EM Algorithm

1.1 Overview

Suppose that the likelihood of the observed data Y_{obs} is difficult to maximize. However, given additional unobserved data Y_{miss}, the likelihood maximization becomes tractable. Let $\ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})$ denote the complete data log likelihood, and $\ln f(Y_{\text{obs}}|\boldsymbol{\theta})$ the observed data log likelihood. The Expectation-Maximization algorithm proceeds as follows:

1. **E-step**: Given a current estimate of the parameter $\boldsymbol{\theta}^{(r)}$, calculate the expectation of the complete data log likelihood given the observed data:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \int \ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta}) f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)}) dY_{\text{miss}}.$$

2. **M-step**: Maximize the current objective function to update the estimate of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(r+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}\right)$$

Remark 1.1.1. Provided differentiation in $\boldsymbol{\theta}$ commutes with integration over $f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta})$, the EM-algorithm may be implemented as follows:

- Derive the complete data score equations $\ell(\boldsymbol{\theta})$:

$$\mathcal{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta}).$$

- Take the expectation of the complete data score equations given Y_{obs} and $\boldsymbol{\theta}^{(r)}$:

$$\mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{E}\left\{\mathcal{U}(\boldsymbol{\theta})\middle|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)}\right\}.$$

- Obtain $\boldsymbol{\theta}^{(r+1)}$ by solving:

$$\mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) \stackrel{\text{Set}}{=} \mathbf{0}.$$

♦

Remark 1.1.2. The overall EM algorithm converges linearly, regardless of the maximization procedure employed in the M-step. Thus, stable linear optimization methods, such as coordinate ascent, are perhaps preferable to faster yet less stable optimization methods, such as Newton-Raphson. ♦

1.2 Information Inequality

Definition 1.2.1. The **cross entropy** between densities f and g is:

$$\mathcal{H}(f||g) = -\int \ln\big\{g(x)\big\}f(x)dx = -\mathbb{E}_f \ln\big\{g(x)\big\}.$$

■

Proposition 1.2.1 (**Gibbs' Inequality**). For densities f and g:

$$\mathcal{H}(f||g) - \mathcal{H}(f||f) \geq 0. \tag{1.2.1}$$

♦

Proof.

$$\begin{aligned}\mathcal{H}(f||g) - \mathcal{H}(f||f) &= -\int \ln\big\{g(x)\big\}f(x)dx + \int \ln\big\{f(x)\big\}f(x)dx \\ &= -\int \ln\left\{\frac{g(x)}{f(x)}\right\}f(x)dx.\end{aligned}$$

For $x \in (0,\infty)$, $\ln(x) \leq x - 1$, or $-\ln(x) \geq -(x-1)$, therefore:

$$\begin{aligned}-\int \ln\left\{\frac{g(x)}{f(x)}\right\}f(x)dx &\geq -\int \left\{\frac{g(x)}{f(x)} - 1\right\}f(x)dx \\ &\geq -\int g(x)dx + \int f(x)dx = -1 + 1 = 0.\end{aligned}$$

■

Definition 1.2.2. The **Kullback-Leibler divergence** between f and g is:

$$\mathbb{KL}(f||g) \equiv \mathcal{H}(f||g) - \mathcal{H}(f||f) = -\int \ln\left\{\frac{g(x)}{f(x)}\right\}f(x)dx. \tag{1.2.2}$$

■

1.3 Verification of the EM Algorithm

Definition 1.3.1. The **EM objective** is defined as:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \int \ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})dY_{\text{miss}}. \tag{1.3.3}$$

■

Proposition 1.3.2. The observed data log likelihood decomposes as the EM objective plus the cross entropy $\mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)})$ between $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}^{(r)})$ and $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})$. That is:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)}), \tag{1.3.4}$$

where:

$$\mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)}) = -\int \ln\left\{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})\right\} f(Y_{\text{miss}}|Y_{\text{obs}},\boldsymbol{\theta}^{(r)})dY_{\text{miss}}.$$

♦

Proof. The conditional density of the missing data given the observed is:

$$f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}) = \frac{f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta})}{f(Y_{\text{obs}}|\boldsymbol{\theta})}$$

Upon taking the logarithm:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = \ln f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta}) - \ln f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}).$$

Taking the expectation with respect to the density $f(Y_{\text{miss}}|Y_{\text{obs}},\boldsymbol{\theta}^{(r)})$:

$$\begin{aligned}\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = &\int \ln\left\{f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta})\right\} f(Y_{\text{miss}}|Y_{\text{obs}},\boldsymbol{\theta}^{(r)})dY_{\text{miss}} \\ &- \int \ln\left\{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})\right\} f(Y_{\text{miss}}|Y_{\text{obs}},\boldsymbol{\theta}^{(r)})dY_{\text{miss}} \\ =&Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)}).\end{aligned}$$

■

Proposition 1.3.3 (Increment Property). Increasing the EM objective causes an increase at least as great in the observed data log likelihood:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) - \ln f(Y_{\text{obs}}|\boldsymbol{\theta}^{(r)}) \ge Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r)}).$$

