An Introduction to the EM Algorithm and its Applications
The EM algorithm is used to compute MLE estimates when data are missing.

In the case of mixture models, missing data can take at least three forms:

- *Missing labels*
- *Unknown number of components*
- *Missing data in the data vectors*
Goal of the EM Algorithm

Find the MLE when direct maximization of the likelihood function is difficult.

One difficult maximization problem \(\rightarrow\) Sequence of simpler maximization problems
Some Applications of the EM Algorithm

- Estimation in the presence of **missing data**
- Estimation of the parameters of a **mixture distribution**
Some Notation

\[ Y = \text{Incomplete data} \]
\[ X = \text{Augmented data} \]
\[ (Y, X) = \text{Complete data} \]
\[ \theta = \text{Parameters to be estimated} \]
The EM algorithm has two main steps:

(i) The **Expectation Step** (E-step)

(ii) The **Maximization Step** (M-step)
Expectation Step

\[ l(\theta|y) = \log(f(y|\theta)) \]  
Incomplete log-likelihood

\[ l(\theta|y, x) = \log(f(y, x|\theta)) \]  
Complete log-likelihood

\[ \tilde{l}(\theta|y, \theta^{(k)}) = E_{x|y, \theta^{(k)}} \{ l(\theta|y, X) \} = \]
Expected log-likelihood

\[ \int \cdots \int \log f(y, x|\theta) h(x|y, \theta^{(k)}) \, dx \]
The Conditional Density

\[ h \left( x \mid y, \theta^{(k)} \right) = \frac{f \left( y, x \mid \theta^{(k)} \right)}{f \left( y \mid \theta^{(k)} \right)} \]

Note that \( h \) is a fully specified density.
Maximization Step

$$\theta^{(k+1)} = \arg \max_\theta \tilde{l} (\theta | y, \theta^{(k)})$$

Each iteration (EM-step) increases the value of the incomplete-data log-likelihood:

$$l (\theta^{(k+1)} | y) \geq l (\theta^{(k)} | y)$$
Overview of the EM algorithm

**Step 1:** Obtain the *complete-data* log-likelihood function.

**Step 2:** Take expectation using the *conditional distribution* of the augmented data, given the incomplete data and the current values of the parameters. This is called the **E-step**.

**Step 3:** Maximize the resulting expected log-likelihood function. This is called the **M-step**.

**Step 4:** Repeat steps 2 and 3 until *convergence*. 
A Simple Example: Mixture Distribution

Incomplete-Data: \( Y_1, Y_2, \ldots, Y_n \) iid with common density

\[
f (y) = (1 - p) f_0 (y) + pf_1 (y)
\]

Incomplete-Data Log-Likelihood:

\[
l (p|y) = \sum_{i=1}^{n} \log [(1 - p) f_0 (y_i) + pf_1 (y_i)]
\]
Complete-Data Log-Likelihood

Augmented Data: $X_1, X_2, \ldots, X_n$, iid $Bin(1, p)$

Complete Data: $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$, independent, with common density

$$f(y, x) = [(1 - p) f_0(y)]^{1-x} [pf_1(y)]^x$$

Complete-Data Log-Likelihood:

$$l(p|y, x) = \sum_{i=1}^{n} (1 - x_i) \log [(1 - p) f_0(y_i)] + x_i \log [pf_1(y_i)]$$
A key observation (for the E-Step)

\[ E \left[ X_i | y_i, p^{(k)} \right] = P \left[ X_i = 1 | y_i, p^{(k)} \right] \]

\[ = \frac{f \left( y_i, 1 | p^{(k)} \right)}{f \left( y_i, 0 | p^{(k)} \right) + f \left( y_i, 1 | p^{(k)} \right)} \]

\[ = \frac{f_1 \left( y_i \right) p^{(k)}}{f_0 \left( y_i \right) \left( 1 - p^{(k)} \right) + f_1 \left( y_i \right) p^{(k)}} \]

\[ = \tilde{p}_i \]
E-Step

\[ l (p | y, X) = \log (1 - p) \sum_{i=1}^{n} (1 - X_i) + \sum_{i=1}^{n} (1 - X_i) \log (f_0 (y_i)) \]

