There are many situations where finding the MLE can be difficult, especially when $\boldsymbol{\theta}$ is a vector. Many of these situations involve some form of missing data. For example, suppose we have observed heights of students at Middlebury, and we might like to estimate the true average height. Unfortunately, our data contain two subpopulations: males and female (sex assigned at birth). It is generally known that heights of females and males are Normally distributed, but with different means. Thus, it makes sense to model each observation as coming from the respective Normal and estimate the true mean of each distribution. However, we didn't ask respondents for their sex and so we don't know which observation corresponds to which Normal! Can we still estimate the two means?

When we have missing data problems, we can use the **EM algorithm** (expectation-maximization) to approximate the MLEs. This algorithm is iterative: we begin with some initial guess of the MLEs, then update our guesses repeatedly according to some rule. It is called EM because we alternate between evaluating an *Expectation* and then *Maximizing* this expected value. The EM algorithm is very commonly used today.

Set-up

Suppose the "complete" data are $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, where only $\mathbf{X}$ is observed. Further suppose that $\mathbf{X}$ and $\mathbf{Y}$ have distributions that are indexed by unknown parameter(s) $\boldsymbol{\theta}$. As usual, we would like to obtain MLEs for $\boldsymbol{\theta}$.

The **complete likelihood** or **full-data likelihood** for $\boldsymbol{\theta}$ is $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$, and the complete log-likelihood is then $\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$. It is so called because it would be the likelihood had we observed the full set of data.

For ease of notation assume:

- The observations are conditionally independent of each other given $\boldsymbol{\theta}$. So

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} f(x_i, y_i|\boldsymbol{\theta}),$$

  where $f(x_i, y_i|\theta)$ is the joint distribution of a complete data observation given $\boldsymbol{\theta}$

- The Y_i are discrete and take any of $m < \infty$ values (which is almost always the case)

So the complete log-likelihood is

$$l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \log f(x_i, y_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}|x_i, y_i) \equiv \sum_{i=1}^{n} l(\boldsymbol{\theta}|x_i, Y_i = y_i)$$

EM Algorithm

Start with some initial guess of $\boldsymbol{\theta}$, call it $\boldsymbol{\theta}^{old}$.

1. E-step: Treat the missing data $\mathbf{Y}$ as random variables and obtain $P(Y_i = j|x_i, \boldsymbol{\theta}^{old})$. This is the conditional distribution of the missing data given the observed data and assuming $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$.

   Then compute the (conditional) expected value of $l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ with respect to this conditional distribution. Call this conditional expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$.

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old}) &= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x},\theta^{old}}[l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})|\boldsymbol{x}, \boldsymbol{\theta}^{old}] \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} l(\boldsymbol{\theta}|x_i, Y_i = j)P(Y_i = j|x_i, \boldsymbol{\theta}^{old}) \qquad \textit{LoTUS and def. of discrete expectation} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} l(\boldsymbol{\theta}|x_i, Y_i = j)P(Y_i = j|x_i, \boldsymbol{\theta}^{old}) \end{aligned}$$

   The purpose of this E-step is that by taking the expectation, we are summing over all the unknown/unobserved Y_i, such that we are left with just a function of $\boldsymbol{\theta}$ and $\mathbf{x}$.

2. M-step: now we simply maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\theta}$ to obtain new estimates of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{new} = arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$$

   Note, this is done using calculus (finding partial derivatives and setting equal to 0), but hopefully, this maximization is much simpler than if we didn't use the complete data.

3. Set $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$.

Repeat these steps until convergence (i.e. until the estimates stop changing).

Example: Mixture of Normals

Suppose $X_1, \ldots, X_n$ are Normal random variables, where each X_i comes from one of m Normal distributions. In particular, $X_i|\mu_j \sim N(\mu_j, 1)$ with probability p_j for $j = 1, \ldots, m$. Assume that $\mu_1 < \mu_2 < \cdots < \mu_m$, $p_j \geq 0 \ \forall \ j$, and $\sum_{j=1}^{m} p_j = 1$.

The unknown parameters are $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_m, p_1, \ldots, p_m)$. Without knowing which particular Normal distribution X_i comes from, the distribution of $X_i|\boldsymbol{\theta}$ is

$$f(x_i|\boldsymbol{\theta}) = \sum_{j=1}^{m} p_j N(\mu_j, 1)$$

Trying to estimate $\boldsymbol{\theta}$ using our usual methods is very difficult, but is made much easier if we incorporate the missing data of group membership: assume there exists another set of unobserved variables $\mathbf{Y} = (Y_1, \ldots Y_n)$ where Y_i denotes which Normal X_i comes from. We have the following distributions:

$$\begin{aligned}
&P(Y_i = j|\boldsymbol{\theta}) = p_j, \quad j = 1, \ldots, m \\
&f(x_i|y_i = j, \boldsymbol{\theta}) = N(\mu_j, 1) \\
&f(x_i, y_i|\boldsymbol{\theta}) = f(x_i|y_i, \theta) P(Y_i = y_i|\boldsymbol{\theta}) = N(\mu_{y_i}, 1) p_{y_i}
\end{aligned}$$

So the complete likelihood and log-likelihood are:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} f(x_i, y_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(p_{y_i} \times \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x_i - \mu_{y_i})^2\} \right)$$

$$l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}|x_i, Y_i = y_i) = \sum_{i=1}^{n} \left[\log(p_{y_i}) - \frac{1}{2}(x_i - \mu_{y_i})^2 - \frac{1}{2}\log(2\pi) \right]$$

