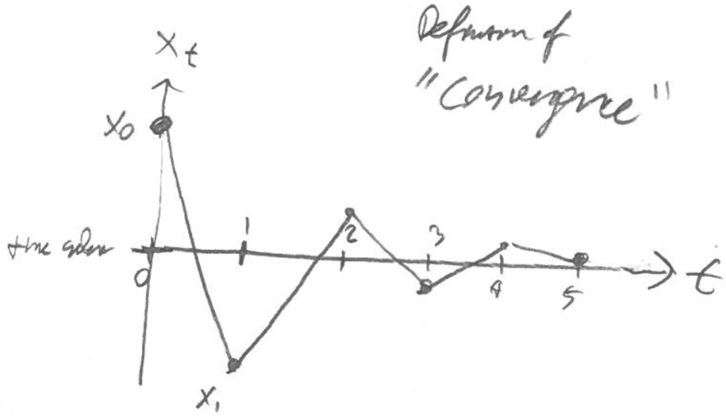
Math 3?? Lec 10 5/9/17

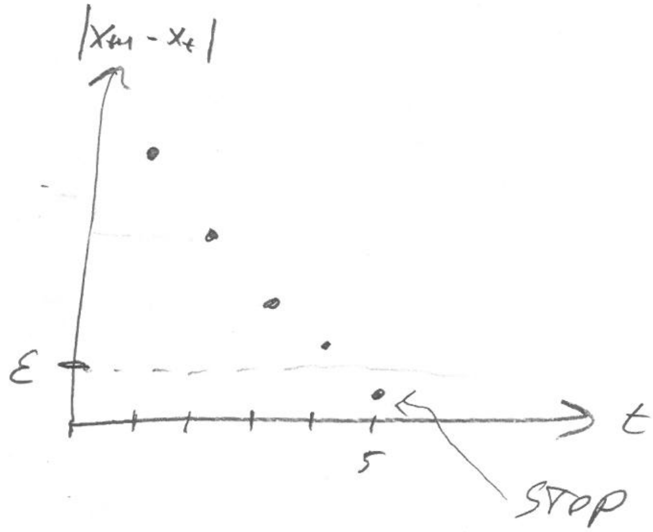
N-R Method: $f(x)=0$ solve for x

① guess solution x_0 } Initialize the algorithm

② Calc. $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ } Iterative algorithm

③ Repeat step 2 until $\underbrace{|x_{t+1} - x_t| < \varepsilon}_{\text{Definition of "Convergence"}}$. Return x^*



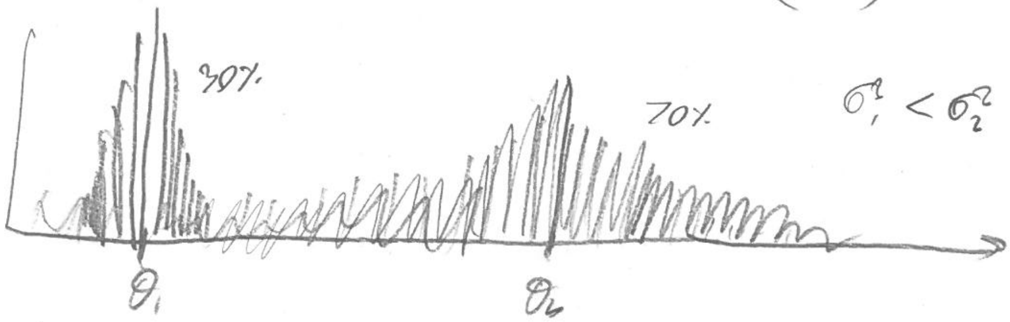


New Problem whose solution is also an ~~iterative~~ algorithm

$$X_1, \ldots, X_n \mid \vec{\theta}_1, \ldots, \vec{\theta}_M, \gamma_1, \ldots, \gamma_M \overset{iid}{\sim} \sum_{m=1}^{M} \gamma_m P_m(X \mid \vec{\theta}_m)$$

