CS 5135/6035 Learning Probabilistic Models

Lecture 10: Latent Variables, Mixture Models, Expectation Maximization

Gowtham Atluri

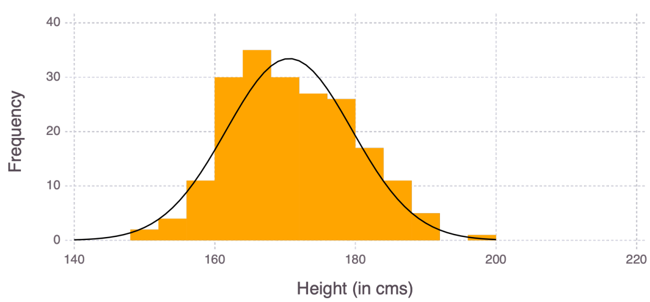
October 2, 2018

Reading Material

- Chapter 9. Mixter Models and EM
  - Bishop, Pattern Recognition and Machine Learning
- R. Sridharan, Gaussian mixture models and the EM algorithm
  - https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf

Parameter Estimation (using MLE)

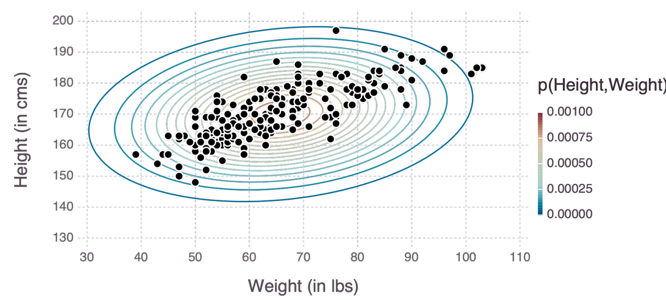
- Fitting Univariate distributions $p(x)$
  - E.g., Height of 200 subjects
  - $$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$



Parameter Estimation (using MLE)

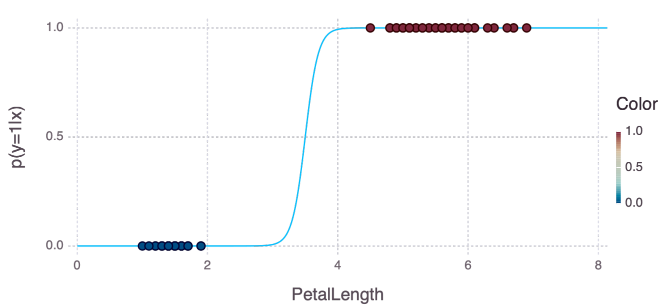
- Fitting Multivariate distributions $p(\boldsymbol{x})$ or $p([x_1, x_2, \ldots, x_d])$
  - E.g., Height and Weight of 200 subjects

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



Parameter Estimation (using MLE)

- Fitting $p(y|x)$ or $p(y|\boldsymbol{x})$
  - E.g., Predicting species from petal length. $p(y=1|x) = \dfrac{1}{1+e^{\beta_0+\beta_1 x}}$
  - $p(\text{species} = \text{virginica}|\text{PetalLength} = 6)$



Latent variables

- Latent variables
- Why model latent variables?
- Identifiability issues
- Modeling latent variables

Observed Data & Modelled variables

Previously, the observed variables are modelled.

Modelling height and weight of subjects in a survey.

Samples	Weight	Height
1	77.4	182.6
2	58.5	161.3
3	63.1	161.2
4	68.6	177.7
5	59.3	157.8
6	76.7	170.4

Estimating parameters to determine p(Weight,Height)

Modelling Species and PetalLength

Samples	PetalLength	Species
1	1.6	setosa
2	1.4	setosa
3	1.3	setosa
4	5.2	virginica
5	5.0	virginica
6	5.2	virginica

Estimating parameters to determine p(Species|PetalLength)

Latent or Hidden variables

Latent Variables

Random variables whose values are not specified in the observed data.

- E.g., An online survey is sent out to employees at a University to collect their height and weight. Gender is a latent variable that is not measured.

Row	Weight	Height	Gender
1	77.4	182.6	M
2	58.5	161.3	F
3	63.1	161.2	F
4	68.6	177.7	M

Why model latent variables?

