CS 5135/6035 Learning Probabilistic Models

Lecture 11: Expectation Maximization for MV Gaussians, Correctness

Gowtham Atluri

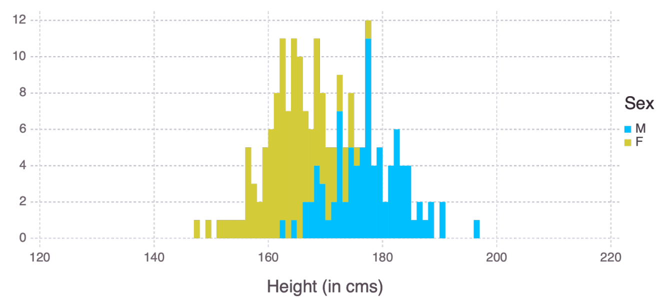
October 4, 2018

Reading Material

- Chapter 9. Mixture Models and EM
  - Bishop, Pattern Recognition and Machine Learning

Parameter Estimation: Mixture of Univariate Gaussians

- Height of 200 subjects



Mixture Models - Expectation Maximization (EM)

- Probability density

$$p(x_i) = \pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)$$

- Log likelihood

$$\ell = \sum_{i=1}^{n} \log\left(\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)\right)$$

- Differentiating ℓ w.r.t. μ_M, we have

$$\sum_{i=1}^{n} \frac{1}{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)} \pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) \frac{x_i - \mu_M}{\sigma^2} = 0$$

- The posterior probability that $z_i = M$

$$p(M|x_i) = \frac{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2)}{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)}$$

- Assuming we know $p(M|x_i)$, and by substituting it in the $\frac{d}{d\mu_M}\ell$

$$\mu_M = \frac{\sum_{i=1}^{n} p(M|x_i) x_i}{\sum_{i=1}^{n} p(M|x_i)} \qquad \mu_F = \frac{\sum_{i=1}^{n} p(F|x_i) x_i}{\sum_{i=1}^{n} p(F|x_i)}$$

Expectation Maximization (EM)

- An elegant and a powerful method for finding Max. Likelihood solutions for models with latent variables
- *Step 1:* Pick initial value μ_M and μ_F
- *Step 2:* $maxIter = 1000$
- *Step 3:* **for** $i = 1 : maxIter$
- *Step 4:* Compute $p(M|x_i)$

$$p(M|x_i) = \frac{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2)}{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)}$$

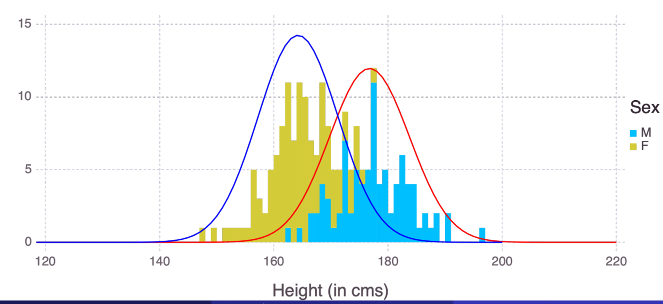
- *Step 5:* Optimize for μ_M and μ_F

$$\mu_M^i = \frac{\sum_{i=1}^{n} p(M|x_i) x_i}{\sum_{i=1}^{n} p(M|x_i)} \qquad \mu_F^i = \frac{\sum_{i=1}^{n} p(F|x_i) x_i}{\sum_{i=1}^{n} p(F|x_i)}$$

- *Step 6:* **if** $|\mu_M^i - \mu_M^{i-1}| < \epsilon$ and $|\mu_F^i - \mu_F^{i-1}| < \epsilon$ terminate; **end**
- *Step 7:* **end for**

Parameter Estimation: Mixture of Univariate Gaussians

- Height of 200 subjects
- $p(x_i) = \pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)$
- Estimating μ_M, μ_F
  - assuming σ^2 is same for the two components and is known.
  - assuming π_M and π_F are known.

