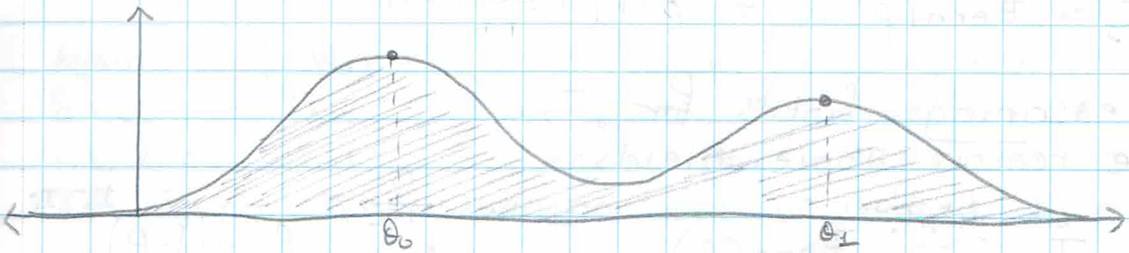
LECTURE 20

Let $X_1, \ldots, X_n \overset{exch}{\sim} \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$

So for X_i, the density looks like



The joint density is the product of the individual densities.

$$\text{So } P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho) = \prod_{i=1}^{n} \rho \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma_0^2}(x_i-\theta_0)^2} + (1-\rho)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2}$$

This is also our likelihood.

We cannot compute $\hat{\theta}_{MLE}$ by hand (i.e. take $\ln$ of the likelihood, compute the derivative with respect to $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2$, & ρ & set it equal to zero). So what is our goal?

GOAL: ESTIMATE $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$

How? ① compute the MLE for $\theta_0, \ldots \rho$ which is hard.

② Grid sample $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$ but this would not be very accurate.

Numerical method ③ 5 Dimensional Newton-Raphson BUT YOU STILL NEED TO TAKE THE DERIVATIVE as in ①.

Let us explore a different option...

What if we knew X_1 belongs to $N(\theta_0, \sigma_0^2)$ &

X_2 belongs to $N(\theta_1, \sigma_1^2)$ &

$\vdots$

X_n belongs to $N(\theta_1, \sigma_1^2)$

Define $I_i := \mathbb{1}_{X_i} = \begin{cases} 1 & \text{if } X_i \text{ belongs to } N(\theta_0, \sigma_0^2) \\ 0 & \text{if } X_i \text{ is NOT } N(\theta_0, \sigma_0^2) \text{ i.e. must be } N(\theta_1, \sigma_1^2) \end{cases}$

Ex: Based off what we defined above, $I_1 = 1$, $I_2 = 0$ & $I_n = 0$

Now what is the distribution of I_1? That is,

What is the probability X_1 belongs to $N(\theta_0, \sigma_0^2)$?
Well,

$$X_1 \sim \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$$

Hence X_1 is $N(\theta_0, \sigma_0^2)$ with probability ρ.
Hence $I_1 \sim \text{Bern}(\rho) \;\; = \rho^{I_1}(1-\rho)^{1-I_1}$

The same reasoning holds for I_2.
Repeating the reason above yields

$$I_1, I_2, \ldots, I_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\rho) \qquad \prod_{i=1}^{n} (\rho^{I_i})(1-\rho)^{1-I_i}$$

Hence $p(\vec{I} \mid \rho) = p(I_1, \ldots, I_n \mid \rho) = \underbrace{p(I_1 \mid \rho) p(I_2 \mid \rho) \ldots p(I_n \mid \rho)}$
$= \text{Product of Bernoulli's}$
$= \text{Bin}(n, \rho)$

So $\vec{I} \mid \rho \sim \text{Bin}(n, \rho)$

$$P(\vec{X}, \vec{I} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho) = P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho, I) P(\vec{I} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$$
$$= P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I) P(\vec{I} \mid \rho)$$

Note $P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho, \vec{I}) = P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I)$. Why?
B/c you are given $\vec{I}$. Remember, $\vec{I}$ is the "INDICATOR RANDOM VARIABLE." It tells you if $X_i \sim N(\theta_0, \sigma_0^2)$ OR if $X_i \sim N(\theta_1, \sigma_1^2)$. So given $\vec{I}$, you do NOT need know ρ.

Also note $P(\vec{I} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho) = P(\vec{I} \mid \rho)$. Why? Because I_j, (the indicator r.v. of X_j) is either 1 w.p. ρ or
0 w.p. $1-\rho$.

Hence $\vec{I}$ depends ONLY on ρ.