- To explain observed data in terms of unobserved concepts.
  - A doctor may group patients into those with a certain *syndrome* and those without
    - grouping makes it easier to understand the relationships between observed symptoms.
  - A biologist may wish to group animals into distinct species
    - grouping makes it easier to explain behavioral or physiological patterns
- Often these distinctions cannot be observed or measured.

Modeling Latent or Hidden variables

- We want to model latent variables, along with observed variables.
- Let $\boldsymbol{v}$ be the observerd/visible variables
- Let $\boldsymbol{h}$ be the hidden/latent variables
- We want to model $p(\boldsymbol{v}, \boldsymbol{h})$
- We can write $p(\boldsymbol{v}|\boldsymbol{\theta}) = \sum_h p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{\theta})$
- To estimate $\boldsymbol{\theta}$, we may perform maximum likelihood on the visible variables alone.

Identifiability Issues

- E.g., Consider two varibles x_1, x_2 where x_1 is observed and x_2 is not.
- Let the distribution $p(x_1, x_2|\theta) = \theta_{x_1,x_2}$
- Marginal likelihood only depends on

$$p(x_1|\theta) = \sum_{x_2} \theta_{x_1,x_2}$$

- Given an MLE solution θ^*, we can find an equivalent MLE solution θ' provided

$$\sum_{x_2} \theta'_{x_1,x_2} = \sum_{x_2} \theta^*_{x_1,x_2}$$

Identifiability Issues

- Let the estimated probability distribution be $\hat{p}$

$\hat{p}(x_1 = 0, x_2 = 0)$ $\hat{p}(x_1 = 0, x_2 = 1)$ $\hat{p}(x_1 = 1, x_2 = 0)$ $\hat{p}(x_1 = 1, x_2 = 1)$

- We can construct a new distribution $\hat{p}'$ such that

$$\hat{p}'(x_1 = 0, x_2 = 0) = \hat{p}(x_1 = 0, x_2 = 1)$$
$$\hat{p}'(x_1 = 0, x_2 = 1) = \hat{p}(x_1 = 0, x_2 = 0)$$
$$\hat{p}'(x_1 = 1, x_2 = 0) = \hat{p}(x_1 = 1, x_2 = 1)$$
$$\hat{p}'(x_1 = 1, x_2 = 1) = \hat{p}(x_1 = 1, x_2 = 0)$$

- There is an inherent symmetry in the solution space.

Modeling Latent Variables

- Different scenarios

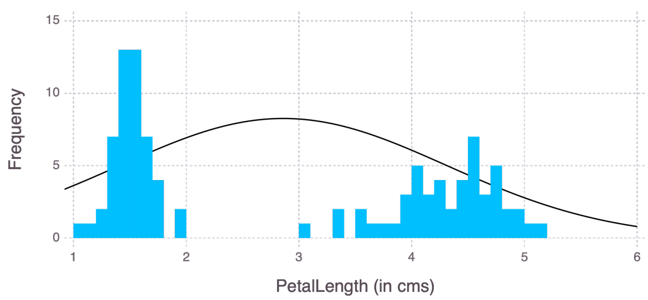
Observed var.	Latent Variable Continuous	Latent Variable Discrete
Continuous	Factor Analysis	Mixture Modeling
Discrete	Latent Trait Analysis	Latent Class Analysis

Mixture Models

- Motivation
- Example
- Formulation
- Limitations of MLE

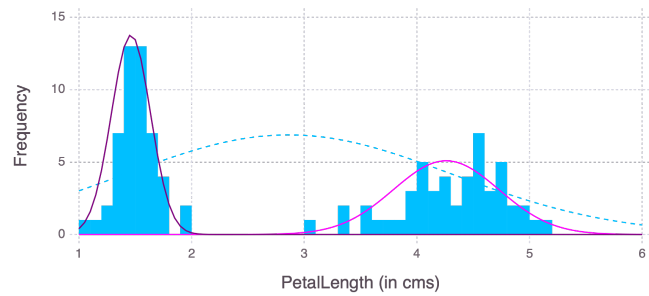
Motivation for Mixture Models

- Parameter estimation based on observed variables is not always ideal for modeling data.
  - E.g., modelling PetalLength using univariate Gaussian
  - A univarite distribution is not suited for modeling a bimodal distribution



