Motivation

Old idea:
- Replace missing values by estimated values
- Estimate parameters using full-data methods
- Re-estimate missing values based on new parameter values
- Repeat previous two steps until convergence

Example: Censored exponentially distributed data
Suppose \( T_1, \ldots, T_n \overset{\text{iid}}{\sim} \text{Exp}(\theta) \) with common censoring time \( C \), that is,
- \( T_i \) is observed \((R_i = 1)\) if \( T_i \leq C \) and
- \( T_i \) is missing \((R_i = 0)\) if \( T_i > C \).
Suppose that only \( T_1, \ldots, T_m \) are observed.
Complete-data log-likelihood
\[
    l_n(\theta|T) = n \log \theta - \theta \sum_{i=1}^{n} T_i
\]
Since \( T_{m+1}, \ldots, T_n \) are unobserved, we replace them by their expectation
\[
    T_i^* = \mathbb{E}(T_i|T_i > C, \theta^*) = C + \frac{1}{\theta^*}
\]
where \( \theta^* \) is the current estimate of \( \theta \). Thus
\[
    l_n(\theta|T_1, \ldots, T_m, T_{m+1}^*, \ldots, T_n^*) = n \log \theta - \theta \sum_{i=1}^{n} Y_i - (n - m) \frac{\theta}{\theta^*},
\]
where \( Y_i = \min(T_i, C) \). Differentiating with respect to \( \theta \) and equating to zero yields
\[
    \frac{n}{\theta} = \sum_{i=1}^{n} Y_i + \frac{n - m}{\theta^*}
\]
At convergence, \( \theta = \theta^* \) and thus
\[
    \frac{n}{\hat{\theta}} - \frac{n - m}{\hat{\theta}} = \frac{m}{\hat{\theta}} = \sum_{i=1}^{n} Y_i \Rightarrow \hat{\theta} = m/\sum_{i=1}^{n} Y_i.
\]
Motivation

**Problem:** Method does not always lead to ML estimates

**Example:** Regression

Suppose that

\[ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \]

with \( \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \) and that only the first \( m \) values have been observed.

The complete-data log-likelihood is

\[ l_n(\beta_0, \beta_1, \sigma^2 | Y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \]

After imputation of predicted values

\[ Y_i^* = \begin{cases} Y_i & \text{if } i = 1, \ldots, m \\ \beta_0^* + \beta_1^* x_i & \text{otherwise} \end{cases} \]

maximization of \( l_n(\beta_0, \beta_1, \sigma^2 | Y_{\text{obs}}, Y_{\text{mis}}^*) \) yields

\[ \beta_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})Y_i^*}{\sum_{i=1}^{m} (x_i - \bar{x})^2} \]

\[ \beta_0 = \bar{Y}^* - \beta_1 \bar{x} \]

At convergence, \( \beta_1 = \beta_1^* \) and \( \beta_0 = \beta_0^* \), which leads to

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})Y_i}{\sum_{i=1}^{m} (x_i - \bar{x})^2} \]

\[ \hat{\beta}_0 = \frac{1}{m} \sum_{i=1}^{m} Y_i - \hat{\beta}_1 \frac{1}{m} \sum_{i=1}^{m} x_i. \]

However, maximization with respect to \( \sigma^2 \) yields

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \]

that is, the variance is underestimated.
Motivation

Example: Regression (contd)

Derivation of expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$:

The LS estimates based on $Y^* = (Y_{\text{obs}}, Y^*_{\text{mis}})$ satisfy

$$
\beta_0 = \bar{Y}^* - \beta_1 \bar{x}^*,
$$

$$
\sum_{i=1}^{n} (x_i - \bar{x}^*)^2 \beta_1 = \sum_{i=1}^{n} (Y_i - \bar{Y}^*)(x_i - \bar{x}^*)
$$

where $\bar{Y}^* = \sum_{i=1}^{n} Y^*_i$ and $\bar{x}^* = \sum_{i=1}^{n} x_i$.

