

Missing data lecture 11: More EM

One more EM convergence result

Another result is that if θ^{t+1} is chosen such that:

1. $\frac{\partial}{\partial \theta} Q(\theta | \theta^t) |_{\theta=\theta^{t+1}} = 0$
2. $\theta^{t+1} \rightarrow \theta^*$
3. $f(Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta)$ is sufficiently smooth in θ

Then

$$\frac{\partial}{\partial \theta} \ell(\theta | Y_{(0)} = \tilde{y}_{(0)}) |_{\theta=\theta^*} = 0$$

$$\frac{\partial}{\partial \theta} \ell(\theta | Y_{(0)} = \tilde{y}_{(0)}) |_{\theta=\theta^*} = \frac{\partial}{\partial \theta} Q(\theta | \theta^t) |_{\theta=\theta^*} - \frac{\partial}{\partial \theta} H(\theta | \theta^t) |_{\theta=\theta^*}$$

The first quantity on the RHS is zero by the condition that we pick θ^{t+1} as that which leads to $\frac{\partial}{\partial \theta} Q(\theta | \theta^t) |_{\theta=\theta^*} = 0$, so if $\theta^t \rightarrow \theta^*$, we must have gotten there by setting these gradients equal to zero. The second quantity on the RHS is

$$\begin{aligned} \frac{\partial}{\partial \theta} H(\theta | \theta^t) |_{\theta=\theta^*} &= \frac{\partial}{\partial \theta} \int_{\mathcal{Y}_{(1)}} \log f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta) f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dy_{(1)} |_{\theta=\theta^*} \\ &= \int_{\mathcal{Y}_{(1)}} \frac{\frac{\partial}{\partial \theta} f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta) |_{\theta=\theta^*}}{f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^*)} f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t) dy_{(1)} |_{\theta=\theta^*} \end{aligned}$$

as $\theta^t \rightarrow \theta^*$ the denominators cancel, leaving

$$\int_{\mathcal{Y}_{(1)}} \frac{\partial}{\partial \theta} f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta) |_{\theta=\theta^*} dY_{(1)}$$

which, if we pull the derivative out of the integral again, equals zero because we're differentiating a constant.

Nontrivial EM example

Here's a nontrivial example: Let's say we have three outcomes, so Y is an $n \times 3$ matrix, and Y_i are distributed trivariate normal with means, μ_1, μ_2, μ_3 and covariance matrix Σ . For simplicity's sake, we assume the missingness is ignorable, and assume we have only three missingness patterns:

$$m_i \in \{(0, 0, 1), (1, 0, 0), (0, 0, 0)\}.$$

The parameters of interest are $\mu_1, \mu_2, \mu_3, \Sigma$, and we'd like to use all the available data to do inference. Let the three groups that have missingness be defined as:

$$\begin{aligned} r_1 &= \#(y_{1i} \text{ obs, } , y_{2i} \text{ obs, } y_{3i} \text{ miss. }) \\ r_2 &= \#(y_{1i} \text{ miss, } , y_{2i} \text{ obs, } y_{3i} \text{ obs }) \\ n - r_1 - r_3 &= \#(y_{1i} \text{ obs, } , y_{2i} \text{ obs, } , y_{3i} \text{ obs }) \end{aligned}$$

, and suppose we have arranged our indices i so $i \in \{1, \dots, r_1\}$ have y_1 observed, $i \in \{r_1 + 1, \dots, r_1 + r_2\}$ have y_2 observed and $i \in \{r_1 + r_2 + 1, \dots, n\}$ have all data observed. z Let $f_{Y_j, Y_k}((y_{ji}, y_{ki}) \mid (\mu_j, \mu_k), \Sigma_{jk}), j = 1, 2$ be the bivariate normal density for indices $j, k, j \neq k$, where

$$\Sigma_{jk} = \mathbb{E} [((Y_j, Y_k)^T - (\mu_j, \mu_k)^T)((Y_j, Y_k) - (\mu_j, \mu_k))]$$

while $f_Y(y_i \mid \mu_1, \mu_2, \mu_3, \Sigma)$ is the trivariate normal density. The observed data likelihood is

$$\begin{aligned} L(\mu_1, \mu_2, \mu_3, \Sigma \mid Y_{(0)}) &= \prod_{i=1}^{r_1} f_{Y_1, Y_2}(y_{1i}, y_{2i} \mid \mu_1, \mu_2, \Sigma_{12}) \\ &\quad \prod_{i=r_1+1}^{r_1+r_2} f_{Y_2, Y_3}(y_{2i}, y_{3i} \mid \mu_2, \mu_3, \Sigma_{23}) \\ &\quad \prod_{i=r_1+r_2+1}^n f_Y(y_i \mid \mu_1, \mu_2, \mu_3, \Sigma) \end{aligned}$$

This is going to be a hard function to maximize in terms of Σ . We can find the MLEs for this expression, but it isn't standard, and it'll involve some thinking, whereas if we had a complete dataset we could just do the maximization very easily.

