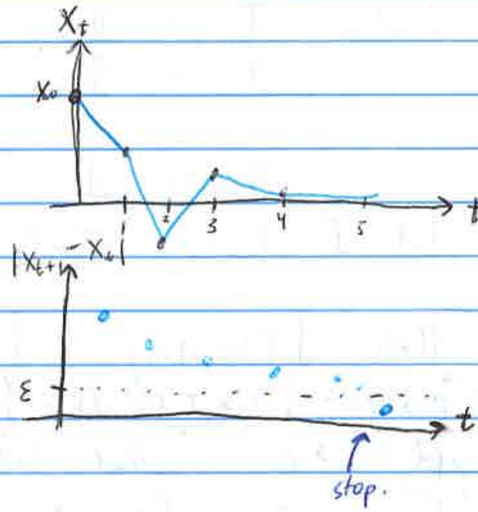
5|4|17

(Newton Raphson?)

N-R Method

$f(x)=0$ solve for x.

1. Guess solution is x_0
2. Calculate $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$
3. Repeat step 2 until $|x_{t+1} - x_t| < \varepsilon$

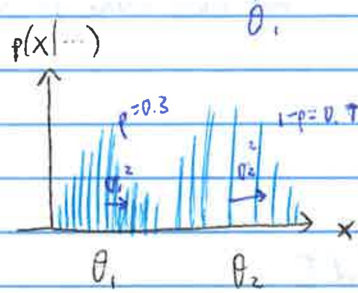


Consider the mixture model:

$$X_1, X_2, \dots, X_n \mid \vec{\theta}_1, \dots, \vec{\theta}_m, \gamma_1, \dots, \gamma_M \overset{iid}{\sim} \sum_{m=1}^{M} \gamma_m P_m(\vec{\theta}_m)$$ (Mixture Distribution)

s.t. $\gamma_1 + \gamma_2 + \dots + \gamma_m = 1$ (otherwise, not a pdf) (likelihood = Mixture)

$$\text{Ex: } X_1, \dots, X_n \mid \underbrace{\theta_1, \sigma_1^2}_{\vec{\theta}_1}, \underbrace{\theta_2, \sigma_2^2}_{\vec{\theta}_2}, \underbrace{\rho}_{\gamma_1, \gamma_2} \overset{iid}{\sim} \underset{\gamma_1}{\rho} \overbrace{N(\theta_1, \sigma_1^2)}^{P_1(\vec{\theta}_1)} + \underset{\gamma_2}{(1-\rho)} \overbrace{N(\theta_2, \sigma_2^2)}^{P_2(\vec{\theta}_2)}$$



Mixture Model

(ex: Heights of Students.
Males = Normal dist.
Females = Normal dist.)

$$P(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid X) \propto \underbrace{p(X \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)}_{\prod_{i=1}^{n} \rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2}} \underbrace{P(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)}_{\substack{P(\theta_1)P(\sigma_1^2)P(\theta_2)P(\sigma_2^2)P(\rho) \\ \propto 1 \quad \propto \frac{1}{\sigma_1^2} \quad \propto 1 \quad \propto \frac{1}{\sigma_2^2} \quad \propto 1}}$$

$$= \left(\prod_{i=1}^{n} \rho \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2} + (1-\rho) e^{-\frac{1}{2\sigma_2^2}(x_i-\theta_2)^2} \right) \underbrace{\frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}}_{\text{Prior}}$$

post.

$$= K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid X)$$

How to ~~make~~ get inference?

Grid search: $G_{\theta_1} = \langle \dots \rangle$, $G_{\sigma_2^2} = \langle \dots \rangle$, ... $\Rightarrow$ Inaccurate, in large.

What if we knew which components each X_i belonged to?