\[ + \log (p) \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} X_i \log (f_1 (y_i)) \]

\[ \tilde{l} (p | y) = \log (1 - p) \sum_{i=1}^{n} (1 - \tilde{p}_i) + \sum_{i=1}^{n} (1 - \tilde{p}_i) \log (f_0 (y_i)) \]

\[ + \log (p) \sum_{i=1}^{n} \tilde{p}_i + \sum_{i=1}^{n} \tilde{p}_i \log (f_1 (y_i)) \]
The M-Step

\[
\frac{\partial}{\partial p} \tilde{l}(p|y) = \sum_{i=1}^{n} \frac{(1 - \tilde{p}_i)}{1 - p} + \sum_{i=1}^{n} \frac{\tilde{p}_i}{p} = 0
\]

\[
p^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p}_i}{n} = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1(y_i)p^{(k)}}{f_0(y_i)(1-p^{(k)}) + f_1(y_i)p^{(k)}}
\]
Extensions

➤ Mixture distribution has more than two components

➤ Distributions include unknown parameters

➤ Data are vector valued
More Than Two Components

In this case the complete-data log-likelihood

\[
l (p | y_1, \ldots, y_n, x_1, \ldots, x_n) = \log \left\{ \prod_{i=1}^{n} \prod_{j=1}^{m} [p_j f_j(y_i)]^{x_{ij}} \right\}
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} [\log (p_j) + \log (f_j(y_i))]
\]

where \( x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \) is Multinomial \((1, p_1, p_2, \ldots, p_m)\)
The conditional probability that the \(i^{th}\) observation came from the \(j^{th}\) mixture component, given the observation value and the current parameter estimates, are

\[
\tilde{p}_{ij} = E \left( X_{ij} | y_i, p^{(k)} \right)

= P \left( X_{ij} = 1 | y_i, p^{(k)} \right) = \frac{p_j^{(k)} f_j(y_i)}{\sum_{\alpha=1}^{m} p_{\alpha}^{(k)} f_{\alpha}(y_i)}
\]

The updated probabilities for the \(j^{th}\) mixture component is

\[
p_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \tilde{p}_{ij}, \quad j = 1, \ldots, m
\]
Distributions Include Unknown Parameters

Densities include unknown parameters. For example:

\[ f_0 (y) = N (\mu_0, \sigma_0^2) \]
\[ f_1 (y) = N (\mu_1, \sigma_1^2) \]

In this case

\[ \theta = (p, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2) \]
“Independent Pieces”

\[ \tilde{l}(\theta|y) = \sum_{i=1}^{n} (1 - \tilde{p}_i) \log(f_0(y_i)) \]

\[ + \sum_{i=1}^{n} \tilde{p}_i \log(f_1(y_i)) \]

\[ + \log(1 - p) \sum_{i=1}^{n} (1 - \tilde{p}_i) + \log(p) \sum_{i=1}^{n} \tilde{p}_i \]

\[ = \tilde{l}_0(\mu_0, \sigma_0^2|y) + \tilde{l}_1(\mu_1, \sigma_1^2|y) + \tilde{l}_2(p|y) \]
The Iteration Steps

\[
\hat{\mu}_0^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - \tilde{p}_i) y_i}{\sum_{i=1}^{n} (1 - \tilde{p}_i)}
\]

\[
\hat{\mu}_1^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p}_i y_i}{\sum_{i=1}^{n} \tilde{p}_i}
\]

\[
(\hat{\sigma}^2_0)^{(k+1)} = \frac{\sum_{i=1}^{n} (1 - \tilde{p}_i) \left( y_i - \hat{\mu}_0^{(k+1)} \right)^2}{\sum_{i=1}^{n} (1 - \tilde{p}_i)}
\]

\[
(\hat{\sigma}^2_1)^{(k+1)} = \frac{\sum_{i=1}^{n} \tilde{p}_i \left( y_i - \hat{\mu}_1^{(k+1)} \right)^2}{\sum_{i=1}^{n} \tilde{p}_i}
\]
Example 2

Estimating the overall infection rate.