From the first equation, we get

$$
n \beta_0 = n \bar{Y}^* - \beta_1 n \bar{x}^*
$$

$$
= m \bar{Y} + \sum_{i=m+1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 m \bar{x} \beta_1 \sum_{i=m+1}^{n} x_i
$$

$$
= m \bar{Y} - \beta_1 m \bar{x} + (n - m) \beta_0
$$

which leads to

$$
\beta_0 = \bar{Y} - \beta_1 \bar{x}
$$

with $\bar{Y} = \sum_{i=1}^{m} Y_i$ and $\bar{x} = \sum_{i=1}^{m} x_i$. From the second equation, we obtain

$$
\beta_1 \sum_{i=1}^{n} (x_i - \bar{x}^*)^2 = \sum_{i=1}^{m} (Y_i - \bar{Y}^*)(x_i - \bar{x}^*) + \sum_{i=m+1}^{n} (\beta_0 + \beta_1 x_i - \bar{Y}^*)(x_i - \bar{x}^*)
$$

$$
= \bar{Y}^* - \beta_1 \bar{x}^* + \beta_1 x_i - \bar{Y}^*
$$

$$
= \beta_1 (x_i - \bar{x}^*)
$$

$$
= \sum_{i=1}^{m} (Y_i - \bar{Y}^*)(x_i - \bar{x}^*) + \sum_{i=m+1}^{n} (x_i - \bar{x}^*)^2 \cdot \beta_1
$$

which leads to

$$
\beta_1 \sum_{i=1}^{m} (x_i - \bar{x}^*)^2 = \sum_{i=1}^{m} (Y_i - \bar{Y}^*)(x_i - \bar{x}^*).
$$
Motivation

Example: Regression (contd)

The left side can be written as

\[ \beta_1 \sum_{i=1}^{m}(x_i - \bar{x} + \bar{x} - \bar{x}^*)^2 = \beta_1 \sum_{i=1}^{m}(x_i - \bar{x})^2 + \beta_1 m (\bar{x} - \bar{x}^*)^2. \]

For the right side, we obtain

\[
\begin{align*}
\sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x} + \bar{x} - \bar{x}^*) &= \sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x}) + (\bar{x} - \bar{x}^*) \sum_{i=1}^{m}(Y_i - \bar{Y}^*) \\
&= \sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x}) + (\bar{x} - \bar{x}^*) m (\bar{Y} - \bar{Y}^*) \\
&= \beta_0 + \beta_1 \bar{x} - \beta_0 - \beta_1 \bar{x}^* \\
&= \beta_1 (\bar{x} - \bar{x}^*) \\
&= \sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x}) + m \beta_1 (\bar{x} - \bar{x}^*)^2
\end{align*}
\]

Altogether, we obtain

\[ \beta_1 \sum_{i=1}^{m}(x_i - \bar{x})^2 = \sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x}) \]

since the terms \( m \beta_1 (\bar{x} - \bar{x}^*)^2 \) cancel on both sides. Thus

\[ \beta_1 = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y}^*)(x_i - \bar{x})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}. \]

is the LS estimator for \( \beta_1 \) based on the first \( m \) observations.
EM Algorithm

Intuition
- Imputation of expected values of missing observations leads to the log-likelihood
  \[ l_n(\theta|Y_{\text{obs}}, \mathbb{E}(Y_{\text{mis}}|Y_{\text{obs}}, \theta^*)) \]
  which is then maximized with respect to \( \theta \).
- Instead estimate directly the complete-data log-likelihood by its conditional expectation given the observed data
  \[ \mathbb{E}(l_n(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}, \theta^*) \]
- Maximize the expected log-likelihood with respect to \( \theta \).

EM algorithm
- **Expectation (E) Step**: Calculate the expectation
  \[
  Q(\theta|\theta^{(k)}) = \mathbb{E}(l_n(\theta|Y)|Y_{\text{obs}}, \theta^{(k)})
  = \int l_n(\theta|Y_{\text{obs}}, y_{\text{mis}}) f_{Y_{\text{mis}}}(y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)}) dy_{\text{mis}}.
  \]
- **Maximization (M) Step**: Maximize \( Q(\theta|\theta^{(k)}) \) with respect to \( \theta \)
  \[ Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta|\theta^{(k)}) \quad \text{for all } \theta \in \Theta. \]
**EM Algorithm**

**Example:** Regression

Suppose that

\[ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \]

with \( \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \) and that only the first \( m \) values have been observed.

The complete-data log-likelihood is

\[
    l_n(\beta_0, \beta_1, \sigma^2 | Y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu_i)^2
    \]

\[
    = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i^2 - 2Y_i \mu_i + \mu_i^2)
    \]

with \( \mu_i = \beta_0 + \beta_1 x_i \).