If we had a complete dataset, we know from an earlier lecture the ML solutions for these quantities:

$$\hat{\mu}_1 = \bar{y}_1, \hat{\mu}_2 = \bar{y}_2, \hat{\mu}_3 = \bar{y}_3, \hat{\Sigma} = 1/n \sum_i y_i y_i^T - \hat{\mu} \hat{\mu}^T$$

How can we get an MLE when there is ignorable missingness as described above?

Multivariate normal with missingness

We can run the EM algorithm, which will require computing the $Q(\theta | \theta^t)$ function. For the multivariate normal example, this is:

$$\mathbb{E} \left[\ell_Y(\mu, \Sigma | Y) | Y_{(0)}, \mu^t, \Sigma^t \right] = \frac{1}{2} \log(\det \Sigma^{-1}) - \frac{1}{2} \text{tr} \sum_i \mathbb{E} \left[((y_i - \mu)(y_i - \mu)^T | Y_{(0)}, \mu^t, \Sigma^t) \Sigma^{-1} \right]$$

If we expand out the cross product, we see we need

$$\mathbb{E} \left[(y_i - \mu)(y_i - \mu)^T | Y_{(0)}, \mu^t, \Sigma^t \right] = \mathbb{E} \left[y_i y_i^T | Y_{(0)}, \mu^t, \Sigma^t \right] - \mathbb{E} \left[\mu y_i^T - y_i^T \mu | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu \mu^T$$

The first term is :

$$\mathbb{E} \left[y_i y_i^T | Y_{(0)}, \mu^t, \Sigma^t \right] = \text{Cov}(y_i | Y_{(0)}, \mu^t, \Sigma^t) + \mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] \mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right]^T$$

Plugging this back in above gives:

$$\text{Cov}(y_i | Y_{(0)}, \mu^t, \Sigma^t) + \mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] \mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right]^T - \mathbb{E} \left[\mu y_i^T - y_i^T \mu | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu \mu^T$$

simplifying to

$$\text{Cov}(y_i | Y_{(0)}, \mu^t, \Sigma^t) + (\mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu)(\mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu)^T$$

Plugging this back in above, we get the expected log-likelihood

$$\begin{aligned} \mathbb{E} \left[\ell_Y(\mu, \Sigma | Y) | Y_{(0)}, \mu^t, \Sigma^t \right] &= \frac{1}{2} \log(\det \Sigma^{-1}) \\ &- \frac{1}{2} \text{tr} \sum_i \left(\text{Cov}(y_i | Y_{(0)}, \mu^t, \Sigma^t) + (\mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu)(\mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right] - \mu)^T \right) \Sigma^{-1} \end{aligned}$$

This leads to the M step estimates:

$$\mu^{t+1} = \frac{1}{n} \sum_i \mathbb{E} \left[y_i | Y_{(0)}, \mu^t, \Sigma^t \right],$$

and for

$$\Sigma^{t+1} = \frac{1}{n} \sum_i \text{Cov}(y_i | \tilde{y}_{i(0)}, \mu^t, \Sigma^t) + (\mathbb{E} \left[y_i | \tilde{y}_{i(0)}, \mu^t, \Sigma^t \right] - \mu^{t+1})(\mathbb{E} \left[y_i | \tilde{y}_{i(0)}, \mu^t, \Sigma^t \right] - \mu^{t+1})^T$$

The key is that the conditional expectation and covariances for y_i are informed by the observed data.

Suppose our observation is in the first group, so $m_i = (0, 0, 1)$. Then we need to compute $\text{Cov}(y_i | y_{i1}, y_{i2}, \mu^t, \Sigma^t)$, which is

$$\begin{bmatrix} \text{Var}(y_{i1} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Cov}(y_{i1}, y_{i2} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Cov}(y_{i1}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) \\ \text{Cov}(y_{i1}, y_{i2} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Var}(y_{i2} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Cov}(y_{i2}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) \\ \text{Cov}(y_{i1}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Cov}(y_{i2}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) & \text{Var}(y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) \end{bmatrix}$$

Most of the elements of this matrix are zero, because of the properties of conditional expectation. Look at $\text{Cov}(y_{i1}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t)$, for instance:

$$\begin{aligned} \text{Cov}(y_{i1}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) &= \mathbb{E}[y_{i1}y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] - \mathbb{E}[y_{i1} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] \mathbb{E}[y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] \\ &= \mathbb{E}[y_{i1}y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] - y_{i1} \mathbb{E}[y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] \end{aligned}$$

and similarly

$$\mathbb{E}[y_{i1} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] = y_{i1}$$

so we get

$$\begin{aligned} \text{Cov}(y_{i1}, y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t) &= y_{i1} \mathbb{E}[y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] - y_{i1} \mathbb{E}[y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t] \\ &= 0 \end{aligned}$$

The same holds for the conditional variances. The only nonzero element is $\text{Var}(y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t)$. We can use the properties of the conditional normal distribution to derive the conditional variance.