$\Rightarrow$ the likelihood is a mixture. The canonical example is $M=2$ $P_m = $ Normal:

$$X_1, \ldots, X_n \mid \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \overset{iid}{\sim} \underbrace{\rho}_{\gamma_1} N(\underbrace{\theta_1, \sigma_1^2}_{\vec{\theta}_1}) + \underbrace{(1-\rho)}_{\gamma_2} N(\underbrace{\theta_2, \sigma_2^2}_{\vec{\theta}_2})$$



Goal: inference for $\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho$

2

let $\underbrace{P(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)}_{\vec{\theta}} = P(\theta_1) P(\sigma_1^2) P(\theta_2) P(\sigma_2^2) P(\rho) = \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}$

$\propto 1, \quad \propto \frac{1}{\sigma_1^2}, \quad \propto 1, \quad \propto \frac{1}{\sigma_2^2}, \quad \propto U(0,1) \propto 1$

$\propto P(x|\theta)P(\theta) =$

$$\Rightarrow P(\vec{\theta}|\vec{X}) \propto \left(\prod_{i=1}^{n} \rho N(\theta_1, \sigma_1^2) + (1-\rho) N(\theta_2, \sigma_2^2)\right) \frac{1}{\sigma_1^2} \frac{1}{\sigma_2^2} = K(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho / X_1 \dots X_n)$$

$\Downarrow$

grid search? Maybe ... but 5 dimensions

Obviously not conjugate nor of known form

$\Rightarrow \hat{\theta}_{MMSE}, \hat{\theta}_{MMP}, \hat{\theta}_{MAE}$??? HARD

$\mathcal{G}_{\theta_1} = \langle \dots \rangle, \; \mathcal{G}_{\sigma_1^2} = \langle \dots \rangle, \; \mathcal{G}_{\theta_2} = \langle \dots \rangle, \; \mathcal{G}_{\sigma_2^2} = \langle \dots \rangle, \; \mathcal{G}_{\rho} = \langle \dots \rangle$

You can get reasonable posterior est's of $\hat{\theta}_{MMSE}$'s.

Is there a Better way?

What if we knew that X_1 "belonged to" $m=2$,

X_2 "belonged to" $m=1$,

X_3 "belonged to" $m=1$,

$\vdots$

X_n "belonged to" $m=1$

Define:

$I_1 = \mathbb{1}_{X_1 \text{ belongs to } m=1}$

$I_2 = \mathbb{1}_{X_2}$ - - - - -

$\vdots$

$I_n = \mathbb{1}_{X_n}$ - - - -

e.g.

$I_1 = 1$

$I_2 = 1$

$I_3 = 0$

$I_4 = 1$

$I_5 = 0$

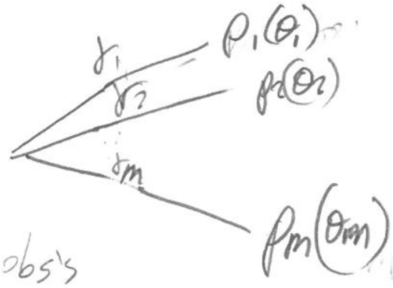
$\vdots$

$I_n = 1$

called latent variables: they are important but unobserved!

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

Mixture distr is



it's as if the obs's "choose a path" (as long as you consider a very large # of obs)

Note $f(z) = \int f(z,y)\,dy = \int f(z|y)\,f(y)\,dy$

let $\theta = \{\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho\}$
let $I = \{I_1, \ldots, I_n\}$

$$P(X|\theta) = \underbrace{\int P(X,I|\theta)\,dI}_{} = \int P(X|I,\theta)\,P(I|\theta)\,dI$$

data augmentation.. add in latent var's!