Now remember $X_1 \sim \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$. What is $P(X_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I)$? We are given I. $I_1 = 0$ OR $I_1 = 1$.
We need $P(X_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I)$ to "switch" from $N(\theta_0, \sigma_0^2)$ if $I_1 = 1$ to $N(\theta_1, \sigma_1^2)$ if $I_1 = 0$. Hence

$$P(X_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I) = \left(N(\theta_0, \sigma_0^2)\right)^{I_1} \left(N(\theta_1, \sigma_1^2)\right)^{1-I_1}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i - \theta_0)^2}\right)^{I_1} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_1 - \theta_1)^2}\right)^{1-I_1}$$

and in general

$$P(X_j \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x_j-\theta_0)^2}\right)^{I_j} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_j-\theta_1)^2}\right)^{1-I_j}$$

Again, why? Well $X_j \sim \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)$. Since in $P(X_j \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I)$, we are given $\vec{I}$, we know what I_j equals to. $I_j = 0$ or $I_j = 1$. If $I_j = 0$, then based on how we defined I_j on page 1, we X_j is $N(\theta_1, \sigma_1^2)$. If $I_j = 1$, then we want $P(X_j \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, I)$ to be $N(\theta_0, \sigma_0^2)$ & the way we switch between them is to use the above equation b/c if $I_j = 0$ you get $N(\theta_1, \sigma_1^2)$ & if $I_j = 1$ you get $N(\theta_0, \sigma_1^2)$.

Hence

$$P(\vec{X}, \vec{I}) = P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \vec{I}) \; P(\vec{I} \mid \rho)$$

$$= \prod_{i=1}^{n} \Bigg(\underbrace{\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x_i-\theta_0)^2}}_{\alpha}\Bigg)^{I_i} \Bigg(\underbrace{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)^2}}_{\beta}\Bigg)^{1-I_1} \left(\rho^{I_i}\right)(1-\rho)^{1-I_i}$$

$$= \prod_{i=1}^{n} (\rho\alpha)^{I_i} \left((1-\rho)\beta\right)^{1-I_i}$$

We then showed what we did is mathematically valid.

So originally, our likelihood would have been $P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$! We then "added $\vec{I}$" in & we computed our new likelihood, $P(\vec{X}, \vec{I} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$. [This process is called DATA AUGMENTATION."] However, the question remains, WHY DID WE DO THIS? The answer is that if wanted $\theta_0^{MLE}, \theta_1^{MLE}, \sigma_0^2{}^{MLE}, \sigma_1^2{}^{MLE}, \rho^{MLE}$, we would need to take the derivative of the log likelihood, $\ln(P(\vec{X} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho))$ & set it equal to 0 & solve for $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$. However, this is too difficult to do by hand.

It turns out, we can take the derivative of the log likelihood, $\ln(P(\vec{X}, \vec{I} \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$ & get our MLE's. We have:

① $\hat{\theta}_0^{MLE} = \dfrac{\sum x_i I_i}{n_0}$

② $\hat{\theta}_1^{MLE} = \dfrac{\sum x_i (1 - I_i)}{n_1}$

③ $\hat{p}_{MLE} = \dfrac{n_0}{n} = \dfrac{n_0}{n_0 + n_1}$

④ $\widehat{\sigma_0^2}_{MLE} = \dfrac{\sum (x_i - \theta_0)^2 I_i}{n_0}$

⑤ $\widehat{\sigma_1^2}_{MLE} = \dfrac{\sum (x_i - \theta_1)^2 (1 - I_i)}{n_1}$

where $n_0 := \sum I_i$

$n_1 := \sum 1 - I_i$

Remark: Note that $\hat{\theta}_0^{MLE}$, $\hat{\theta}_1^{MLE}$, $\hat{p}_{MLE}$ are like our "$\bar{X}$"! These MLE's are "natural averages" in the following sense:

① for $\hat{\theta}_0^{MLE}$, add up the # of X_i's that belong to the $N(\theta_0, \sigma_0^2)$ mixture & divide by the number of such terms.

② Similar interpretation for $\hat{\theta}_1^{MLE}$ & $\hat{p}_{MLE}$.

③ $\widehat{\sigma_K^2}_{MLE}$ is the varience if θ_K is known for $k = 0, 1$.

Okay, let's recap...
We augmented in some data which allowed us to get the MLE's, which was something we were NOT able to do before. However, looking at our MLE's, there appears to be a problem: Our MLE's REQUIRES US TO KNOW I. BUT WE DO NOT KNOW $\vec{I}$ AHEAD OF TIME! Hence, the above equations, as it stands now, are useless.

BUT, what if we can estimate $\vec{I}$? That is, what if we can estimate I_i?
Before answering this question, we note for an INDICATOR R.V. X, we have $P(X) = E(X)$.
We now return to the question.

Define $\hat{I}_1 := E[I_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho]$

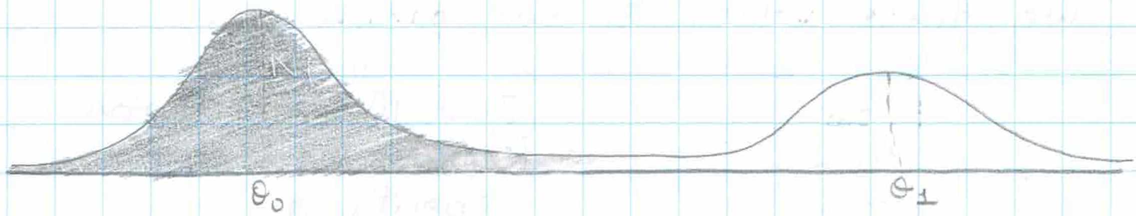
$= P(I_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$

since $\hat{I}_1$ is an indicator r.v.