E-step

Let's find the conditional distribution of the missing data assuming $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$:

$$\begin{aligned} P(Y_i = j|x_i, \boldsymbol{\theta}^{old}) &= \frac{f(x_i, Y_i = j|\boldsymbol{\theta}^{old})}{f(x_i|\boldsymbol{\theta}^{old})} \qquad \textit{(Bayes)} \\ &= \frac{f(x_i|Y_i = j, \boldsymbol{\theta}^{old})P(Y_i = j|\boldsymbol{\theta}^{old})}{\sum_{k=1}^{m} f(x_i|Y_i = k, \boldsymbol{\theta}^{old})P(Y_i\ k|\boldsymbol{\theta}^{old})} \quad \textit{(dist. theory and LoTP)} \\ &= \frac{dnorm(x_i; \mu_j^{old}, 1)p_j^{old}}{\sum_{k=1}^{m} dnorm(x_i; \mu_k^{old}, 1)p_k^{old}} \\ &\equiv \tau_{ij}^{old} \qquad \textit{(note that this is just a number! Nothing unknown)} \end{aligned}$$

Now we obtain the conditional expectation:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} l(\boldsymbol{\theta}|x_i, Y_i = j)P(Y_i = j|x_i, \boldsymbol{\theta}^{old}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\log(p_j) - \frac{1}{2}(x_i - \mu_j)^2 - \frac{1}{2}\log(2\pi)\right)\tau_{ij}^{old} \end{aligned}$$

M-step

Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\theta}$, but with the constraint of $\sum_{j=1}^{m} p_j = 1$. (Lagrange multipliers coming back!) Note: no need to confirm maximum.

$$\boldsymbol{\theta}^{new} = arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\log(p_j) - \frac{1}{2}(x_i - \mu_j)^2 - \frac{1}{2}\log(2\pi)\right)\tau_{ij}^{old} + \lambda(\sum_{j=1}^{m} p_j - 1)$$

$$\begin{aligned} \frac{\partial}{\partial \mu_j} &= \sum_{i=1}^{n} \tau_{ij}^{old}(x_i - \mu_j) \overset{set}{=} 0 \Rightarrow \hat{\mu}_j = \frac{\sum_{i=1}^{n} x_i \tau_{ij}^{old}}{\sum_{i=1}^{n} \tau_{ij}^{old}}, \qquad j = 1, \ldots, m \\ \frac{\partial}{\partial \lambda} &= \sum_{j=1}^{m} p_j - 1 \overset{set}{=} 0 \Rightarrow \sum_{j=1}^{m} p_j = 1 \\ \frac{\partial}{\partial p_j} &= \sum_{i=1}^{n} \frac{\tau_{ij}^{old}}{\lambda} \overset{set}{=} 0 \Rightarrow p_j = \frac{-\sum_{i=1}^{n} \tau_{ij}^{old}}{\lambda} \end{aligned}$$

Now,

$$\sum_{j=1}^{m} \frac{-\sum_{i=1}^{n} \tau_{ij}^{old}}{\lambda} = 1 \Rightarrow \lambda = -\sum_{i=1}^{n}\sum_{j=1}^{m} \tau_{ij}^{old} = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(Y_i = j|x_i, \boldsymbol{\theta}^{old}) = -\sum_{i=1}^{n} 1 = -n$$

$$\hat{p}_j = \frac{\sum_{i=1}^{n} \tau_{ij}^{old}}{n}, \quad j = 1, \ldots, m$$

This yields, $\boldsymbol{\theta}^{new} = (\hat{\mu}_1, \ldots, \hat{\mu}_m, \hat{p}_1, \ldots, \hat{p}_m)$.

Proof of why EM works

Recall Jensen's inequality: If X is a random variable and $g()$ is a convex function, then $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$. An analogous inequality holds for concave functions: if $h()$ is a concave function, then $h(\mathbb{E}[X]) \geq \mathbb{E}[h(X)]$.

$$\begin{aligned}
l(\boldsymbol{\theta}|\mathbf{x}) = \log f(\mathbf{x}|\boldsymbol{\theta}) &= \log \int_{S_\mathbf{y}} f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\
&= \log \int_{S_\mathbf{y}} f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \cdot \frac{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})} d\mathbf{y} \\
&= \log \mathbb{E}\left[\frac{f(\mathbf{x}, \mathbf{Y}|\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})} \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right] \\
&\geq \mathbb{E}\left[\log \left(\frac{f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})} \right) \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right] \qquad \log() \text{ is concave } (*) \\
&= \mathbb{E}\left[\log f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right] - \mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old}) \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right] \\
&= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old}) - \mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old}) \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right] \equiv q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})
\end{aligned}$$

Any value of $\boldsymbol{\theta}$ that increases $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ beyond $q(\boldsymbol{\theta}^{old}|\boldsymbol{\theta}^{old})$ must also increase $l(\boldsymbol{\theta}|\mathbf{x})$ beyond $l(\boldsymbol{\theta}^{old}|\mathbf{x})$. Since $\mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old}) \Big| \mathbf{x}, \boldsymbol{\theta}^{old} \right]$ is a constant, maximizing the $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ is equivalent to maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ over $\boldsymbol{\theta}$. If $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$, then

$$\frac{f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})} = \frac{f(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}^{old})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}^{old})},$$

which is a constant (no unknowns). Since the log of the expectation of a constant is equal to the expectation of the log of that constant, we have equality instead of inequality in $(*)$ if $\boldsymbol{\theta} = \boldsymbol{\theta}^{old}$ (convergence).