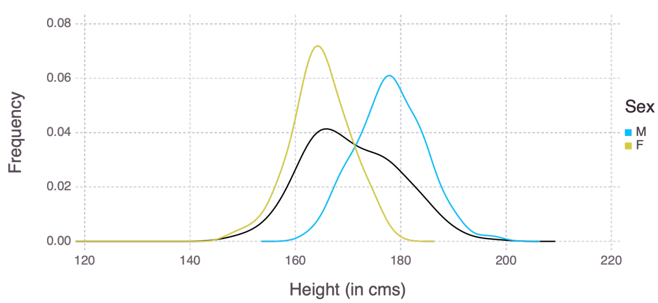
Mixture Models

- Data is modelled as a mixture of several components
  - Each component has a simple parametric form (such as a Gaussian)

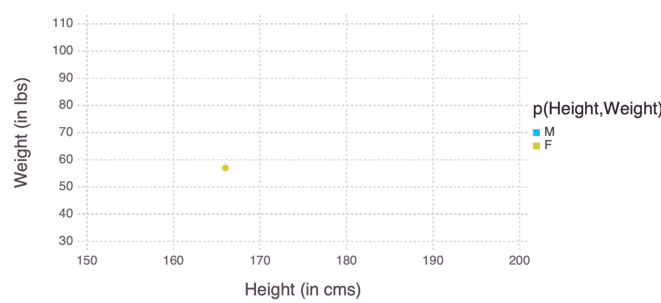


- Mixture Model is not 'aware' of the underlying interpretation

Mixture Models - univariate example



Mixture Models - bivariate example



Mixture Models - formally

Mixture Models

A distribution f is a **mixture** of k component distributions $f_1, f_2, \ldots, f_k$ if

$$f(x) = \sum_{i=1}^{k} \lambda_i f_i(x)$$

where λ_i are the **mixing weights**, $\lambda_i > 0, \sum_i \lambda_i = 1$

The above definition is a complete stochastic model

- It provides a recipe for generating new data points
- First pick a component prob. distribution
  - With probabilities given by the **mixing weights**
- Then generate one observation according to that distribution

$$z_i \sim Multinomial(\lambda_1, \lambda_2, \ldots, \lambda_k)$$
$$x_i | z_i \sim f_{z_i}$$

Mixture Models - formally

Mixture Models

A distribution f is a **mixture** of k component distributions $f_1, f_2, \ldots, f_k$ if

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x)$$

where π_i are the **mixing weights**, $\pi_i > 0, \sum_i \pi_i = 1$

- In principle, f_is can be arbitrary distributions
- In practice, we prefer **parametric mixture models**
  - All distributions belong to the same parametric family, with different parameters
- Gaussian mixture model is a popular mixture model

Scenario

Scenario

Heights of 200 individuals are collected in a survey. Estimate the means of the heights for males and females (μ_M and μ_F).
- Assume the two groups have the same known variance σ^2.

```
#data = dataset("car","Davis");
data[:Height]
```

```
## 200-element Array{Int32,1}:
##  182
##  161
##  161
##  177
##  157
##  170
##  167
##  186
##  178
```


The Model

- Let x be a random variable representing the height of individuals
  - $x_i \in \mathbb{R}, i = 1, 2, \ldots, n$
- Let z be a random variable representing gender
  - $z_i \in M, F$
  - In general, z can be a categorical varibles.
- $$p(z_i) = p^{\mathbb{1}(z_i = M)}(1-p)^{\mathbb{1}(z_i = M)} \qquad p(z_i) = \prod_{c \in \{M,F\}} \pi_c^{\mathbb{1}(z_i = c)}$$

  where $\pi_M = p$, $\pi_F = 1 - p$ (assume p is known)
- Conditional distributions within each class are Gaussian

$$p(x_i | z_i) = \prod_c \mathcal{N}(x_i; \mu_c, \sigma^2)^{\mathbb{1}(z_i = c)}$$

Parameter Estimation - first attempt (MLE)

- We observe i.i.d heights $D = \{x_1, x_2, \ldots, x_n\}$, and want to find MLE estimates for parameters μ_M and μ_F.
  - Unsupervised learning problem: gender is not observed, but parameters are estimated based on gender.
- Probability density for one data point x_i

$$\begin{aligned} p(x_i) &= \sum_{z_i} p(z_i) p(x_i | z_i) \\ &= \sum_{z_i} \prod_c (\pi_c^{\mathbb{1}(z_i = c)}) (\mathcal{N}(x_i; \mu_c, \sigma^2)^{\mathbb{1}(z_i = c)}) \\ &= \sum_{z_i} \prod_c (\pi_c \mathcal{N}(x_i; \mu_c, \sigma^2)^{\mathbb{1}(z_i = c)}) \\ &= \pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2) \end{aligned}$$