Define: $I_1 := \mathbb{1}_{X_1 \text{ is in } m=1}$
$I_2 := \mathbb{1}_{X_2 \text{ is in } m=1}$
$\vdots$
$I_n := \mathbb{1}_{X_n \text{ " " "}}$

called "latent variables / information"
↓
The I_i's are unobserved but important (can't see them)

Let $I = \{I_1, \dots, I_n\}$

Recall: $f(z) = \int f(z,y)\,dy = \int f(z|y)\,f(y)\,dy$.

very difficult to compute → $\underbrace{p(x|\theta) = \int p(x, I|\theta)\,dI}_{\cancel{1st}} = \int p(x|I,\theta)\,p(I|\theta)\,dI$. easy to compute if get I

→ this is called "Data Augmentation"

The 1st part from before → augmenteng X with the I_i's. (Adding more data to the data)

So $p(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho|X) \propto \int p(X|I, \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)\, p(I|\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)\, \cancel{dI} \cdot p(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \sigma_2^2)\, dI$

$= K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho|X)$

$= \int K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho|X, I)\, dI$

Model Goal: Get $\hat{\theta}_{MAP} = \operatorname{argmax}\{K(\theta|X)\}$ most likely value of the 5 parameters.

Expectation - Maximization Algorithm (1977)

step 1: Guess $\hat{\theta}_{MAP} = \theta_0$ to start

Expectation step → step 2: Compute $I_0 = E[I_0|X, \theta = \theta_0]$

Maximization step → step 3: Consider $\mathcal{L}(\theta; I_0, X) = K(\underbrace{\theta|X, I}_{\text{Body of Integral above.}})\Big|_{I=I_0} dI$

and find $\hat{\theta}_1 = \operatorname{argmax}\{\mathcal{L}(\theta; I, X)\}$ i.e. the MLE procedure

step 4: Repeat steps 2 & 3 until $\|\theta_{t+1} - \theta_t\| < \varepsilon$, where ε is the predefined tolerance level.

E-M Implementation for our 2-normal mixture

step 1: Initialize $\theta_{1,0} = 0$
$\sigma_{1,0}^2 = 1$
$\theta_{2,0} = 0$
$\sigma_{2,0}^2 = 1$
$\rho = 0.5$

step 2: $I_{1,0} = E\left[I_1 \mid X, \theta_1 = \theta_{1,0}, \sigma_1^2 = \sigma_{1,0}^2, \theta_2 = \theta_{2,0}, \sigma_2^2 = \sigma_{2,0}^2, \rho = \rho_0\right]$

$= p(I_1 = 1 \mid x, \cdots) \overset{\text{Bayes Rule}}{=} \dfrac{p(x \mid I_1 = 1, \cdots)\, p(I_1 = 1 \mid \cdots)}{\underbrace{p(x \mid \cdots)}_{p(x \mid I_1 = 1, \cdots) + p(x \mid I_1 = 0, \cdots)\cdot(p(I_1 = 0 \mid \cdots) \cdot (p(I_1 = 1 \mid \cdots))}}$

$I_1 \sim \underset{\text{depends on } \rho}{\text{Bern}}\left(p(I_1 = 1 \mid \cdots)\right)$

$$= \frac{\frac{1}{\sqrt{2\pi\sigma_{1,0}^2}} \cdot e^{-\frac{1}{2\sigma_{1,0}^2}(x_1 - \theta_{1,0})^2} \cdot \rho}{\rho \frac{1}{\sqrt{2\pi\sigma_{1,0}^2}} e^{-\frac{1}{2\sigma_{1,0}^2}(x_1 - \theta_{1,0})^2} + (1-\rho)\frac{1}{\sqrt{2\pi\sigma_{2,0}^2}} e^{-\frac{1}{2\sigma_{2,0}^2}(x_1 - \theta_{2,0})^2}}$$

