

HW3 - Selected Solutions

Recap: To maximize $l(\theta|Y_{obs}) = \log P(Y_{obs}|\theta)$, we construct,

$$P(Y_{obs}, Y_{mis}|\theta) \text{ s.t. } \int P(Y_{obs}, Y_{mis}|\theta).dY_{mis} = P(Y_{obs}|\theta)$$

and use EM: $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta^{(t)})$

where: $Q(\theta|\theta^{(t)}) = \mathbb{E}[\log P(Y_{obs}, Y_{mis}|\theta)|Y_{obs}, \theta^{(t)}]$

Last time: $l(\theta^{(t+1)}) \geq l(\theta^{(t)})$ (corresponding to monotone convergence)

Today:

- (1) Some more theory
- (2) What do when the maximization is hard
- (3) What to do when it is hard to compute expectation

From the monotonicity proof, we saw that -

$$l(\theta^{(t+1)}) - l(\theta^{(t)}) = [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})] + [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})],$$

where the latter term is non-negative for any $\theta^{(t)}$

So, we just need to ensure that,

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

to guarantee monotone convergence.

Thus, we don't need to maximize Q; we just need to increase it. This approach is referred to as Generalized EM (GEM).

Convergence rate of EM:

Idea: EM gives an update - $\vec{\theta}^{(t+1)} = M(\vec{\theta}^{(t)})$

Here, M is the update/mapping operator: $M : \mathbb{R}^p \rightarrow \mathbb{R}^p$

To study convergence rate - let θ^* be the MLE. Then:

$$M(\theta^{(t)}) = \theta^* + \left[\frac{\partial}{\partial \theta} M(\theta) \right]_{\theta=\theta^*} \cdot [\theta^{(t)} - \theta^*] + O(\|\theta^{(t)} - \theta^*\|^2)$$

$$\Rightarrow \theta^{(t+1)} - \theta^* = M(\theta^{(t)}) - \theta^* \simeq DM(\theta^*) \cdot (\theta^{(t)} - \theta^*)$$

$$\Rightarrow \lim_{t \rightarrow \infty} \frac{\|\theta^{(t+1)} - \theta^*\|}{\|\theta^{(t)} - \theta^*\|} = \rho$$

which means we have a *linear rate of convergence* to ρ , the maximal eigenvalue of DM.

Idea: It can be shown that DM also has a representation as the ‘fraction of missing information’ (not covered here).

What to do when the E-step is hard:

Example: Probit regression -

We saw that,

$$Z_i^{(t+1)} \Big| \beta^{(t)}, y_i = \begin{cases} TN(x_i^T \beta^{(t)}, 1; (-\infty, 0]), & \text{if } y_i = 0 \\ TN(x_i^T \beta^{(t)}, 1; [0, \infty)), & \text{if } y_i = 1 \end{cases}$$

Suppose we did not know the conditional expectation of a truncated normal - then what do we do? We have the following options:

1. Use Monte Carlo to simulate from $N(x_i^T \beta^{(t)}, 1)$, throw away any samples outside the range, and compute the mean
2. Simulate from $N(x_i^T \beta^{(t)}, 1)$; flip sign if necessary and compute the mean.
3. Use inverse CDF sampling.
4. If independent samples are not possible, we know that, $P(Y_{mis} | Y_{obs}, \theta^{(t)}) \propto P(Y_{mis}, Y_{obs} | \theta^{(t)})$.

So, can sample from $Y_{mis} | (Y_{obs}, \theta^{(t)})$ using MCMC; use samples to get the desired conditional expectations.

5. For 1-D, we can use Numerical Integration (with the Trapezoidal rule, Quadrature, etc.)

Sampling truncated Random Variables:

Let $X \sim F_X$. Let $Z \sim X \mid X \in (a, b)$

To sample from Z , sample from $U \sim U(F_X(a), F_X(b))$

Then, $Z = F_X^{-1}(U)$

It can be shown that $Z \sim X \mid X \in (a, b)$

Note: EM using a Monte Carlo E-step is called MCEM or MCMCEM (when using MCMC in the E-step).