Parameter Estimation - first attempt (MLE)

- Probability density for one data point x_i

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

- Joint density or Likelihood for $D = \{x_1, x_2, \ldots, x_n\}$

$$L = p(D) = \prod_{i=1}^{n} (\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2))$$

- Log likelihood

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

- In earlier secenarios, taking a log of likelihood resulted in sum of log terms
  - easier to differentiate
  - here we have a sum in the log that we cannot separate

Parameter Estimation - first attempt (MLE)

- Log likelihood

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

- By symmetry, we only need to look at one of the means; the other will almost follow the same process

$$\begin{aligned}\frac{d}{du}\mathcal{N}(x; \mu, \sigma^2) &= \frac{d}{du}\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right] \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \frac{2(x-\mu)}{2\sigma^2} \\ &= \mathcal{N}(x; \mu, \sigma^2). \frac{(x-\mu)}{\sigma^2}\end{aligned}$$

Parameter Estimation - first attempt (MLE)

- Log likelihood

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

- We know

$$\frac{d}{du}\mathcal{N}(x; \mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2). \frac{(x-\mu)}{\sigma^2}$$

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

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

- This derivative has ratios of exponentials and linear terms
  - Not possible to arrive at a closed form solution for μ_M
  - Gradient-based approaches are applicable

Expectation Maximixation Approach

- Motivation
- Approach
- Julia code
- Examples

Parameter Estimation - first attempt (MLE)

- 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 obtain

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

- This derivative has ratios of exponentials and linear terms
  - Not possible to arrive at a closed form solution for μ_M

Motivation for Expectation Maximization (EM)

- If we knew which subjects were male and female, computing μ_M and μ_F is straightforward.
  - Simplifies to univariate Gaussian parameter estimation
- With Bayes rule, the posterior probability

$$\begin{aligned}p(z_i|x_i) &= \frac{p(x_i|z_i)p(z_i)}{p(x_i)} \\ &= \frac{\prod_c \left(\pi_c \mathcal{N}(x_i; \mu_c, \sigma^2)\right)^{\mathbb{1}(z_i=c)}}{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)}\end{aligned}$$

- The posterior probability that $z_i = M$

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

Motivation for Expectation Maximization (EM)

- Bringing the $\frac{d}{d\mu_M}\ell$ and the posterior probability together
- 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 $\frac{d}{d\mu_M}\ell$

$$\sum_{i=1}^{n} p(M|x_i) \frac{x_i - \mu_M}{\sigma^2} = 0 \implies \mu_M = \frac{\sum_{i=1}^{n} p(M|x_i) x_i}{\sum_{i=1}^{n} p(M|x_i)}$$

Motivation for Expectation Maximization (EM)

- Bringing the $\frac{d}{d\mu_M}\ell$ and the posterior probability together
- 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 $\frac{d}{d\mu_M}\ell$

$$\sum_{i=1}^{n} p(M|x_i)\frac{x_i - \mu_M}{\sigma^2} = 0 \implies \mu_M = \frac{\sum_{i=1}^{n} p(M|x_i)x_i}{\sum_{i=1}^{n} p(M|x_i)}$$

- μ_M is a weighted average of the heights
  - each height is weighted by how likely that person is to be male
- By symmetry, μ_F is weighted ($p(F|x_i)$) average of the heights

Motivation for Expectation Maximization (EM)

- To compute posterior prob. $p(M|x_i)$, we need μ_M and μ_F

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

- To compute μ_M and μ_F, we need $p(M|x_i)$ and $p(F|x_i)$

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

- Strategy: We will fix one and solve for the other, iteratively.
- **EM Algorithm**
  - **E Step:**, we fix parameters μ_M and μ_F, and compute the posterior distribution $p(M|x_i)$ and $p(F|x_i)$
  - **M Step:**, we fix posteriori distribution $p(M|x_i)$ and $p(F|x_i)$ and optimize for μ_M and μ_F
  - Repeat the two steps until the values converge

