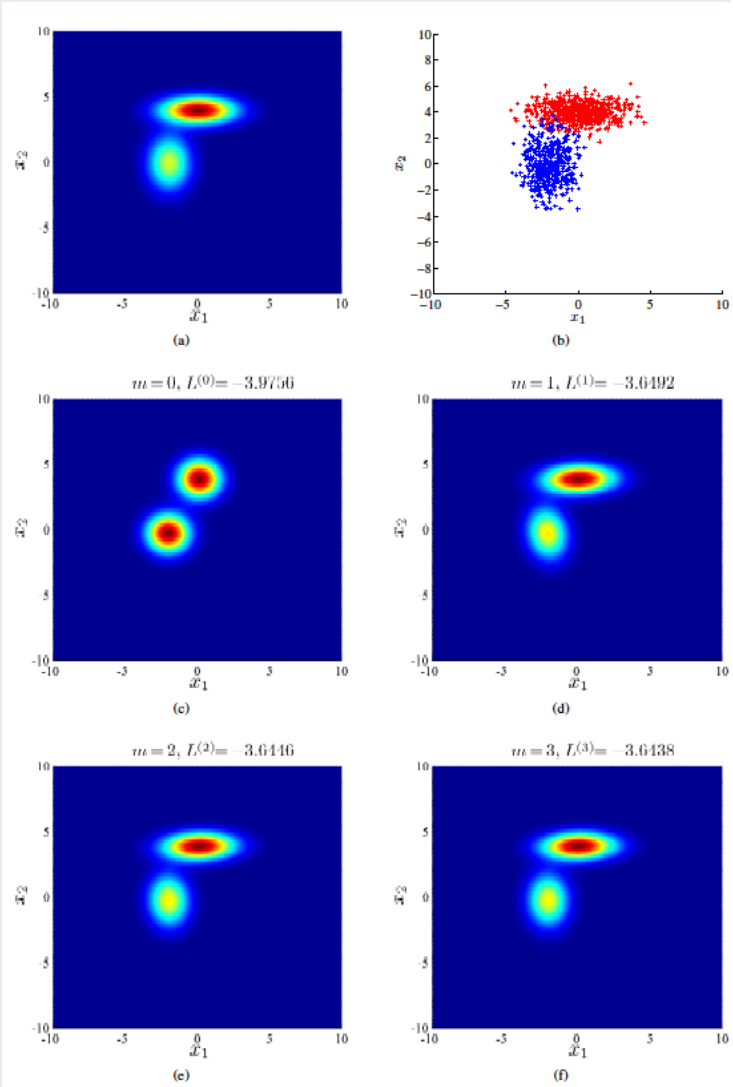
After 3 iterations

$$\mu_1^3 = \begin{bmatrix} 0.0806 \\ 3.9445 \end{bmatrix}, \mu_2^3 = \begin{bmatrix} -2.0181 \\ -0.1740 \end{bmatrix},$$

$$\Sigma_1^3 = \begin{bmatrix} 2.7452 & 0.0568 \\ 0.0568 & 0.4821 \end{bmatrix},$$

$$\Sigma_2^3 = \begin{bmatrix} 0.8750 & -0.0153 \\ -0.0153 & 1.7935 \end{bmatrix},$$

$\pi_1^3 = 0.5966$, $\pi_2^3 = 0.4034$.



GMM Fitting Data Generation

```
N=1000;
mu1 = [0 ; 4];  mu2 = [-2 ;0];
Sigma1 = [3 0;0 0.5];  Sigma2 = [1 0 ; 0 2];
S1 = chol(Sigma1);   S2 = chol(Sigma2);
pi1 = 0.6 ; pi2=0.4;
Yi = ones(N,1);
for i=1:N;
    if ( rand() < 0.6 )
       Y(:,i)= S1*randn(2,1) + mu1;
   else
       Y(:,i)= S2*randn(2,1) + mu2;  Yi(i) =2;
   end;
end;
```