Then $I_{2,0} = E[I_2 \mid X_2, \cdots]$

$I_{3,0} = E[I_3 \mid X_3, \cdots]$

$\vdots$

$I_{n,0} = E[I_n \mid X_n, \cdots]$

step 3: Consider $\mathcal{L}(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho; I, x) = \overset{\text{generally}}{p(x \mid I, \theta)\, p(I \mid \theta)\, p(\theta)}$

$$= \left(\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i - \theta_1)^2}\right)^{I_i} \cdot \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x_i - \theta_2)^2}\right)^{1 - I_i}\right) \cdot \left(\prod_{i=1}^{n} \rho^{I_i}(1-\rho)^{1-I_i}\right) \cdot \left((\sigma_1^2)^{-1}(\sigma_2^2)^{-1}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n (\sigma_1^2)^{-1}(\sigma_2^2)^{-1}(\sigma_1^2)^{-\frac{1}{2}\Sigma I_i}(\sigma_2^2)^{-\frac{1}{2}\Sigma(1 - I_i)} e^{-\frac{1}{2\sigma_1^2}\Sigma I_i (x_i - \theta_1)^2 - \frac{1}{2\sigma_2^2}\Sigma(1 - I_i)(x_i - \theta_2)^2} \cdot \rho^{\Sigma x_i}(1-\rho)^{\Sigma(1 - I_i)}$$

same as MLE

take log $= l(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho; I, x)$

$$= n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \left(1 + \tfrac{1}{2}\Sigma I_i\right)\ln(\sigma_1^2) - \left(1 + \tfrac{1}{2}\Sigma 1 - I_i\right)\ln(\sigma_2^2) - \frac{1}{2\sigma_1^2}\Sigma I_i (x_i - \theta_1)^2 - \frac{1}{2\sigma_2^2}\Sigma(1 - I_i)(x_i - \theta_2)^2$$

take derivative $= \cdots$

Get $\hat{\theta}_1$ by $\frac{\partial}{\partial\theta_1}[\text{log likelihood}] \overset{\text{set}}{=} 0$

$\hookrightarrow \dfrac{\Sigma x_i I_i}{\sigma_1^2} - \dfrac{2\theta_1 \Sigma I_i}{2\sigma_1^2} = 0$

$\boxed{\hat{\theta}_1 = \dfrac{\Sigma x_i I_i}{\Sigma I_i}}$ like $\bar{X}_{\text{mixture 1}}$

Same for $\boxed{\hat{\theta}_2 = \dfrac{\Sigma x_i (1 - I_i)}{\Sigma(1 - I_i)}}$ like $\bar{X}_{\text{mixture 2}}$ ✓✓

Get $\hat{\sigma}_1^2$ by $\frac{\partial}{\partial\sigma_1^2}[\text{log likelihood}] \overset{\text{set}}{=} 0$

$\hookrightarrow -\dfrac{1 + \frac{1}{2}\Sigma I_i}{\sigma_1^2} + \dfrac{1}{2(\sigma_1^2)^2} \cdot \Sigma I_i (x_i - \theta_1)^2 = 0$

$1 + \frac{1}{2}\Sigma I_i = \frac{1}{2\sigma_1^2}\Sigma I_i (x_i - \theta_1)^2$

$\boxed{\hat{\sigma}_1^2 = \dfrac{\Sigma I_i (x_i - \theta_1)^2}{2\left(1 + \frac{1}{2}\Sigma I_i\right)}}$

likewise, $\boxed{\hat{\sigma}_2^{\,2} = \dfrac{\Sigma(1-I_i)(x_i-0)^2}{2+\Sigma(1-I_i}}$

$\hat{p}$ by $\frac{\partial}{\partial p}[\text{loglik}] \overset{\text{set}}{=} 0$

$\hookrightarrow \dfrac{\Sigma I_i}{p} - \dfrac{\Sigma 1 - I_i}{1-p} = 0$

$\Sigma I_i - p\Sigma I_i = pn - p\Sigma I_i$

$\boxed{\hat{p} = \dfrac{\Sigma I_i}{n}}$

Better versions of I's until we converge