○ **E-step:**

\[
    \mathbb{E}(Y_i | \theta^*) = \mu_i^*
    \]
\[
    \mathbb{E}(Y_i^2 | \theta^*) = \sigma^*^2 + (\mu_i^*)^2
    \]

Thus

\[
    Q(\theta | \theta^*) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i^* - \mu_i)^2 - \frac{n - m}{2\sigma^2} \sigma^*^2
    \]

○ **M-step:** As before we obtain at convergence

\[
    \hat{\beta}_1 = \frac{\sum_{i=1}^{m} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{m} (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \frac{1}{m} \sum_{i=1}^{m} Y_i - \beta_1 \frac{1}{m} \sum_{i=1}^{m} x_i.
    \]

For \( \sigma^2 \) we get at convergence \( (\sigma^2 = \sigma^*^2) \)

\[
    \frac{\partial Q(\theta | \theta^*)}{\partial \theta} = -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{m} (Y_i - \mu_i)^2 - \frac{n - m}{2\sigma^2}
    \]

and thus

\[
    \hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^{m} (Y_i - \mu_i)^2.
    \]
**Example**

**Example:** $t$ distribution
Suppose that $Y_1, \ldots, Y_n$ are independently sampled from the density

$$f_{Y_i}(y|\mu) = \frac{1}{\sqrt{\pi \Gamma(\frac{1}{2})}} (1 + (y - \mu)^2)^{-1}$$

($t$ distribution with one degree of freedom and noncentrality parameter $\mu$).

Suppose that $X_i \overset{iid}{\sim} \chi^2_1$ such that

$$Y_i|X_i \sim \mathcal{N}(\mu, X_i^{-1}).$$

The complete-data log-likelihood is

$$l_n(\mu|Y, X) = \frac{1}{2} \sum_{i=1}^n \log X_i - \frac{1}{2} \sum_{i=1}^n X_i(Y_i - \mu)^2 + \sum_{i=1}^n \log f_{X_i}(X_i).$$

- **E-step:** Note that
  $$X_i|Y_i \sim \Gamma(1, \frac{1}{2} [1 + (Y_i - \mu)^2])$$

  and thus
  $$x_i^{(k)} = \mathbb{E}(X_i|Y_i, \mu^{(k)}) = \frac{2}{1 + (Y_i - \mu^{(k)})^2}.$$

  This leads to
  $$Q(\mu|\mu^{(k)}) = -\frac{1}{2} \sum_{i=1}^n x_i^{(k)} (Y_i - \mu)^2 + \text{constant}$$

- **M-step:** Maximization with respect to $\mu$ leadsto
  $$\mu^{(k+1)} = \frac{\sum_{i=1}^n x_i^{(k)} Y_i}{\sum_{i=1}^n x_i^{(k)}}.$$
Example

Derivation of conditional distribution of $X_i$ given $Y_i$:

Since

$$Y_i \sim t_1(\mu)$$
$$X_i \sim \chi^2_1 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$
$$Y_i | X_i \sim N\left(\mu, \frac{1}{x_i}\right)$$

we have

$$f_{Y_i}(y) = \frac{1}{\sqrt{\pi} \Gamma\left(\frac{1}{2}\right)} \left(1 + (y - \mu)^2\right)^{-1}$$
$$f_{X_i}(x) = \frac{1}{\sqrt{2}} \Gamma\left(\frac{1}{2}\right) x^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x\right)$$
$$f_{Y_i | X_i}(y | x) = \frac{1}{\sqrt{2\pi / x}} \exp\left(-\frac{1}{2} x (y - \mu)^2\right)$$

Thus the density of the conditional distribution of $X_i$ given $Y_i$ is

$$f_{X_i | Y_i}(x | y) = \frac{f_{Y_i | X_i}(y | x) f_{X_i}(x)}{f_{Y_i}(y)}$$
$$= \frac{x^{\frac{1}{2}} x^{-\frac{1}{2}} \exp\left[-\frac{1}{2} x (1 + (y - \mu)^2)\right]}{\sqrt{\pi} \Gamma\left(\frac{1}{2}\right) (1 + (y - \mu)^2)^{-1}}$$
$$= \frac{1}{2} \left(1 + (y - \mu)^2\right) \exp\left[-\frac{1}{2} x (1 + (y - \mu)^2)\right],$$

that is,

$$X_i | Y_i \sim \Gamma\left(1, \frac{1}{2} \left(1 + (Y_i - \mu)^2\right)\right)$$