(Example code and path shown on projector for MCEM applied to the probit EM example when $Z_i^{(t+1)} = \mathbb{E}[Z_i \mid Y_i, \beta^{(t)}]$ is computed using inverse CDF sampling)

What do to when the M-step is hard:

Suppose $\theta \in \mathbb{R}^p$ and finding $\underset{\theta}{\text{Argmax}} Q(\theta \mid \theta^{(t)})$ is hard. Then,

1. Increase Q - that is, ensure that, $Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$ to get GEM.

2. Conditionally maximize $Q(\theta \mid \theta^{(t)})$. For example, let $\theta \in \mathbb{R}^2$.

Set, $\theta_1^{(t+1)} = \underset{\theta_1}{\text{argmax}} Q(\theta_1, \theta_2^{(t)} \mid \theta_1^{(t)}, \theta_2^{(t)})$

Set, $\theta_2^{(t+1)} = \underset{\theta_2}{\text{argmax}} Q(\theta_1^{(t+1)}, \theta_2 \mid \theta_1^{(t)}, \theta_2^{(t)})$

Note:

- E-step is not recomputed between maximizations.
- Not guaranteed to give us the joint maximum over both θ_1 and θ_2

Example:

Let $Y_i \mid \alpha, \beta \sim \text{Gamma}(\alpha, \beta)$

Let $i = 1, \dots, n = n_{\text{obs}} + n_{\text{mis}}$

(Some Y_i s are missing - this is independent of all model components)

$$P(Y_i \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i} \quad (y_i, \alpha, \beta > 0)$$

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E} \left[n(\alpha \log \beta - \log \Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^n \log(y_i) - \beta \sum_{i=1}^n y_i \mid Y_{\text{obs}}, \theta^{(t)} \right]$$

This is harder to maximize w.r.t α .
 We maximize w.r.t. β as follows:

$$\frac{\partial Q}{\partial \beta} = \frac{n\alpha}{\beta} - \left(\sum_{i=1}^n y_i + \sum_{i=n_{obs}+1}^{n_{obs}+n_{mis}} \mathbb{E}[Y_i|Y_{obs}, \theta^{(t)}] \right)$$

Set $\alpha = \alpha^{(t)}$. Solving for $\frac{\partial Q}{\partial \beta} = 0$, we get,

$$\beta^{(t+1)} = \frac{n\alpha^{(t)}}{\sum_{i=1}^{n_{obs}} y_i + n_{mis} \left(\frac{\alpha^{(t)}}{\beta^{(t)}} \right)}$$

Now, to maximize w.r.t α :

$$\frac{\partial Q}{\partial \alpha} = n \log \beta - n\psi_o(\alpha) + \sum_{i=1}^{n_{obs}} \log(y_i) + n_{mis} (\psi_o(\alpha^{(t)}) - \log \beta^{(t)})$$

where, $\psi_r(\alpha) = \frac{\partial^{r+1}}{\partial \alpha^{r+1}} (\log \Gamma(\alpha))$

If $\frac{\partial Q}{\partial \alpha} = 0$, then use Newton-Raphson

Next, $\frac{\partial^2 Q}{\partial \alpha^2} = -n(\psi_1(\alpha))$

Let $\alpha_{NR}^{(0)} = \alpha^{(t)}$ (here, we set $j=0$)

Set $\alpha_{NR}^{j+1} = \alpha_{NR}^{(j)} + \frac{g(\alpha_{NR}^{(j)})}{n\psi_1(\alpha_{NR}^{(j)})}$

Increment $j \rightarrow j + 1$ until convergence.

Finally, set $\alpha^{(t+1)} = \alpha_{NR}^*$,

where $g(\cdot)$ is $\frac{\partial Q}{\partial \alpha}$, the function we are seeking the root of.