What is $P(I_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho)$?

Remember,

$$I_1 = \begin{cases} 1 & \text{if } X_1 \text{ belongs to } N(\theta_0, \sigma_0^2) \\ 0 & \text{if } X_1 \text{ belongs to } N(\theta_1, \sigma_1^2) \end{cases}$$

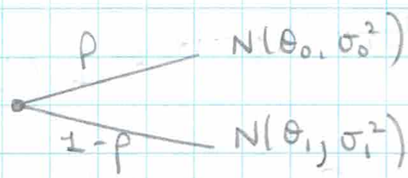


That is, how likely is it X_1 belongs to $N(\theta_0, \sigma_0^2)$ vs belonging to $N(\theta_1, \sigma_1^2)$?

From the picture, it is clear

$$\hat{I}_1 = P(I_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho) = \frac{\rho N(\theta_0, \sigma_0^2)}{\rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2)} = \frac{\text{frequency}}{\text{sum}}$$

i.e., the probability of being in the $N(\theta_0, \sigma_0^2)$ distribution over our entire sample space. Note we multiply by ρ since



Hence

$$\hat{I}_1 = P(I_1 \mid \theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho) = \frac{\rho\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_1-\theta_0)^2}\right)}{\rho\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_1-\theta_0^2)}\right) + (1-\rho)\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot e^{-\frac{1}{2\sigma_1^2}(X_1-\theta_1)^2}\right)}$$

AND MORE GENERALLY, we call the following equation ⑥

$$\boxed{\hat{I}_i = \frac{\rho\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i-\theta_0)^2}\right)}{\rho\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i-\theta_0)^2}\right) + (1-\rho)\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i-\theta_1)^2}\right)}}$$

BUT wait... to compute $\hat{I}_i$, we need $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$. This is bad because these are the parameters we WANT

TO ESTIMATE! We do NOT have them ahead of time. So again it seems like this is useless... But wait! Here is a solution to this called the E-M algorithm

EXPECTATION-MAXIMATION

STEP ONE

Guess values for $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$. We denote these initial guesses by indexing via 0 (i.e guess 0) That is, we guess values for our parameters:

$$\begin{bmatrix} \theta_{0,0} \\ \theta_{1,0} \\ \sigma_{0,0}^2 \\ \sigma_{1,0}^2 \\ \rho_0 \end{bmatrix}$$

(Just like in Newton Raphson we first specify a guess | starting point)

STEP TWO (called E-step)

Calculate $\hat{I}_i$, the augmented data, by using equation (6). So we compute

$$\hat{I}_i = \frac{\rho_0 \frac{1}{\sqrt{2\pi\sigma_{0,0}^2}} e^{-\frac{1}{2\sigma_{0,0}^2}(X_i - \theta_{0,0})^2}}{\rho_0 \frac{1}{\sqrt{2\pi\sigma_{0,0}^2}} e^{-\frac{1}{2\sigma_{0,0}^2}(X_i - \theta_{0,0})^2} + (1-\rho_0) e^{-\frac{1}{2\sigma_{1,0}^2}(X_i - \theta_{1,0})^2}}$$

$\forall\, i = 1, 2, 3, \ldots, n$. So we get $\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_n$ which we can really think of as $\hat{\hat{I}}_1, \hat{\hat{I}}_2, \ldots, \hat{\hat{I}}_n$ since we used estimates for $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2, \rho$.

STEP THREE (called M-step)

Compute $\hat{\theta}_{0\,MLE}, \hat{\theta}_{1\,MLE}, \hat{\sigma}^2_{0\,MLE}, \hat{\sigma}^2_{1\,MLE}, \hat{\rho}_{MLE}$ (i.e, equations (1) - (5) by using the $\hat{\hat{I}}_i$'s from step two.) we really

get $\hat{\hat{\theta}}_{0\,MLE}, \hat{\hat{\theta}}_{1\,MLE}, \hat{\hat{\sigma}}_0^2{}_{MLE}, \hat{\hat{\sigma}}_1^2{}_{MLE}, \hat{\hat{\rho}}_{MLE}$ since we used estimates of I_i's.

STEP FOUR: REPEAT STEP TWO WITH PARAMETER GUESSES FROM STEP THREE AND ITERATE UNTIL $\|\theta_{t+1} - \theta_t\| < \varepsilon$, a predetermined TOLERANCE LEVEL.

Note that one may show this process does converge.

At the end of the E-M, we get estimates of our parameters $\theta_0, \theta_1, \sigma_0^2, \sigma_1^2$, & ρ. Thus we have achieved our goal listed on page 1.

As a last remark, note that the E-M is useful for when the typical MLE case breaks down b/c of the complexity of taking the derivative of the log likelihood.