♦

Proof. Substituting $\boldsymbol{\theta}^{(r)}$ for $\boldsymbol{\theta}$ in (1.3.4):

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}^{(r)}) = Q(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}(\boldsymbol{\theta}^{(r)}||\boldsymbol{\theta}^{(r)}). \tag{1.3.5}$$

Subtracting (1.3.5) from (1.3.4):

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) - \ln f(Y_{\text{obs}}|\boldsymbol{\theta}^{(r)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)}) - \mathcal{H}(\boldsymbol{\theta}^{(r)}||\boldsymbol{\theta}^{(r)}).$$

From Gibbs' inequality (1.2.1), $\mathcal{H}(\boldsymbol{\theta}||\boldsymbol{\theta}^{(r)}) - \mathcal{H}(\boldsymbol{\theta}^{(r)}||\boldsymbol{\theta}^{(r)}) \ge 0$, therefore:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) - \ln f(Y_{\text{obs}}|\boldsymbol{\theta}^{(r)}) \ge Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) - Q(\boldsymbol{\theta}^{(r)}|\boldsymbol{\theta}^{(r)}).$$

■

1.4 Connection to Variational Inference

Proposition 1.4.4. The observed data log likelihood is expressible as:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = \mathcal{L}\big\{\boldsymbol{\theta}, g(Y_{\text{miss}})\big\} + \mathbb{KL}\big\{g(Y_{\text{miss}})||f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})\big\},$$

where $g(Y_{\text{miss}})$ is any density over Y_{miss},

$$\mathcal{L}\big\{\boldsymbol{\theta}, g(Y_{\text{miss}})\big\} = \int \ln\left\{\frac{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}}.$$

is the **evidence lower bound** (ELBO), and:

$$\mathbb{KL}\big\{g(Y_{\text{miss}})||f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})\big\} = -\int \ln\left\{\frac{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}}$$

is the Kullback-Leibler divergence (1.2.2) between $g(Y_{\text{miss}})$ and $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})$. ♦

Proof. Recall from before:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = \ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta}) - \ln f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}).$$

Let $g(Y_{\text{miss}})$ denote some density on Y_{miss}. Adding and subtracting $\ln\{g(Y_{\text{miss}})\}$:

$$\begin{aligned}\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) &= \{\ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta}) - \ln g(Y_{\text{miss}})\} - \big\{\ln f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}) - \ln g(Y_{\text{miss}})\big\} \\ &= \ln\left\{\frac{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} - \ln\left\{\frac{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\}.\end{aligned}$$

Taking the expectation with respect to $g(Y_{\text{miss}})$:

$$\begin{aligned}\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) = &\int \ln\left\{\frac{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}} \\ &- \int \ln\left\{\frac{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}}.\end{aligned}$$

The first term is the evidence lower bound:

$$\mathcal{L}\big\{\boldsymbol{\theta}, g(Y_{\text{miss}})\big\} = \int \ln\left\{\frac{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}}.$$

The second term is the Kullback-Leibler divergence between $g(Y_{\text{miss}})$ and $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})$:

$$\mathbb{KL}\big\{g(Y_{\text{miss}})||f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})\big\} = -\int \ln\left\{\frac{f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})}{g(Y_{\text{miss}})}\right\} g(Y_{\text{miss}})dY_{\text{miss}}.$$

■

Corollary 1.4.1. Since, from Gibbs' inequality (1.2.1), $\mathbb{KL}(\cdot||\cdot) \geq 0$, the observed data log likelihood is not less than the evidence lower bound:

$$\ln f(Y_{\text{obs}}|\boldsymbol{\theta}) \geq \mathcal{L}\big\{\boldsymbol{\theta}, g(Y_{\text{miss}})\big\}.$$

♣

Corollary 1.4.2. Substituting $f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})$ for $g(Y_{\text{miss}})$ in the expression for the evidence lower bound gives:

$$\mathcal{L}\big\{\boldsymbol{\theta}, f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})\big\} = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}\big\{\boldsymbol{\theta}^{(r)}||\boldsymbol{\theta}^{(r)}\big\}.$$

♣

Proof.

$$\begin{aligned}\mathcal{L}\big\{\boldsymbol{\theta}, f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})\big\} &= \int \ln\left\{\frac{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})}{f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})}\right\} f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})dY_{\text{miss}}. \\ &= \int \ln\big\{f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta})\big\} f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})dY_{\text{miss}} \\ &\quad - \int \ln\big\{f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})\big\} f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})dY_{\text{miss}} \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) + \mathcal{H}\big\{\boldsymbol{\theta}^{(r)}||\boldsymbol{\theta}^{(r)}\big\}.\end{aligned}$$

■

1.5 Information Matrix

Definition 1.5.1. The **complete data score** is:

$$\mathcal{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(Y_{\text{obs}}, Y_{\text{miss}}|\boldsymbol{\theta}).$$

The **EM score** is:

$$\mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}). \tag{1.5.6}$$

■

Proposition 1.5.5. If differentiation in $\boldsymbol{\theta}$ commutes with integration over $f(Y_{\text{miss}}|Y_{\text{obs}}, \boldsymbol{\theta})$, then:

$$\mathbb{E}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}}, \boldsymbol{\theta}^{(r)}\big\} = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}).$$