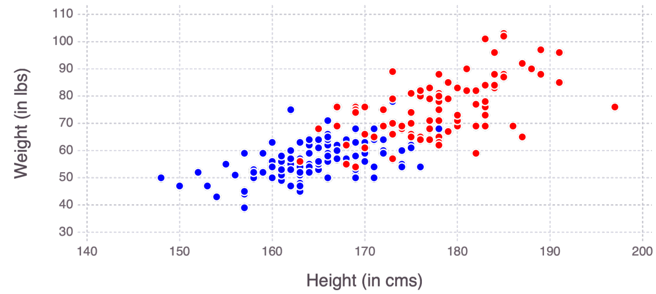


Mixture of Multivariate Gaussians

- Motivation
- Assumptions (Univarite vs. Bivariate)
- Maximizing Likelihood
- Update Equations
- EM Approach

Parameter Estimation: Mixture of Bivariate Gaussians

- Height and Weight of 200 subjects



Mixture of MV Gaussians

Univariate case:

$$p(x_i) = \pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)$$

- EM approach for estimating μ_M, μ_F
  - assuming σ^2 is same for the two components and is known.
  - assuming π_M and π_F are known.

Multivariate case:

$$p(\boldsymbol{x}) = \pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$$

- Goal is to estimate $(\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$, $(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$, and (π_M, π_F).
  - using Maximum Likelihood Estimation

Mixture of MV Gaussians

- $\boldsymbol{x}$ is the observed random variable
- Let z be a binary latent variable.
  - In general, z can be a categorical variable.
- $$p(z) = p^{\mathbb{1}(z=M)}(1-p)^{\mathbb{1}(z=M)} \qquad p(z) = \prod_{c \in \{M,F\}} \pi_c^{\mathbb{1}(z=c)}$$

  where $\pi_M = p$, $\pi_F = 1 - p$
- Alternatively,

$$p(\boldsymbol{z}) = \prod_{c \in \{M,F\}} \pi_c^{\mathbb{1}(z=c)}$$

- Conditional distribution of $\boldsymbol{x}$, given a value of $\boldsymbol{z}$

$$p(\boldsymbol{x}|z) = \prod_{c \in \{M,F\}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)^{\mathbb{1}(z=c)} \qquad p(\boldsymbol{x}|z = M) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$$

Mixture of MV Gaussians

$$p(z) = \prod_{c \in \{M,F\}} \pi_c^{\mathbb{1}(z=c)} \qquad p(\boldsymbol{x}|z) = \prod_{c \in \{M,F\}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)^{\mathbb{1}(z=c)}$$

- Marginal distribiuon of $\boldsymbol{x}$ is obtained as

$$p(\boldsymbol{x}) = \sum_{z \in \{M,F\}} p(z)p(\boldsymbol{x}|z) = \pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$$

We will use this to write the likelihood of observed variables $\boldsymbol{x}$.

- Conditional probability of z given $\boldsymbol{x}$, denoted as $p(z = M|\boldsymbol{x})$ or $\gamma(M)$

$$\begin{aligned}\gamma(M) \equiv p(z = M|\boldsymbol{x}) &= \frac{p(z = M)p(\boldsymbol{x}|z = M)}{\sum_{j \in \{M,F\}} p(z = j)p(\boldsymbol{x}|z = j)} \\ &= \frac{\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)}{\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)}\end{aligned}$$

Likelihood for a Mixture of MV Gaussians

- Assuming N data points $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are sampled from the Mixture of MV Gaussians
- Density for one data point $\boldsymbol{x}_i$ is

$$p(\boldsymbol{x}_i) = \pi_M \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$$

- Likelihood is

$$L(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n) = \prod_{n=1}^{N} (\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F))$$

- Log-likelihood is

$$\ell = \sum_{n=1}^{N} \log(\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F))$$

- Compute partial derivatives and solve for the parameters

MV Gaussian - partial derivative w.r.t. $\boldsymbol{\mu}$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\mu}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{\partial}{\partial\boldsymbol{\mu}}\Big[\frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\Big] \\ &= \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.\frac{2}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}).\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

MV Gaussian - partial derivative w.r.t. $\boldsymbol{\Sigma}$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{\partial}{\partial\boldsymbol{\Sigma}}\Big[\frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\Big] \\ &= \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.\frac{-1}{2}\boldsymbol{\Sigma}^{-1} \\ &+ \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.\Big[\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} \\ &= \frac{-1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\Big[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\Big]\end{aligned}$$