Let our covariance matrix Σ be defined:

$$\Sigma = \begin{bmatrix} \Sigma_{12} & \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \\ \begin{bmatrix} \sigma_{13} & \sigma_{23} \end{bmatrix} & \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix} \end{bmatrix}, \Sigma_{12} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Then the conditional variance, evaluated at $\mu^{(t)}, \Sigma^{(t)}$ is:

$$(\sigma_3^2)^{(t)} - \begin{bmatrix} \sigma_{13}^{(t)} & \sigma_{23}^{(t)} \end{bmatrix} (\Sigma_{12}^{(t)})^{-1} \begin{bmatrix} \sigma_{13}^{(t)} \\ \sigma_{23}^{(t)} \end{bmatrix}$$

The conditional mean, $\mathbb{E}[y_{i3} | y_{i1}, y_{i2}, \mu^t, \Sigma^t]$ is

$$\mu_3^{(t)} + \begin{bmatrix} \sigma_{13}^{(t)} & \sigma_{23}^{(t)} \end{bmatrix} (\Sigma_{12}^{(t)})^{-1} \begin{bmatrix} y_{i1} - \mu_1^{(t)} \\ y_{i2} - \mu_2^{(t)} \end{bmatrix}$$

Similar formulas can be used for the second group, where you have observed y_{i2}, y_{i3} , but not y_{i1} .

Measuring uncertainty in EM parameter estimates

We might want to find something like standard errors for our parameter estimates. We can get asymptotic standard errors from the inverse of the information matrix: $I(\theta | Y_{(0)})^{-1}$: Starting with our standard decomposition of the observed data log likelihood:

$$\ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) = \log(f(Y_{(0)} = \tilde{y}_{(0)} | \theta)) + \log(f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta))$$

Rearranging

$$\log(f(Y_{(0)} = \tilde{y}_{(0)} | \theta)) = \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) - \log(f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta))$$

we can differentiate twice to get:

$$-\nabla_{\theta}^2 \log(f(Y_{(0)} = \tilde{y}_{(0)} | \theta)) = -\nabla_{\theta}^2 \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) + \nabla_{\theta}^2 \log(f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta))$$

which is equivalent to:

$$I(\theta | Y_{(0)} = \tilde{y}_{(0)}) = I(\theta | Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}) + \nabla_{\theta}^2 \log(f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta))$$

where $I(\theta | Y_{(0)} = \tilde{y}_{(0)})$ is the observed information. Taking expectations with respect random variable $Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t$ gives

$$\begin{aligned} I(\theta | Y_{(0)} = \tilde{y}_{(0)}) &= \mathbb{E}_{Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t} [I(\theta | Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)})] + \mathbb{E}_{Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t} [\nabla_{\theta}^2 \log(f(Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta))] \\ &= -\nabla_{\theta}^2 Q(\theta | \theta^t) + \nabla_{\theta}^2 H(\theta | \theta^t) \end{aligned}$$

When we have reached a stationary point θ^* , we can evaluate the first term on the right easily:

$$-\nabla_{\theta}^2 Q(\theta | \theta^t) |_{\theta = \theta^*}$$

We can get the second term in EM from Louis' Identity:

$$\begin{aligned} -\nabla_{\theta}^2 H(\theta | \theta^t) &= \mathbb{E}_{Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^t} [\nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) \nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)})^T] \\ &\quad - \nabla_{\theta} \ell_Y(\theta | Y_{(0)} = \tilde{y}_{(0)}) \nabla_{\theta} \ell_Y(\theta | Y_{(0)} = \tilde{y}_{(0)})^T \end{aligned}$$

where we note that at the stationary point, the gradient is zero, so we get:

$$-\nabla_{\theta}^2 H(\theta | \theta^t) |_{\theta = \theta^*} = \mathbb{E}_{Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}} [\nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) \nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)})^T] |_{\theta = \theta^*}$$

The final expression is:

$$\begin{aligned} I(\theta^* | Y_{(0)} = \tilde{y}_{(0)}) &= -\nabla_{\theta}^2 Q(\theta | \theta^*) |_{\theta = \theta^*} \\ &\quad - \mathbb{E}_{Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}} [\nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) \nabla_{\theta} \ell_Y(\theta | Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)})^T] |_{\theta = \theta^*} \end{aligned}$$

Variants of EM

The standard EM algorithm has two important characteristics that may make it hard to apply in practice:

1. The expectation step of the log-likelihood might be intractable.
2. The M step might not be able to be done exactly.

There are variants of the EM algorithm that can handle both of these issues.