That is, the EM score is the expectation of the complete data score given $(Y_{\text{obs}}, \boldsymbol{\theta}^{(r)})$. ♦

Proof.

$$\mathbb{E}\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\} = \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta})\big|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\right\}$$
$$= \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\left\{\ln f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta})\big|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\right\} = \frac{\partial}{\partial\boldsymbol{\theta}}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}).$$

■

Definition 1.5.2. The **complete data expected information** is:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\big\{\mathcal{U}(\boldsymbol{\theta})\otimes\mathcal{U}(\boldsymbol{\theta})\big\}.$$

The **EM expected information** is the variance of the EM score:

$$\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\big\}.$$

■

Proposition 1.5.6 (**Total Variance Decomposition**).

$$\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathcal{I}(\boldsymbol{\theta}) - \mathbb{E}\Big[\mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\big\}\Big]. \tag{1.5.7}$$

♦

Proof. Observe that:

$$\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\big\} = \mathbb{V}\Big[\mathbb{E}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\big\}\Big].$$

By law of total variance:

$$\begin{aligned}\mathcal{I}(\boldsymbol{\theta}) =& \mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta})\big\}\\ =&\mathbb{E}\Big[\mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\big\}\Big] + \mathbb{V}\Big[\mathbb{E}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\big\}\Big]\\ =&\mathbb{E}\Big[\mathbb{V}\big\{\mathcal{U}(\boldsymbol{\theta})|Y_{\text{obs}},\boldsymbol{\theta}^{(r)}\big\}\Big] + \mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}).\end{aligned}$$

■

Remark 1.5.3. The EM information may be found either by:

1. Directly finding the variance of the EM score in (1.5.6).
2. Applying the total variance decomposition in (1.5.7).

To obtain a final estimate of the information for use in inference, $\mathcal{I}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ will be evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r)}$, so the distinction between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(r)}$ may be dropped. ♦

Extensions

2.1 Monte Carlo EM

Monte Carlo (MC) EM is useful when the expectation in the E-step is intractable, but it is possible to simulate from $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta})$. In MC-EM, on the E-step, a sample $(Y_{\text{miss}}^{(1)},\cdots,Y_{\text{miss}}^{(M)})$ of size M is drawn from $f(Y_{\text{miss}}|Y_{\text{obs}};\boldsymbol{\theta}^{(r)})$, and the EM objective (1.3.3) is approximated by:

$$\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \frac{1}{M}\sum_{m=1}^{M}\ln f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta}).$$

Similarly, the score may be approximated by:

$$\hat{\mathcal{U}}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M}\dot{\ell}(Y_{\text{obs}},Y_{\text{miss}}^{(m)}|\boldsymbol{\theta}),$$

where:

$$\dot{\ell}(Y_{\text{obs}},Y_{\text{miss}}^{(m)}|\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ln f(Y_{\text{obs}},Y_{\text{miss}}|\boldsymbol{\theta}).$$

Finally, the observed information is approximated as:

$$\hat{\mathcal{J}}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M}\dot{\ell}(Y_{\text{obs}},Y_{\text{miss}}^{(m)}|\boldsymbol{\theta})\otimes\dot{\ell}(Y_{\text{obs}},Y_{\text{miss}}^{(m)}|\boldsymbol{\theta}).$$

2.2 Expectation Conditional Maximization

The expectation conditional maximization (ECM) algorithm is useful when jointly maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ with respect to all of $\boldsymbol{\theta}$ challenging; however, the maximization may be split into a sequence of simpler, conditional maximizations. For example, suppose there is a natural partition of the parameter as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_K)$. The M-step of ECM proceeds as follows:

$$\begin{aligned}
\boldsymbol{\theta}_1^{(r+1)} &\leftarrow \arg\max_{\boldsymbol{\theta}_1\in\Theta_1} Q(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2^{(r)},\cdots,\boldsymbol{\theta}_K^{(r)}|\boldsymbol{\theta}^{(r)}) \\
\boldsymbol{\theta}_2^{(r+1)} &\leftarrow \arg\max_{\boldsymbol{\theta}_2\in\Theta_2} Q(\boldsymbol{\theta}_1^{(r+1)},\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_K^{(r)}|\boldsymbol{\theta}^{(r)}) \\
&\vdots \\
\boldsymbol{\theta}_K^{(r+1)} &\leftarrow \arg\max_{\boldsymbol{\theta}_K\in\Theta_K} Q(\boldsymbol{\theta}_1^{(r+1)},\boldsymbol{\theta}_2^{(r+1)},\cdots,\boldsymbol{\theta}_K|\boldsymbol{\theta}^{(r)})
\end{aligned}$$

Moreover, for each conditional maximization, either the EM objective $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ or the observed data log likelihood $\ln f(Y_{\text{obs}}|\boldsymbol{\theta})$ may be maximized. This is useful when the observed data log likelihood admits closed form solutions for some components of $\boldsymbol{\theta}$.