We observe $Y_1, Y_2, \ldots, Y_n$ and $Z_1, Z_2, \ldots, Z_n$ (mutually independent)

$Y_i = \text{Number of infected people in the } i^{th} \text{ region}$ $\text{Poisson}(\beta \tau_i)$

$Z_i = \text{Measure of population density in the } i^{th} \text{ region}$ $\text{Poisson}(\tau_i)$

$\tau_i = \text{Factor influencing population density in the } i^{th} \text{ region}$ \text{Unknown}

$\beta = \text{Overall infection rate}$ \text{constants}
The complete-data likelihood is:

\[
f (y, z | \beta, \tau_1, \ldots, \tau_n) = \prod_{i=1}^{n} e^{-\beta \tau_i} \frac{(\beta \tau_i)^{y_i}}{y_i!} \times e^{-\tau_i} \frac{(\tau_i)^{z_i}}{z_i!}
\]

The MLE estimates are:

\[
\hat{\beta} = \frac{\bar{y}}{\bar{z}}
\]

\[
\hat{\tau}_i = \frac{z_i + y_i}{\hat{\beta} + 1}, \quad i = 1, \ldots, n
\]
Incomplete-Data Likelihood

Suppose now that $z_1$ is missing

The incomplete-data likelihood is:

$$f(y, z|\beta, \tau_1, \ldots, \tau_n) = \prod_{i=1}^{n} \frac{e^{-\beta \tau_i} (\beta \tau_i)^{y_i}}{y_i!} \prod_{i=2}^{n} \frac{e^{-\tau_i} (\tau_i)^{z_i}}{z_i!}$$
Incomplete-Data MLE Equations

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{\tau}_i}
\]

\[
y_1 = \hat{\tau}_1 \hat{\beta}
\]

\[
z_i + y_i = \hat{\tau}_i \left( \hat{\beta} + 1 \right) \quad i = 2, \ldots, n
\]

This system of equations is not easy to solve
Notations

Incomplete Data: \( y = (y_1, \ldots, y_n), \quad z_{(-1)} = (z_2, \ldots, z_n) \)

Augmented Data: \( z_1 \)

Complete Data: \( y = (y_1, \ldots, y_n), \quad z = (z_1, \ldots, z_n) \)
Complete data log-likelihood

\[ l (\beta, \tau_1, \ldots, \tau_n | y, z) = \log [f (y, z | \beta, \tau_1, \ldots, \tau_n)] \]

\[ = C - (1 + \beta) \sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} y_i \log (\beta \tau_i) + \sum_{i=1}^{n} z_i \log (\tau_i) \]

\[ = C - \sum_{i=1}^{n} [(1 + \beta) \tau_i + y_i \log (\beta \tau_i)] + \sum_{i=2}^{n} z_i \log (\tau_i) + z_1 \log (\tau_1) \]
The E-Step

\[ \tilde{l} (\beta, \tau | y, z_{(-1)}) = E_{Z_1 | y, z_{(-1)}, \hat{\tau}^{(k)}, \hat{\beta}^{(k)}} \{ l (\beta, \tau_1, \ldots, \tau_n | y, z) \} \]

\[ = C_1 - \sum_{i=1}^{n} [(1 + \beta) \tau_i + y_i \log (\beta \tau_i)] + \sum_{i=2}^{n} z_i \log (\tau_i) + \hat{\tau}_1^{(k)} \log (\tau_1) \]

\[ = C_2 + l (\beta, \tau_1, \ldots, \tau_n | y, \hat{\tau}_1^{(k)}, z_2, \ldots, z_n) \]
The M-Step

\[
\hat{\beta}^{(k+1)} = \frac{\sum_{i=1}^{n} y_i}{\hat{\tau}_1^{(k)} + \sum_{i=2}^{n} z_i}
\]

\[
\hat{\tau}_1^{(k+1)} = \frac{\hat{\tau}_1^{(k)} + y_1}{\hat{\beta}^{(k+1)} + 1}
\]

\[
\hat{\tau}_i^{(k+1)} = \frac{z_i + y_i}{\hat{\beta}^{(k+1)} + 1}, \quad i = 2, \ldots, n
\]