Maximum Likelihood Expectation (estimating $\boldsymbol{\mu}_M$, $\boldsymbol{\mu}_F$)

$$\ell = \sum_{n=1}^{N}\log(\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F))$$

$$\frac{\partial\ell}{\partial\boldsymbol{\mu}_M} = -\sum_{n=1}^{N}\underbrace{\frac{\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M)}{\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F)}}_{\gamma(M)}\boldsymbol{\Sigma}_M^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_M) = 0$$

$$\sum_{n=1}^{N}\gamma(M)\boldsymbol{\Sigma}_M^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_M) = 0 \implies \sum_{n=1}^{N}\gamma(M)\boldsymbol{\Sigma}_M^{-1}\mathbf{x}_n = \sum_{n=1}^{N}\gamma(M)\boldsymbol{\Sigma}_M^{-1}\boldsymbol{\mu}_M$$

$$\implies \sum_{n=1}^{N}\gamma(M)\mathbf{x}_n = \boldsymbol{\mu}_M\sum_{n=1}^{N} \implies \boldsymbol{\mu}_M = \frac{\sum_{n=1}^{N}\gamma(M)\mathbf{x}_n}{\sum_{n=1}^{N}\gamma(M)}$$

Maximum Likelihood Expectation (estimating $\boldsymbol{\Sigma}_M$, $\boldsymbol{\Sigma}_F$)

$$\ell = \sum_{n=1}^{N}\log(\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F))$$

$$\frac{\partial\ell}{\partial\boldsymbol{\Sigma}_M} = \sum_{n=1}^{N}\gamma(M)[\Sigma_M^{-1} - \Sigma_M^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})(\mathbf{x_M}-\boldsymbol{\mu}_\mathbf{M})^T\Sigma_M^{-1}] = 0$$

$$\implies \sum_{n=1}^{N}\gamma(M)\boldsymbol{\Sigma}_M^{-1} = \sum_{n=1}^{N}\gamma(M)\boldsymbol{\Sigma}_M^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})^T\boldsymbol{\Sigma}_M^{-1}$$

$$\implies \boldsymbol{\Sigma}_M\sum_{n=1}^{N}\gamma(M) = \sum_{n=1}^{N}\gamma(M)(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})^T$$

$$\implies \boldsymbol{\Sigma}_M = \frac{\sum_{n=1}^{N}\gamma(M)(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})(\mathbf{x}-\boldsymbol{\mu}_\mathbf{M})^T}{\sum_{n=1}^{N}\gamma(M)}$$

Maximum Likelihood Expectation (estimating π_M, π_F)

- We need to maximize ℓ under the constraint $\pi_M + \pi_F = 1$.

$$\ell = \sum_{n=1}^{N}\log(\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F))$$

- Achieved using Lagrange multiplier and maximizing the following quantity

$$\ell + \lambda\Big(\pi_M + \pi_F - 1\Big)$$

- Compute the derivative w.r.t π_M and equate it to 0.

$$\sum_{n=1}^{N}\frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M)}{\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F)} + \lambda = 0$$

- Multiplying both slides by π_M

$$\sum_{n=1}^{N}\frac{\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M)}{\pi_M\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M) + \pi_F\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_F,\boldsymbol{\Sigma}_F)} + \pi_M\lambda = 0 \implies \sum_{n=1}^{N}\gamma(M) = -\lambda\pi_M$$

Maximum Likelihood Expectation (estimating π_M, π_F)

$$\sum_{n=1}^{N}\gamma(M) = -\lambda\pi_M$$

- Taking sum over the two labels $\{M, F\}$ of z

$$\sum_{c\in\{M,F\}}\sum_{n=1}^{N}\gamma(M) = \sum_{c\in\{M,F\}} -\lambda\pi_c \implies N\sum_{c\in\{M,F\}}\gamma(M) = -\sum_{c\in\{M,F\}}\lambda\pi_c$$

$$\implies N = -\lambda(\pi_M + \pi_F) \implies \lambda = -N$$

- Substituting $\lambda = -N$, in the above equation we have.

$$\sum_{n=1}^{N}\gamma(M) = N\pi_M$$

$$\pi_M = \frac{\sum_{n=1}^{N}\gamma(M)}{N}$$

EM Approach

E Step:

$$\gamma(M) = \frac{\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)}{\pi_M \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M) + \pi_F \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)} \qquad \gamma(F) = 1 - \gamma(M)$$

