STA250 Lecture 14

November 18th, 2013

Recap: To maximize $l(\theta|y_{obs}) = \log P(y_{obs}|\theta)$, we construct $P(y_{obs}, y_{mis}|\theta)$, 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 $Q(\theta|\theta^{(t)}) = E[logP(y_{obs}, y_{mis}|\theta)|y_{obs}, \theta^{(t)}]$

Last time: $l(\theta^{(t+1)}) \geq l(\theta^{(t)})$ [monotone convergence]
Today:

1. A bit more theory
2. What to do when maximization is hard
3. What to do when the expectation is hard to compute

Note: From the proof for monotonicity:

$$0 \leq l(\theta^{(t+1)}) - l(\theta^{(t)}) = [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})] + [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})].$$

Since $H(\theta^{(t+1)}|\theta^{(t)}) \geq H(\theta^{(t)}|\theta^{(t)})$ always holds, one can obtain $l(\theta^{(t+1)}) - l(\theta^{(t)})$ as long as $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$, i.e., we still get monotone convergence! This suggests that we don't need to maximize Q, but rather simply increase it.

This is called **Generalized EM (GEM)**.

Convergence rate of EM: idea: EM gives an update $\theta^{(t+1)} = M(\theta^{(t)})$, i.e., a function of $\theta^{(t)}$. Here M is the update mapping/operator, where $\theta \in \mathrm{R}^p$, $M : \mathrm{R}^p \to \mathrm{R}^p$.

To study convergence rate, let θ_* be the MLE, then:

$$(\text{near } \theta_*) \quad M(\theta^{(t)}) = \theta^* + (\theta^{(t)} - \theta_*)\frac{\partial}{\partial \theta} M(\theta)|_{\theta=\theta_*} + o(||\theta^{(t)} - \theta_*||^2)$$

We see:

$$\begin{aligned} \theta^{(t+1)} - \theta_* &= M(\theta^{(t)}) - \theta_* \\ &\approx \mathrm{DM}(\theta_*) \times (\theta^{(t)} - \theta_*), \end{aligned}$$

then we see that:

$$\lim_{t \to \infty} \frac{||\theta^{(t+1)} - \theta_*||}{||\theta^{(t)} - \theta_*||} = \rho,$$

where ρ is the maximal eigenvalue of DM.

Aside: It can also be shown that DM has a representation as the 'fraction of missing information'.

When the E-step is hard:

example:[Probit regression]

We saw that $Z_i^{(t+1)}|(\beta^{(t+1)}, y_i) = \begin{cases} TN(X_i^T\beta^{(t)}, 1; (-\infty, 0]) \text{ if } y_i = 0 \\ TN(X_i^T\beta^{(t)}, 1; [0, \infty)) \text{ if } y_i = 1 \end{cases}$.

Suppose we didn't know the expected value of a truncated normal, what could we do?

Monte Carlo :

1. Simulate from $N(X_i^T\beta^{(t)}, 1)$, then throw away any samples outside the range and compute the mean.
2. Simulate from $N(X_i^T\beta^{(t)}, 1)$, and flip sign if needed (works only for truncation at 0).
3. Inverse-CDF sampling
4. Rejection sampling
5. Sample from $p(y_{\text{mis}}|y_{\text{obs}}, \theta^{(t)})$ using MCMC, and use samples to approximate the desired conditional expectation.

Sampling truncated r.v.'s: (Monte Carlo method 3 (above))

Let $X \sim F_x$, $Z \sim X|X \in (a,b)$, to sample from Z:

$$U \sim \mathrm{U}(F_x(a), F_x(b))$$

Let $Z = F_x^{-1}(U)$, then we can show that $Z \sim X|X \in (a,b)$. Note that in general this method works if you can compute F_x^{-1}, $F_x(a)$, $F_y(b)$ stably, which is not always the case.

Numerical integration $\begin{cases} \text{Trapezadal} & \text{usually restricted to univariate} \\ \text{Quadrature} & \text{or lower dimensional settings} \end{cases}$

EM using a Monte Carlo E-step (as Monte Carlo method 4 listed above) is called **MCEM** (or **MCMCEM**).