$$P(\theta|X) \propto P(X|\theta)\,P(\theta) = \int P(X|I,\theta)\,P(I|\theta)\,dI\;P(\theta) = \int P(X|I,\theta)\,P(I|\theta)\,P(\theta)\,dI = k(\theta|X)$$

$$k(\theta|X) = \int k(\theta, I|X)\,dI$$

$$\cancel{k(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho|X) = \int\cdots\int P(X|I_1,\ldots,I_n,\theta_1,\sigma_1^2,\theta_2,\sigma_2^2)\,P(I_1,\ldots,I_n|\theta_1,\sigma_1^2,\theta_2,\sigma_2^2)\,P(\theta_1,\sigma_1^2,\theta_2,\sigma_2^2,\rho)\,dI_1\ldots dI_n}$$

$$\cancel{= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1^2}}}$$

All this doesn't really help us too much. But what if you cared about $\hat{\theta}_{MAP} = \arg\max\{k(\theta|X)\}$

New!

It turns out there was an algorithm called the Expectation-Maximization Algorithm (E-M) (1977) that can converge to $\hat{\theta}_{MAP}$. How?

Step 1: let $\theta = \theta_0$ (a guess of θ... just like N-R!)

Step 3: Define: $\mathcal{L}(\theta; I, X) = P(X|I,\theta)\,P(I|\theta)\,P(\theta)$

Find $\hat{\theta}_1$ via $\ell'(\theta|I,X) = 0$ (i.e. MLE). Note: $\hat{\theta}_{MLE}$ is a function of I, X the max step

Step 2: Let $I_0 = E[I|X, \theta = \theta_0]$ the expectation step

Step 4: Repeat Steps 2-3 until $\|\hat{\theta}_{n+1} - \hat{\theta}_n\| < \epsilon$, i.e. until "convergence".

What would the EM algorithm be in our mixture of two models case ?

Step 1: $\theta_{1,0} = 0,\ \sigma^2_{1,0} = 1,\ \theta_{2,0} = 0,\ \sigma^2_{2,2} = 0,\ \rho = 0.5$

Step 2: $I_{1,0} = E\left[I_1 \mid X_1 \dots X_n,\ \theta_1 = \theta_{1,0},\ \sigma_1^2 = \sigma^2_{1,0},\ \theta_2 = \theta_{2,0},\ \sigma_2^2 = \sigma^2_{2,0},\ \rho = \rho_0\right]$

If $\quad = P(I_1 = 1 \mid X_1 \dots) = \dfrac{P(X \mid I = 1, \dots)\, P(I = 1 \mid \dots)}{P(X \mid \dots)}$

$Q = \mathbb{1}_A \sim \text{Bern}(P(A))$

$E[Q] = P(A)$

$P(X \mid \dots) = P(X \mid I = 1, \dots) + P(X \mid I = 0, \dots)$

$P(A \mid B) = \dfrac{\overbrace{P(B \mid A)\, P(A)}^{P(A,B)}}{P(B)}$

$$= \frac{\frac{1}{\sqrt{2\pi\sigma^2_{1,0}}} e^{-\frac{1}{2\sigma^2_{1,0}}(X_1 - \theta_{1,0})^2}\, \rho}{\frac{1}{\sqrt{2\pi\sigma^2_{1,0}}} e^{-\frac{1}{2\sigma^2_{1,0}}(X_1 - \theta_{1,0})^2}\, \rho + \frac{1}{\sqrt{2\pi\sigma^2_{2,0}}} e^{-\frac{1}{2\sigma^2_{2,0}}(X_1 - \theta_{2,0})^2}\, (1-\rho)}$$

$I_{2,0} = \dots$

$\vdots$

$I_{n,0} = \dots$

Step 3: $$\mathcal{L}(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho \mid I_1 \dots I_n, X_1, \dots, X_n) = \left(\prod_{i=1}^n \overset{P(X \mid I, \theta)}{\overset{\parallel}{P(X_i \mid I_i, \theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \rho)}}\right) \left(\overset{P(I \mid \theta)}{\overset{\parallel}{\prod_{i=1}^n \rho^{I_i}(1-\rho)^{1-I_i}}}\right) \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2}$$

$$= \frac{1}{\sigma_1^2} \frac{1}{\sigma_2^2} \prod_{i=1}^n N(\theta_1, \sigma_1^2)^{I_i}\, N(\theta_2, \sigma_2^2)^{1-I_i} \prod \rho^{I_i}(1-\rho)^{1-I_i}$$