M Step:

$$\boldsymbol{\mu}_M = \frac{\sum_{n=1}^N \gamma(M)\boldsymbol{x}_n}{\sum_{n=1}^N \gamma(M)} \qquad \boldsymbol{\mu}_F = \frac{\sum_{n=1}^N \gamma(F)\boldsymbol{x}_n}{\sum_{n=1}^N \gamma(F)}$$

$$\boldsymbol{\Sigma}_M = \frac{\sum_{n=1}^N \gamma(M)(\boldsymbol{x}-\boldsymbol{\mu}_M)(\boldsymbol{x}-\boldsymbol{\mu}_M)^T}{\sum_{n=1}^N \gamma(M)} \qquad \boldsymbol{\Sigma}_F = \frac{\sum_{n=1}^N \gamma(F)(\boldsymbol{x}-\boldsymbol{\mu}_F)(\boldsymbol{x}-\boldsymbol{\mu}_F)^T}{\sum_{n=1}^N \gamma(F)}$$

$$\pi_M = \frac{\sum_{n=1}^N \gamma(M)}{N} \qquad \pi_M = 1 - \pi_F$$

Mixture of Multivariate Gaussians (Julia code)

- E-step; M-Step
- Visualization of the estimated components
- Singularities

E-step (Julia code)

```
function E_step(x,mu_M,mu_F,sigma_M,sigma_F,pi_M)
    numerator = zeros(size(x,1));
    denominator = zeros(size(x,1));
    post_x = zeros(size(x,1));
    for i=1:size(x,1)
        numerator[i] = pi_M.*pdf(MvNormal(mu_M,sigma_M),x[i,:]);
        denominator[i] = numerator[i]
                + (1-pi_M)* pdf(MvNormal(mu_F,sigma_F),x[i,:]);
        post_x[i] = numerator[i] ./denominator[i];
    end

    return post_x;
end
```

```
## E_step (generic function with 2 methods)
```


E-step (Julia code)

```
function M_step(x,post_x)
    mu_M = sum(post_x.*x,1)./sum(post_x);
    mu_M = Vector(mu_M[:]);
    mu_F = sum((1.-post_x).*x,1)./sum((1.-post_x));
    mu_F = Vector(mu_F[:]);
    sigma_M = round.((post_x.*(x.-mu_M'))'*(x.-mu_M')
              /sum(post_x),5);
    sigma_F = round.(((1.-post_x).*(x.-mu_F'))'*(x.-mu_F')
              /sum(1.-post_x),5);
    pi_M = sum(post_x)/size(x,1);
    return mu_M, mu_F, sigma_M, sigma_F, pi_M;
end
```

```
## M_step (generic function with 1 method)
```


EM (Julia code)

```
function EM(x,mu_M,mu_F,sigma_M, sigma_F,pi_M)
    maxIter = 1000;
    for i=1:maxIter
        print(i,"\n");
        post_x = E_step(x,mu_M,mu_F,sigma_M,sigma_F,pi_M);
        mu_M_new, mu_F_new,sigma_M_new, sigma_F_new, pi_M_new =
              M_step(x,post_x);
        if(sum(abs.(mu_M-mu_M_new))<0.001
                && sum(abs.(mu_F-mu_F_new))<0.001
                && sum(abs.(sigma_M-sigma_M_new))<0.001
                && sum(abs.(sigma_F-sigma_F_new))<0.001)
            break;
        end;
        mu_M = mu_M_new; mu_F = mu_F_new;
        sigma_M = sigma_M_new; sigma_F = sigma_F_new;
        pi_M = pi_M_new;
    end
    return mu_M, mu_F, sigma_M, sigma_F, pi_M;
end
```


EM approach on a real dataset

```
data = dataset("car","Davis");
data = data[[1:11; 13:end],:]; #dropping an outlier
x = convert(Array,data[:,[:Height,:Weight]]);
mu_M=[180, 78];
mu_F=[160, 50];
sigma_M = [10.0 0; 0 10.0];
sigma_F = [10.0 0; 0 10.0];
pi_M = 0.5;
mu_M, mu_F, sigma_M, sigma_F, pi_M = EM(x,mu_M,mu_F,sigma_M,sigma_F
```