with

$$\mathbb{E}(X_i | Y_i) = \frac{2}{1 + (Y_i - \mu)^2}.$$
Example

Implementation in R

Now suppose that
\[ Y = (-1.318, 0.613, -6.004, -22.687)^T. \]

\[
estep<-function(Y,p) {
  2/(1+(Y-p)^2)
}
\]

\[
mstep<-function(Y,X) {
  sum(X*Y)/sum(X)
}
\]

\[
em<-function(Y,p,n) {
  for (i in (1:n)) {
    X<-estep(Y,p)
    p<-mstep(Y,X)
  }
  p
}
\]

\[
Y<-c(-1.3181159, 0.6131288, -6.0042818, -22.6870591)
\]

\[
p<-0
\]
\[
X<-estep(Y,p)
\]
\[
p<-mstep(Y,X)
\]

Iterate until convergence
Convergence Properties

Density of the complete data \( Y \)

\[ f_Y(y|\theta) = f_{Y_{\text{obs}},Y_{\text{mis}}}(y_{\text{obs}}, y_{\text{mis}}|\theta) = f_{Y_{\text{obs}}}(y_{\text{obs}}|\theta)f_{Y_{\text{mis}}|Y_{\text{obs}}}(y_{\text{mis}}|y_{\text{obs}}, \theta) \]

Thus the corresponding log-likelihood can be written as

\[ l_n(\theta|Y) = l_n(\theta|Y_{\text{obs}}) + \log f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta) \]

or equivalently

\[ l_n(\theta|Y_{\text{obs}}) = l_n(\theta|Y) - \log f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta). \]

Note: The left side does not depend on \( Y_{\text{mis}} \). Thus

\[ l_n(\theta|Y_{\text{obs}}) = \mathbb{E}_Y \left( l_n(\theta|Y_{\text{obs}}) \big| Y_{\text{obs}}, \theta^{(k)} \right) = Q(\theta|\theta^{(k)}) - H(\theta|\theta^{(k)}) \]

where

\[ Q(\theta|\theta^{(k)}) = \mathbb{E}_Y \left( l_n(\theta|Y_{\text{obs}}) \big| Y_{\text{obs}}, \theta^{(k)} \right) \]

and

\[ H(\theta|\theta^{(k)}) = \mathbb{E}_Y \left( \log f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta) \big| Y_{\text{obs}}, \theta^{(k)} \right) \]

By Jensen’s inequality

\[ H(\theta|\theta^{(k)}) - H(\theta^{(k)}|\theta^{(k)}) = \mathbb{E}_Y \left( \log \frac{f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta)}{f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})} \big| Y_{\text{obs}}, \theta^{(k)} \right) \]

\[ \leq \log \left( \mathbb{E}_Y \left[ \frac{f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta)}{f_{Y_{\text{mis}}|Y_{\text{obs}}}(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})} \big| Y_{\text{obs}}, \theta^{(k)} \right] \right) \]

\[ = \log \left[ \int f_{Y_{\text{mis}}|Y_{\text{obs}}}(y_{\text{mis}}|Y_{\text{obs}}, \theta) \, dy_{\text{mis}} \right] = \log(1) = 0 \]

Hence

\[ l_n(\theta^{(k+1)}|Y_{\text{obs}}) - l_n(\theta^{(k)}|Y_{\text{obs}}) = [Q(\theta^{(k+1)}|\theta^{(k)}) - Q(\theta^{(k)}|\theta^{(k)})] + [H(\theta^{(k)}|\theta^{(k)}) - H(\theta^{(k+1)}|\theta^{(k)})] \geq 0 \]
Jensen’s Inequality

Suppose that
  ◦ $X$ is a real-valued random variable that takes in $(a, b)$;
  ◦ $\psi$ is a convex function on $(a, b)$;
  ◦ $E(X)$ and $E(\psi(X))$ both exist.

Then
$$\psi(E(X)) \leq E(\psi(X)).$$

Geometric proof:

- The line $\alpha(x)$ is tangent to the convex function $\psi(x)$ at $x_0 = E(X)$.
- By convexity $\alpha(x) \leq \psi(x)$ for all $x$
- Thus
  $$E(\alpha(X)) \leq E(\psi(X))$$
- Linearity of $\alpha(x)$ implies
  $$E(\alpha(X)) = \alpha(E(X)) = \psi(E(X))$$