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

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{2n} (\sigma_1^2)^{-1} (\sigma_2^2)^{-1} (\sigma_1^2)^{-\sum I_i/2}\, (\sigma_2^2)^{\frac{\sum(1-I_i)}{2}}\, e^{-\frac{1}{2\sigma_1^2}\sum I_i (X_i - \theta_1)^2}\ \wedge\ e^{-\frac{1}{2\sigma_2^2}\sum (1-I_i)(X_i - \theta_2)^2}\, \rho^{\sum I_i} (1-\rho)^{\overset{n - \sum I_i}{\overset{\parallel}{\sum 1 - I_i}}}$$

$$\ell(\dots \mid \dots) = 2n \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \left(\frac{n}{2}+1\right) \ln(\sigma_1^2) - \left(\frac{n}{2}+1\right) \ln(\sigma_2^2) - \quad - \quad + \sum I_i \ln(\rho) + \left(n - \sum I_i\right) \ln(1-\rho)$$

$$\frac{1}{2\sigma_1^2} \sum I_i (X_i^2 - 2X_i\theta + \theta^2) = \frac{\sum I_i X_i^2}{2\sigma_1^2} - \frac{\theta \sum X_i I_i}{\sigma_1^2} + \frac{\theta^2 \sum I_i}{2\sigma_1^2}$$

[illegible] $\hat{\theta}_1$ set $\frac{\partial}{\partial\theta_1}[\ell\cdots]=0$

$$\Rightarrow \frac{\sum X_i I_i}{\cancel{\sigma_1^2}} + \frac{\theta_1 \sum I_i}{\cancel{\sigma_1^2}} = 0 \Rightarrow \hat{\theta}_1 = \frac{\overbrace{\sum X_i I_i}^{\text{sum of } x\text{'s for } m=1}}{\underbrace{\sum I_i}_{\#\text{ in } m=1}} \approx \bar{X}_1 \quad \text{makes sense..}$$

$$\hat{\theta}_2 = \ldots \frac{\sum X_i (1-I_i)}{\sum 1-I_i} \approx \bar{X}_2$$

$\hat{\sigma}_1^2$ set $\frac{\partial}{\partial\sigma_1^2}[\ell\ldots]=0$

$$\Rightarrow \frac{-\left(\frac{\sum I_i}{2}+1\right)2}{\cancel{\sigma_1^2}} + \frac{1}{\cancel{2}(\sigma_1^2)^{\cancel{2}}} \sum I_i (X_i-\theta_1)^2 = 0$$

$$\Rightarrow -(\sum I_i + 2) + \frac{\sum I_i (X_i-\theta_1)^2}{\sigma_1^2} = 0 \Rightarrow \hat{\sigma}_1^2 = \frac{\overbrace{\sum I_i (X_i-\theta_1)^2}^{SSE \text{ for } m=1}}{\sum I_i + 2} \approx \text{sample var for } m=1$$

$$\hat{\sigma}_2^2 = \ldots \frac{\sum (1-I_i)(X_i-\theta_2)^2}{\sum(1-I_i)+2} \approx \text{sample var for } m=2$$

$\hat{\rho}$ set $\frac{\partial}{\partial\rho}[\ell(\ldots)]=0$

$$\frac{\sum I_i}{\rho} - \frac{n-\sum I_i}{1-\rho} = 0$$

$$\Rightarrow \frac{\sum I_i}{\rho} = \frac{n-\sum I_i}{1-\rho} \Rightarrow \sum I_i - \cancel{\rho \sum I_i} = \rho n - \cancel{\rho \sum I_i} \Rightarrow \hat{\rho} = \frac{\sum I_i}{n} \quad \text{why? obvious...}$$

Now iterate!!!