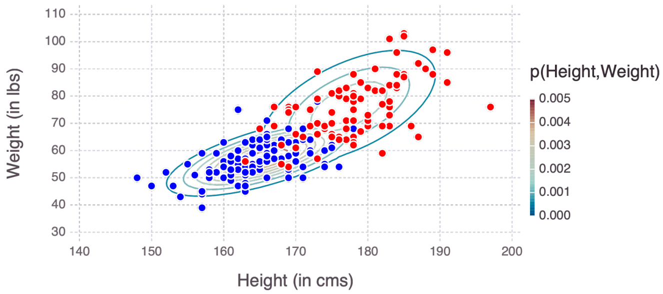
$$\boldsymbol{\mu}_M = [177.37, \quad 76.19] \qquad \boldsymbol{\mu}_F = [165.701, \quad 57.4504]$$

$$\boldsymbol{\Sigma}_M = \begin{bmatrix} 52.5834 & 50.4828 \\ 50.4828 & 155.457 \end{bmatrix} \qquad \boldsymbol{\Sigma}_M = \begin{bmatrix} 42.1344 & 29.5521 \\ 29.5521 & 45.7133 \end{bmatrix}$$

$$\pi_M = 0.4186$$

Parameter Estimation: Mixture of Bivariate Gaussians

- Height and Weight of 200 subjects



Limitation of MLE for Mixture Models

- PDF for a Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

- If the mean of one of the components is exactly equal to the data point

$$\mathcal{N}(x|\mu,\sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}}$$

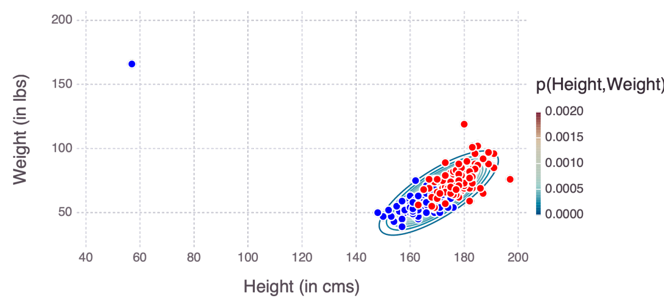
- If we consider the limit $\sigma \to 0$,
  - then this term goes to infinity
  - log-likelihood also goes to infinity
- So an MLE will result in a component with one data point

Limitation of MLE for Mixture Models

- In the case of MLE based univariate parameter estimation
  - When a Gaussian 'collapses' to a data point
    - other data points contribute 0s, resulting in 0 likelihood.
- When there are two (or more) components
  - One component can have finite variance and assign finite probability to all data points
  - other component can shrink to one specific data point, and contribute to increasing likelihood
- This issue of 'singularities' is an example of overfitting that can occur in MLE.

Parameter Estimation: Mixture of Bivariate Gaussians

- Height and Weight of 200 subjects



EM Algorithm

- An abstract view
- Correctness
- KL Divergence

Abstract view of EM

- Goal of EM is to find max. likelihood solutions for models with latent variables
- Let $\boldsymbol{X}$ be the set of all observed data
- Let $\boldsymbol{Z}$ be the set of all latent variables
- Set of model parameters is denoted using $\boldsymbol{\theta}$
- Log-likelihood function is

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \left(\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right)$$

- Note that this discussion is relevant to continuous latent variables as well.
  - Simply replace sum over $\boldsymbol{Z}$ with an integral

Abstract view of EM

- Suppose that for each observation in $\boldsymbol{X}$, we were told the corresponding value of the latent variable $\boldsymbol{Z}$
- Let us call $\{\boldsymbol{X}, \boldsymbol{Z}\}$ the **complete dataset**
- Let us call the actual observed data $\boldsymbol{X}$ the **incomplete dataset**
- The likelihood of the complete dataset takes the form $\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$.
- We are not given $\{\boldsymbol{X}, \boldsymbol{Z}\}$, but only $\boldsymbol{X}$.
  - Our knowledge of latent variables $\boldsymbol{Z}$ is only through the posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$
- We cannot use the complete likelihood
  - we consider instead the expected value under the posterior of the latent variable $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$
- The expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

Abstract view of EM

- The expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- In the E step, we use the current parameter values $\boldsymbol{\theta}^{old}$ to find the posterior distribution of the latent variables $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$.
- We then use this posterior distribution to find the expectation of the compelte-data log-likelihood evaluated for some parameter value $\boldsymbol{\theta}$

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- In the M step, we determine the revised parameter estimate $\boldsymbol{\theta}^{new}$ by maximizing this function

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

A general EM algorithm

- Given a joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$
- **Step 1:** Choose an initial setting for parameters $\boldsymbol{\theta}^{old}$.
- **Step 2:** E Step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$.
- **Step 3:** M Step: Evaluate $\boldsymbol{\theta}^{new}$ given by

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- **Step 4:** Check for convergence of either the log-likelihood or the parameter values. If convergence criteria is not met, then

$$\theta^{old} \leftarrow \theta^{new}$$

and return to step2.

