1.1 Convergence of the EM algorithm

We will show that $\ell\left(\hat{\boldsymbol{\theta}}^{(k+1)}\right) \geq \ell\left(\hat{\boldsymbol{\theta}}^{(k)}\right)$.

In other words, each step of the EM algorithm leads to an improvement of the log-likelihood value.

Thus, if the likelihood is well behaved, it will achieve the MLE, otherwise the EM will achieve a local maxima (if there is one).

↳ bounded, unimodal.

y = observed data
z = hidden data

We know $f_{Z|Y}(z|y;\boldsymbol{\theta}) = \frac{f_{YZ}(y,z;\boldsymbol{\theta})}{f_Y(y|\boldsymbol{\theta})}$. def'n of conditional density true for any y, z

$$\Rightarrow f_Y(y;\underline{\theta}) = \frac{f_{YZ}(y,z;\underline{\theta})}{f_{Z|Y}(z|y;\underline{\theta})}$$ just rewritten (not clear why).

Assume we observe $\boldsymbol{y} = (y_1, \ldots, y_n)$, then

$$L(\underline{\theta}|\underline{Y}) = f_{\underline{Y}}(\underline{Y};\underline{\theta}) = \frac{f_{\underline{Y}\underline{Z}}(\underline{Y},\underline{Z};\underline{\theta})}{f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\underline{\theta})}$$ (if iid, product of univariate densities)

holds for any $\underline{Z}$!

$$\Rightarrow \ell(\underline{\theta}) = \log f_{\underline{Y}}(\underline{Y};\underline{\theta}) = \log f_{\underline{Y}\underline{Z}}(\underline{Y},\underline{Z};\underline{\theta}) - \log f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\underline{\theta}) = \underbrace{\ell_c(\underline{\theta}|\underline{Y},\underline{Z})}_{\text{log likelihood of "complete data" } \underline{Y},\underline{Z}} - \underbrace{\ell(\underline{\theta}|\{\underline{Z}|\underline{Y}\})}_{\text{"conditional likelihood"}}$$

log likelihood of data $\underline{Y}$
want to optimize

So, in order to show that $\ell\left(\hat{\boldsymbol{\theta}}^{(k+1)}\right) \geq \ell\left(\hat{\boldsymbol{\theta}}^{(k)}\right)$, this is the same as

$$Q\left(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}\right) - H\left(\hat{\theta}^{(k+1)}, \hat{\theta}^{(k)}\right) \geq Q\left(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}\right) - H\left(\hat{\theta}^{(k)}, \hat{\theta}^{(k)}\right).$$

$\Rightarrow$ take expected value wrt $\underline{Z}|\underline{Y};\hat{\theta}^{(k)}$

$$\int \ell(\underline{\theta}|\underline{Y}) f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\hat{\theta}^{(k)})\,d\underline{Z} = \int \log f_{\underline{Y}\underline{Z}}(\underline{Y},\underline{Z};\theta) f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\hat{\theta}^{(k)})\,d\underline{Z} - \int \log f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\theta) f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\hat{\theta}^{(k)})\,d\underline{Z}$$

$$\ell(\underline{\theta}|\underline{Y}) \underbrace{\int f_{\underline{Z}|\underline{Y}}(\underline{Z}|\underline{Y};\hat{\theta}^{(k)})\,d\underline{Z}}_{=1} = \ldots$$

$$\Rightarrow \ell(\underline{\theta}|\underline{Y}) = Q(\underline{\theta}, \hat{\theta}^{(k)}) - H(\underline{\theta}, \hat{\theta}^{(k)})$$

(function of θ!)

Step 1: Show that $H(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(k)})$ is maximized when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$.

i.e. $H(\underline{\hat{\theta}}^{(k)}, \underline{\hat{\theta}}^{(k)}) \geq H(\underline{\theta}, \hat{\theta}^{(k)})$ for any $\underline{\theta} \in \Theta$.

Recall: Jensen's Inequality. A function Φ is convex if $\Phi(\frac{x_1+x_2}{2}) \leq \frac{1}{2}\Phi(x_1) + \frac{1}{2}\Phi(x_2)$. Then

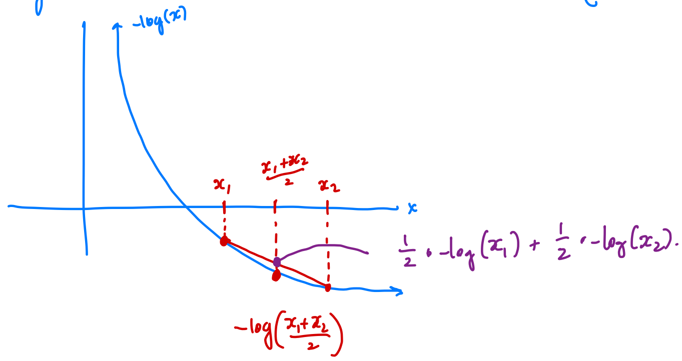
$$\Phi(\mathrm{E}[g(X)]) \leq \mathrm{E}[\Phi(g(X))],$$

where g is a real-valued integrable function.