Let's see MCEM for the Probit EM example where $Z_i^{(t+1)} = E[Z_i|y_i, \beta^{(t)}]$ is computed using inverse CDF sampling method.

Remark: MCEM can't achieve monotone increasing property of EM, it only produces an approximate version of Q.

It is trickier to decide the convergence criterion for MCEM. See Levine & Casella (2001) on webpage for more on MCEM.

When the M-step is hard

Suppose $\theta \in \mathrm{R}^p$, and finding $\mathrm{argmax}_\theta\, Q(\theta|\theta^{(t)})$ is hard, what to do?

–Option 1: Just increase Q (i.e., let $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$ and we get a **GEM**

–Option 2: Conditionally maximize $Q(\theta|\theta^{(t)})$,i.e.,

e.g., $\theta \in \mathrm{R}^2$, $\theta = (\theta_1, \theta_2)^T$, set

$$\theta_1^{(t+1)} = \underset{\theta_1}{\mathrm{argmax}}\, Q\big((\theta_1, \theta_2^{(t)})|(\theta_1^{(t)}, \theta_2^{(t)})\big)$$

$$\theta_2^{(t+1)} = \underset{\theta_1}{\mathrm{argmax}}\, Q\big((\theta_1^{(t+1)}, \theta_2)|(\theta_1^{(t)}, \theta_2^{(t)})\big)$$

Note: the E-step is not re-computed between the maximizations.

See: Meng & Rubin(1993) for ECM + convergence properties.

example: $y_i|\alpha, \beta \sim \mathrm{Gamma}(\alpha, \beta)$, $i = 1, 2, \ldots, n_{\mathrm{obs}} + n_{\mathrm{mis}} = n$. Denote $\theta = (\alpha, \beta)$.

Some $y_i's$ are missing (assuming missingness is independent of all model components – to make things simple)

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

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

$$\frac{\partial Q}{\partial \beta} = \frac{n\alpha}{\beta} - \big(\sum_{i=1}^{n_{\text{obs}}} y_i + \sum_{i=n_{\text{obs}+1}}^{n_{\text{obs}}+n_{\text{mis}}} E[y_i|y_{\text{obs}}, \theta^{(t)}]\big)$$

Set $\alpha = \alpha^{(t)}$, solving for $\frac{\alpha Q}{\alpha \beta} = 0$, one gets:

$$\beta^{(t+1)} = n\alpha^{(t)} / \big(\sum_{i=1}^{n_{\text{obs}}} y_i + n_{\text{mis}} \frac{\alpha^{(t)}}{\beta^{(t)}}\big).$$

To maximize w.r.t α:

$$\frac{\partial Q}{\partial \alpha} = n \log \beta - n\Psi_0(\alpha) + [\sum_{i=1}^{n_{\text{obs}}} \log y_i + n_{\text{mis}} \big(\Psi_0(\alpha^{(t)} - \log(\beta^{(t)})\big)] = g(\alpha),$$

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

<u>FACT:</u> $y \sim Gamma(\alpha, \beta)$, $E[\log y] = \Psi_0(\alpha) - \log \beta$

Set $\frac{\partial Q}{\partial \alpha} = 0$, use Newton-Raphson (NR), $\frac{\partial^2 Q}{\partial \alpha^2} = -n\Psi_1(\alpha)$.
<u>Use NR</u>: Let $\alpha_{\text{NR}}^{(0)} = \alpha^{(t))}$, set $j = 0$
set $\alpha_{\text{NR}}^{(j+1)} = \alpha_{\text{NR}}^{(j)} + \frac{g(\alpha_{\text{NR}}^{(j)})}{n\Psi_1(\alpha_{\text{NR}}^{(j)})}$, increment $j \to j+1$ until convergence.
Set $\alpha^{(t+1)} = \alpha_{\text{NR}}^*$ – final value from NR.