Correctness of EM algorithm

- Our goal is to maximize

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- We introduce a distribution $q(\boldsymbol{Z})$ defined over the latent variables
- **Claim:** For any choice of $q(\boldsymbol{Z})$, the following decomposition holds

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

where we define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

$$KL(q||p) = -\sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

Note that $\mathcal{L}(q, \boldsymbol{\theta})$ is a functional of the distribution $q(\boldsymbol{Z})$, and a function of parameters $\boldsymbol{\theta}$.

Verify the claim using $\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \log p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{X}|\boldsymbol{\theta})$

Kullback-Leibler (KL) Divergence

- KL divergence is a measure for comparing two probability distributions
  - has origins in information theory
- Defining entropy of a probability distribution
  - using log2 helps with interpretation
    - Minimum # bits to encode the information
    - Does not tell us about the optimal encoding scheme

$$H = \sum_{i=1}^{N} p(x_i). \log p(x_i)$$

- Defining KL divergence $D_{KL}(p||q)$: Divergence from q to p (not symmetric)

$$\text{Discrete case: } D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i). \log \frac{p(x_i)}{q(x_i)}$$

$$\text{Continuous case: } D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x). \log \frac{p(x)}{q(x)} dx$$

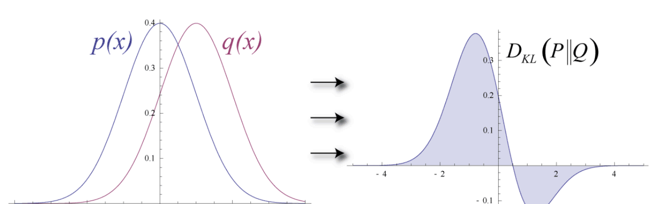
KL Divergence

- Defining KL divergence $D_{KL}(p||q)$: Divergence from q to p

$$\text{Discrete case: } D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i). \log \frac{p(x_i)}{q(x_i)}$$

$$\text{Continuous case: } D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x). \log \frac{p(x)}{q(x)} dx$$

- $KL(q||p) \geq 0$
- $KL(q||p) == 0$, if, and only if, p(x) = q(x).



Correctness of EM algorithm

- For any choice of $q(\boldsymbol{Z})$, the following decomposition holds

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

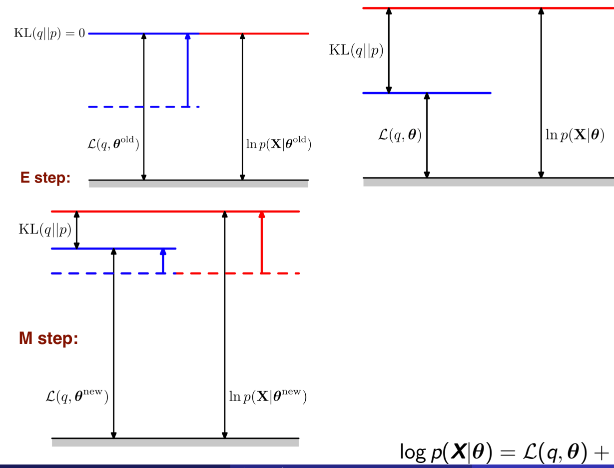
where we define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

$$KL(q||p) = -\sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

- As $\mathrm{KL}(p||q) \geq 0$, $\mathcal{L}(q, \boldsymbol{\theta}) \leq \log p(\boldsymbol{X}|\boldsymbol{\theta})$.
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound on $\log p(\boldsymbol{X}|\boldsymbol{\theta})$.
- In E-Step: Lowed bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $q(\boldsymbol{Z})$, fixing $\boldsymbol{\theta}^{old}$
- In M-Step: $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{new}$

EM approach visually



$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

EM approach visually

$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$