$\Longleftrightarrow$

$\Phi\left(\int g(x) f(x)\,dx\right) \leq \int \Phi(g(x)) f(x)\,dx$ where $f(x)$ is density of X.

Fact : $-\log$ is convex



Consider $H(\underline{\hat{\theta}}^{(k)}, \underline{\hat{\theta}}^{(k)}) - H(\underline{\theta}, \underline{\hat{\theta}}^{(k)})$. WTS this is positive (non negative). $\forall \underline{\theta}$

$$H(\underline{\theta}, \hat{\theta}^{(k)}) = \int \log(f_{Z|Y}(\underline{z}|\underline{y}; \underline{\theta}))\, f_{Z|Y}(\underline{z}|\underline{y}; \underline{\hat{\theta}}^{(k)})\,dz$$

$$\Rightarrow H(\underline{\hat{\theta}}^{(k)}, \underline{\hat{\theta}}^{(k)}) - H(\underline{\theta}, \underline{\hat{\theta}}^{(k)}) = \int \left(\log(f_{Z|Y}(\underline{z}|\underline{y}; \underline{\hat{\theta}}^{(k)})) - \log(f_{Z|Y}(\underline{z}|\underline{y}; \underline{\theta})) \right) f_{Z|Y}(\underline{z}|\underline{y}; \underline{\hat{\theta}}^{(k)})\,dz$$

$$= \int -\log\left(\frac{f_{Z|Y}(\underline{z}|\underline{y}; \underline{\theta})}{f_{Z|Y}(\underline{z}|\underline{y}; \underline{\hat{\theta}}^{(k)})} \right) f_{Z|Y}(\underline{z}|\underline{y}; \underline{\hat{\theta}}^{(k)})\,dz$$

$$\geq -\log \int \frac{f_{Z|Y}(\underline{z}|\underline{y}; \underline{\theta})}{f_{Z|Y}(\underline{z}|\cancel{\underline{y}}; \underline{\hat{\theta}}^{(k)})} f_{Z|Y}(\underline{z}|\cancel{\underline{y}}; \underline{\hat{\theta}}^{(k)})\,dz$$

$$= -\log \underbrace{\int f_{Z|Y}(\underline{z}|\underline{y}; \underline{\theta})\,dz}_{1}$$

$$= 0$$

$$\Rightarrow H(\underline{\hat{\theta}}^{(k)}, \underline{\hat{\theta}}^{(k)}) \geq H(\underline{\theta}, \underline{\hat{\theta}}^{(k)}) \quad \forall \underline{\theta}. //$$

Step 2: Find a $\hat{\boldsymbol{\theta}}^{k+1}$ that will optimize $\underset{\sim}{Q}$.

Recall goal is to find $\hat{\theta}^{(k+1)}$ s.t. $\ell(\hat{\underline{\theta}}^{(k+1)}) \geq \ell(\hat{\underline{\theta}}^{(k)})$ & $\ell(\underline{\theta}) = Q(\underline{\theta}, \hat{\underline{\theta}}^{(k)}) - H(\underline{\theta}, \hat{\underline{\theta}}^{(k)})$

This is the EM algorithm.

Let $\hat{\theta}^{(k+1)} = \underset{\underline{\theta}}{\operatorname{argmax}}\ Q(\underline{\theta}, \hat{\theta}^{(k)})$.

We know $H(\hat{\underline{\theta}}^{(k+1)}, \hat{\underline{\theta}}^{(k)}) \leq H(\hat{\underline{\theta}}^{(k)}, \hat{\underline{\theta}}^{(k)})$ because true for all $\underline{\theta}$

& $Q(\hat{\underline{\theta}}^{(k+1)}, \hat{\underline{\theta}}^{(k)}) \geq Q(\hat{\underline{\theta}}^{(k)}, \hat{\underline{\theta}}^{(k)})$ by optimization.