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**

Writing code for E step in Julia

- Computing $p(M|x_i)$

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

```
function E_step(x,mu_M,mu_F,sigma,p)
    numerator = p*pdf.(Normal(mu_M,sigma),x)
    denom = numerator .+ (1-p).* pdf.(Normal(mu_F,sigma),x);
    post_x = numerator ./denom;
    return post_x;
end
```

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


Writing code for M step in Julia

- Computing μ_M by maximizing the likelihood

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

```
function M_step(x,post_x)
    mu_M = (post_x'*x)./sum(post_x);
    mu_F = ((1.-post_x)'*x)./sum((1.-post_x));
    return mu_M, mu_F;
end
```

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


Writing code for EM approach in Julia

```
function EM(x,mu_M,mu_F,p,sigma)
    maxIter = 1000;
    for i=1:maxIter
        print(i,"\n");
        post_x = E_step(x,mu_M,mu_F,sigma,p);
        mu_M_new, mu_F_new = M_step(x,post_x);
        if(abs(mu_M-mu_M_new)<0.001
           && abs(mu_F-mu_F_new)<0.001)
            break;
        end;
        mu_M = mu_M_new;
        mu_F = mu_F_new;
    end
    return mu_M, mu_F;
end
```

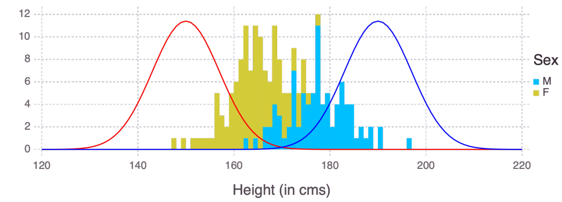
```
## EM (generic function with 2 methods)
```

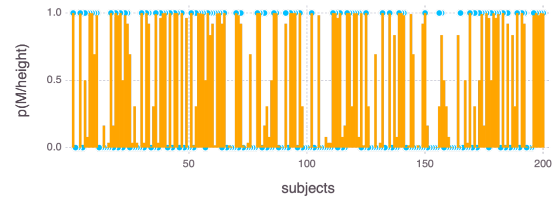
Using EM approach for estimating μ_M and μ_F in Julia

```
#data = dataset("car","Davis");
x = data[:Height];
mu_M=190;
mu_F=150;
p = 0.5;
sigma=7;
EM(x,mu_M,mu_F,p,sigma)
```

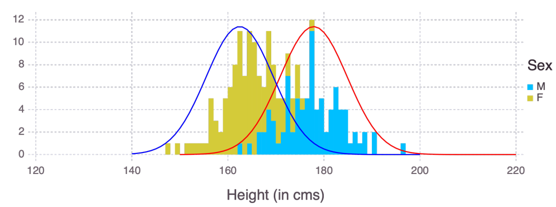
```
## (176.38366334360896, 163.46901164793636)
```

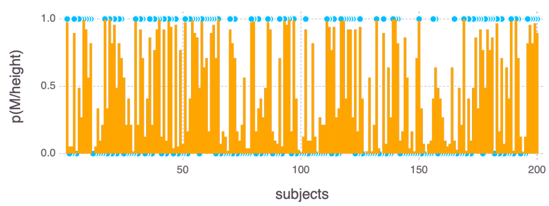

Exploring EM approach: 1st iteration



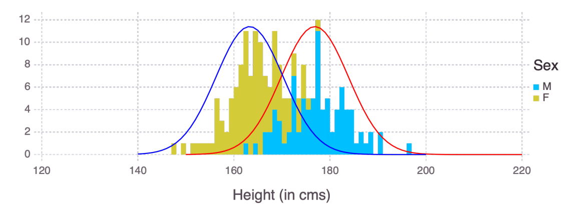


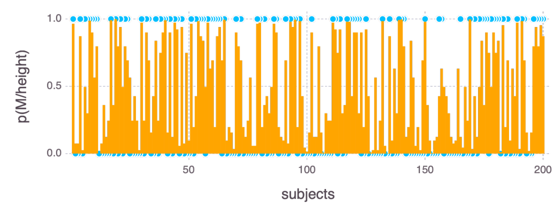
Exploring EM approach: 2nd iteration





Exploring EM approach: 3rd iteration





Exploring EM

- posterior distribution for distinguishable components