GEM: Generalized EM

If we can't maximize the $Q(\theta | \theta^t)$ exactly, we can instead set θ^{t+1} so that

$$Q(\theta^t | \theta^t) \leq Q(\theta^{t+1} | \theta^t).$$

We showed last time that any value of θ^{t+1} for which the above holds will lead to an increase in the observed likelihood, which is ultimately what we're trying to maximize.

One can show that we still reach a stationary point for the observed likelihood under a GEM algorithm.

ECM

The first variant of EM is called ECM, or Expectation-Conditional Maximization. This is a GEM algorithm, so if we can't do the maximization step exactly, we can instead do conditional maximization, which at least increases the Q function each iteration. We have an example of this sort of model from the first few weeks of class: the repeated measures model:

$$\begin{aligned} y_i | X_i &= X_i \beta + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \Sigma) \\ \epsilon_i &\perp\!\!\!\perp \epsilon_j \forall i \neq j \end{aligned}$$

We can look back in our notes to see that we have an iterative maximization scheme for this model:

$$\beta^{(t+1)} = \left(\sum_i X_i^T (\Sigma^{(t)})^{-1} X_i \right)^{-1} \sum_i X_i^T (\Sigma^{(t)})^{-1} y_i$$

and

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_i (y_i - X_i \beta^{(t+1)})(y_i - X_i \beta^{(t+1)})^T$$

Each step of the conditional maximization increased the complete data likelihood:

$$\begin{aligned}\ell_Y(\beta^{t+1}, \Sigma^t | Y) &\geq \ell_Y(\beta^t, \Sigma^t | Y) \\ \ell_Y(\beta^{t+1}, \Sigma^{t+1} | Y) &\geq \ell_Y(\beta^{t+1}, \Sigma^t | Y)\end{aligned}$$

Thus, if we had missing outcome data in our regression, we could run an ECM algorithm with the following steps:

1. Initialize with β^1, Σ^1
2. For $t = 2, \dots$ compute
 - a. Find the conditional expectations: $\mathbb{E} [y_i | Y_{(0)}, X, \beta^t, \Sigma^t]$, $\mathbb{E} [y_i y_i^T | Y_{(0)}, X, \beta^t, \Sigma^t]$

- b. Set β^{t+1} to:

$$\beta^{(t+1)} = \left(\sum_i X_i^T (\Sigma^{(t)})^{-1} X_i \right)^{-1} \sum_i X_i^T (\Sigma^{(t)})^{-1} \mathbb{E} [y_i | Y_{(0)}, X, \beta^t, \Sigma^t]$$

- c. Set Σ^{t+1} to:

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_i \mathbb{E} [(y_i - X_i \beta^{(t+1)})(y_i - X_i \beta^{(t+1)})^T | Y_{(0)}, X, \beta^t, \Sigma^t]$$

- d. If $Q(\theta^{t+1} | \theta^t) - Q(\theta^t | \theta^t) \geq \epsilon$, continue, otherwise end

Monte Carlo EM

Monte Carlo EM (MCEM) algorithms can be used when the conditional expectation is not possible to do analytically. That is, if we cannot do this integral:

$$Q(\theta | \theta^{(t)}) = \int_{\mathcal{Y}_{(1)}} \ell_Y(\theta | Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}) f(Y_{(1)} = y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^{(t)}) dy_{(1)}$$

we can instead approximate it with a Monte Carlo estimator. Assuming we can draw samples from $y_{(1)}^s \sim f(Y_{(1)} | Y_{(0)} = \tilde{y}_{(0)}, \theta^{(t)})$, we can compute:

$$\hat{Q}(\theta | \theta^{(t)}) = \frac{1}{S} \sum_{s=1}^S \ell_Y(\theta | Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}^s)$$

which we know:

$$\lim_{S \rightarrow \infty} \hat{Q}(\theta | \theta^{(t)}) = Q(\theta | \theta^{(t)})$$

This change in the algorithm isn't without its complications, because we now have to determine how to measure convergence. While it is true that $Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)})$, we now have to contend with the fact that there is noise in our assessment of convergence, so we need something more like:

$$P(\hat{Q}(\theta^{(t+1)} | \theta^{(t)}) - \hat{Q}(\theta^{(t)} | \theta^{(t)}) \geq \epsilon)$$