So, $\ell(\hat{\underline{\theta}}^{(k)}) = Q(\hat{\underline{\theta}}^{(k)}, \hat{\underline{\theta}}^{(k)}) - H(\hat{\underline{\theta}}^{(k)}, \hat{\underline{\theta}}^{(k)})$

$\leq Q(\hat{\underline{\theta}}^{(k+1)}, \hat{\underline{\theta}}^{(k)}) - H(\hat{\underline{\theta}}^{(k+1)}, \hat{\underline{\theta}}^{(k)}) = \ell(\hat{\theta}^{(k+1)})$ ✓

Example (Two-Component Mixture, Cont'd):

$Q(\theta, \hat{\theta}^{(k)}) = \int \log f_{YZ}(Y, z; \underline{\theta}) f_{Z|Y}(z|y; \hat{\theta}^{(k)})\, dz$

For the Gaussian mixture, the complete log-likelihood:

$$\log f_{YZ}(\underline{Y}, \underline{Z}; \theta) = \sum_{i=1}^{n} \{ Z_i \log f_1(Y_i; \mu_1, \Sigma_1) + (1-Z_i) \log f_2(Y_i; \mu_2, \Sigma_2) + Z_i \log p + (1-Z_i) \log(1-p) \}.$$

To get the conditional density, $f_{Z|Y}(\underline{Z}|\underline{y}; \hat{\underline{\theta}}^{(k)}) = \prod_{i=1}^{n} f_{Z|Y}(Z_i|y_i; \hat{\underline{\theta}}^{(k)})$

$$f_{Z|Y}(Z_i|y_i; \hat{\underline{\theta}}^{(k)}) = \frac{f_{YZ}(y_i, z_i; \hat{\underline{\theta}}^{(k)})}{f_Y(y_i; \hat{\theta}^{(k)})}$$

complete density contribution

observed density.

$$= \frac{\left[\hat{p}^{(k)} f_1(y_i; \hat{\mu}_1^{(k)}, \hat{\Sigma}_1^{(k)})\right]^{z_i} \left[(1-\hat{p}^{(k)}) f_2(y_i; \hat{\mu}_2^{(k)}, \hat{\Sigma}_2^{(k)})\right]^{1-z_i}}{\hat{p}^{(k)} f_1(y_i; \hat{\mu}_1^{(k)}, \hat{\Sigma}_1^{(k)}) + (1-\hat{p}^{(k)}) f_2(y_i; \hat{\mu}_2^{(k)}, \hat{\Sigma}_2^{(k)})}$$

Z_i can only take values 0 or 1 $\Rightarrow$ Bernoulli!

$$P(Z_i = 1 | Y_i = y_i, \hat{\theta}^{(k)}) = \frac{\hat{p}^{(k)} f_1(y_i; \hat{\mu}_1^{(k)}, \hat{\Sigma}_1^{(k)})}{\hat{p}^{(k)} f_1(y_i; \hat{\mu}_1^{(k)}, \hat{\Sigma}_1^{(k)}) + (1-\hat{p}^{(k)}) f_2(y_i; \hat{\mu}_2^{(k)}, \hat{\Sigma}_2^{(k)})} \overset{\text{define.}}{=} \hat{w}_i^{(k)} \quad \text{Then}$$

$$Z_i | Y_i = y, \hat{\theta}^{(k)} \sim \text{Bern}(\hat{w}_i^{(k)}).$$

$$\Rightarrow Q(\underline{\theta}, \hat{\theta}^{(k)}) = \sum_{i=1}^{n} E_{Z_i|Y_i|_{\theta=\hat{\theta}^{(k)}}} \left[\log f_{YZ}(y_i, z_i; \underline{\theta})\right] \text{ and } E_{Z_i|Y_i|_{\theta=\hat{\theta}^{(k)}}}\left[g(z_i)\right] = g(1)\hat{w}_i^{(k)} + g(0)(1-\hat{w}_i^{(k)}) \Rightarrow$$

$$Q(\underline{\theta}, \hat{\theta}^{(k)}) = \sum_{i=1}^{n} \Big\{ \hat{w}_i^{(k)} \left[\overset{1}{\cancel{z_i}} \log f_1(Y_i; \mu_1, \Sigma_1) + (1-\overset{0}{Z_i}) \log f_2(Y_i; \mu_2, \Sigma_2) + \overset{1}{Z_i} \log p + (1-\overset{0}{Z_i}) \log(1-p)\right]\Big|_{Z_i=1} +$$

$$(1-\hat{w}_i^{(k)}) \left[\underset{0}{Z_i} \log f_1(Y_i; \mu_1, \Sigma_1) + (1-\underset{1}{Z_i}) \log f_2(Y_i; \mu_2, \Sigma_2) + \underset{0}{Z_i} \log p + (1-\underset{1}{Z_i}) \log(1-p)\right]_{Z=0}$$

which yields the intuitive expression from before!

So "plugging in the weights" makes sense from an optization standpoint in this example.

In general can't always separate E + M in this way for Q.

The EM algorithm allows us to obtain $\hat{\boldsymbol{\theta}}_{\text{EM}}$, the parameter estimate which optimizes the algorithm.

Which, if the likelihood is "nice" can $= \hat{\theta}_{MLE}$.