Consider the complete log-likelihood:

$$\ell(p, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2 | \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{i=1}^{n} \{Z_i \log f_1(Y_i; \boldsymbol{\mu}_1, \Sigma_1) + (1 - Z_i) \log f_2(Y_i; \boldsymbol{\mu}_2, \Sigma_2)$$
$$+Z_i \log p + (1 - Z_i) \log(1 - p)\} .$$

We could consider the Z_i's as "weights" which represent our *current* beliefs in which density each datum come from.

If we were given this information (the complete data) our belief weights are 0s or 1s.

Instead, we have current beliefs based on other parameters...?

based on current knowledge of model parameters

Given what our belief is in the weights of the data, what is our estimate of the model parameters?

this seems circular (and it is → iterative procedure)

$$\underset{\text{}}{\underline{\theta}^{(k)}} \longrightarrow \underset{\text{"E"}}{\text{weights}} \longrightarrow \underset{\text{"M"}}{\underline{\theta}^{(k+1)}}$$

$$\hat{\underline{\mu}}_1^{(k+1)} = \frac{\sum_{i=1}^{n} w_i^{(k)} \underline{y}_i}{\sum_{i=1}^{n} w_i^{(k)}}$$

$$\hat{\underline{\mu}}_2^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - w_i^{(k)}) \underline{y}_i}{\sum_{i=1}^{n} (1 - w_i^{(k)})}$$

$$\hat{\Sigma}_1^{(k+1)} = \frac{1}{\sum_{i=1}^{n} w_i^{(k)}} \sum_{i=1}^{n} w_i^{(k)} (\underline{y}_i - \hat{\underline{\mu}}_1^{(k)})(\underline{y}_i - \hat{\underline{\mu}}_1^{(k)})^T$$

$$\hat{\Sigma}_2^{(k+1)} = \frac{1}{\sum_{i=1}^{n} (1 - w_i^{(k)})} \sum_{i=1}^{n} (1 - w_i^{(k)}) (\underline{y}_i - \hat{\underline{\mu}}_2^{(k)})(\underline{y}_i - \hat{\underline{\mu}}_2^{(k)})^T$$

$$\hat{p}^{(k+1)} = \frac{\sum_{i=1}^{n} w_i^{(k)}}{n}$$

This is the basic intuition for the EM algorithm. We will view our data $\boldsymbol{Y}$ as incomplete and imagine there is missing data $\boldsymbol{Z}$ that would make the problem simpler if we had it. The EM algorithm then follows:

we don't actually have $\underline{Z}$! need to maximize something that is only a function of $\underline{\theta}$ and $\underline{Y}$.
$\Rightarrow$ conditional expectation!

① Write down the joint likelihood of the "complete" data $(\underline{Y}, \underline{Z})$, $L_c(\underline{\theta} | \underline{Y}, \underline{Z})$

② E-step: compute conditional expectation of $\log L_c(\underline{\theta} | \underline{Y}, \underline{Z})$ given $\underline{Y}$ assuming true parameter is $\underline{\theta}^{(v)}$

$$Q(\underline{\theta}, \underline{\theta}^{(v)}, \underline{Y}) = E_{\theta^{(v)}}\left[\log L_c(\underline{\theta} | \underline{Y}, \underline{Z}) \,|\, \underline{Y} \right]$$

current value of est. of θ in this iterative procedure.

$$= \int \log L(\underline{\theta} | \underline{Y}, \underline{Z}) \, f_{\underline{Z}|\underline{Y}}(\underline{Z} | \underline{Y}, \underline{\theta}^{(v)}) \, d\underline{Z}$$

③ M-step: Maximize $Q(\underline{\theta}, \underline{\theta}^{(v)}, \underline{Y})$ wrt $\underline{\theta}$ ($\underline{\theta}^{(v)}$ fixed).

$$\theta^{(v+1)} = \underset{\underline{\theta}}{\operatorname{argmax}} \; Q(\underline{\theta}, \underline{\theta}^{(v)}, \underline{Y}).$$

repeat ② + ③ until convergence (values of $\underline{\theta}^{(v+1)}$ aren't changing much).

Example (Two-Component Mixture, Cont'd): The EM algorithm for the two-component Gaussian mixture model is

⓪ Start with $\hat{\underline{\theta}}^{(0)}$ (initialize).

For $v = 0, 1, 2, \ldots$

① E-step: $Q(\underline{\theta}, \underline{\theta}^{(v)}, \underline{Y}) = E_{\theta^{(v)}}\left[\log L_c(\underline{\theta} | \underline{Y}, \underline{Z}) \underline{Y}\right] = \sum_{i=1}^{n} w_i^{(v)} \log f_1(\underline{Y}_i; \underline{\mu}_1, \Sigma_1) + (1 - w_i^{(v)}) \log f_2(Y_i; \underline{\mu}_2, \Sigma_2) +$
$w_i^{(v)} \log p + (1 - w_i^{(v)}) \log(1-p)$

where $w_i = E_{\theta^{(v)}}(Z_i | Y) = \dfrac{p^{(v)} f_1(Y_i; \underline{\mu}_1^{(v)}, \Sigma_1^{(v)})}{p^{(v)} f_1(Y_i; \underline{\mu}_1^{(v)}, \Sigma_1^{(v)}) + (1 - p^{(v)}) f_2(Y_i; \underline{\mu}_2^{(v)}, \Sigma_1^{(v)})}$

② M-step: see page 6

Your Turn: Implement the EM algorithm for the two-component mixture model on our example data.