Lecture note 13 for STA 250

Shuyang Li

November 13, 2013

Recap: We saw that EM can be used to maximize certain forms of complicated likelihood. The algorithm is as follows:

$$\theta = arg\max_{\theta} Q(\theta|\theta^{(t)}) \tag{1}$$

$$Q(\theta|\theta^{(t)}) = E[\log\left[P(y_{obs}, y_{mis}|\theta)\right]|y_{obs}, \theta^{(t)}] \tag{2}$$

Note: EM maximizes $P(y_{obs}|\theta)$ by using $P(y_{obs}, y_{mis}|\theta)$, which satisfies:

$$\int P(y_{obs}, y_{mis}|\theta) dy_{mis} = P(y_{obs}|\theta) \tag{3}$$

Two key points:

- y_{mis} doesn't have to correspond to "real" missing data;
- choice of y_{mis} is not unique.

One example is as follows:

$$Model : y_{obs}|\theta \sim N(\theta, V+1) \tag{4}$$

The goal is to find MLE for θ(the answer is y_{obs}). There is no missing data in this example. Consider a "complete" model:

$$y_{obs}|y_{mis} \sim N(y_{mis}, 1) \tag{5}$$

$$y_{mis} \sim N(\theta, V) \tag{6}$$

First we need to check whether $\int P(y_{obs}, y_{mis}|\theta)dy_{mis} = P(y_{obs}|\theta)$. We can show (standard result) that this model fulfils this equation. Here y_{mis} is not "real" missing data. The way to check this equation is to do the expansion as follows:

$$\int P(y_{obs}, y_{mis}|\theta) dy_{mis} = \int P(y_{obs}|y_{mis}) P(y_{mis}|\theta) dy_{mis} \tag{7}$$

The model stated here is noted as Model(1). We can also use a different "complete" data model as follows:

$$y_{obs}|\tilde{y_{mis}}, \theta \sim N(\tilde{y_{mis}} + \theta, V) \tag{8}$$

$$\tilde{y_{mis}} \sim N(0, 1) \tag{9}$$

$$\tag{10}$$

we can demonstrate $\int P(y_{obs}, \tilde{y_{mis}}|\theta)d\tilde{y_{mis}} = P(y_{obs}|\theta)$. We denote this model as Model(2).
The EM algorithm for Model(1) is:

$$\theta^{t+1} = \frac{1}{V+1}\theta^t + \frac{V}{V+1}y_{obs} \tag{11}$$

The EM algorithm for Model(2) is:

$$\theta^{t+1} = \frac{V}{V+1}\theta^t + \frac{1}{V+1}y_{obs} \tag{12}$$

Hence we have two EMs corresponding to two complete data model, both of which give the same MLE. But which is better?

Model 1 has linear convergence rate $\frac{1}{V+1}$

Model 2 has linear convergence rate $\frac{V}{V+1}$

For the convergene rate, the lower, the better. Hence we choose the one with lower convergence rate as the optimimum model. Model 1 is known as a sufficient augmentation scheme as y_{mis} is a sufficient statistics for θ in the "complete" data model. Model 2 is known as an ancillary augmentation scheme as $\tilde{y_{mis}}$ doesn't depend on θ.
It turns out that the EM algorithm has an important property: *Monotone convergence*:

$$l(\theta^{t+1}) \geq l(\theta^t) \tag{13}$$

Where

$$l(\theta) = \log P(y_{obs}|\theta)$$

This property makes EM very stable and popular. NR, bisection, scoring don't have this property. Below is the proof of this property.
Note $P(y_{obs}, y_{mis}|\theta) = P(y_{obs}|\theta)P(y_{mis}|y_{obs}, \theta)$. Let $l(\theta) = \log P(y_{obs}|\theta)$, we can get:

$$l(\theta) = \log P(y_{obs}, y_{mis}|\theta) - \log P(y_{mis}|y_{obs}, \theta). \tag{14}$$

Integrate both sides with respect to $P(y_{mis}|y_{obs}, \theta^t)$. Left side $= \int [\log P(y_{obs}|\theta)]P(y_{mis}|y_{obs}, \theta^t)dy_{mis}$. As $\log P(y_{obs}|\theta)$ is note related to y_{mis},

$$\int [\log P(y_{obs}|\theta)]P(y_{mis}|y_{obs}, \theta^t)dy_{mis} = \log P(y_{obs}|\theta) = l(\theta)) \tag{15}$$

For the right side of equation (14), after integration, the first iterm in the right side is $Q(\theta|\theta^t)$. Denote the second item after integration as $H(\theta|\theta^t)$, where:

$$H(\theta|\theta^t) = -\int \log P(y_{mis}|y_{obs}, \theta)P(y_{mis}|y_{obs}, \theta^t)dy_{mis} \tag{16}$$

So

$$l(\theta^{t+1}) - l(\theta^t) = \left[Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t\right] + \left[H(\theta^{t+1}|\theta^t) - H(\theta^t|\theta^t)\right] \tag{17}$$

In equation (17), on the right hand side, $\Delta Q = \left[Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t\right]$ is always ≥ 0 because in the M step, we are maximizing Q . So we only need to show $\Delta H = \left[H(\theta^{t+1}|\theta^t) - H(\theta^t|\theta^t)\right] \geq 0$.
Here

$$\Delta H = \int \log\left(\frac{p(y_{mis}|y_{obs}, \theta^t)}{p(y_{mis}|y_{obs}, \theta^{t+1})}\right) P(y_{mis}|y_{obs}, \theta^t)dy_{mis} \tag{18}$$

We know equation (18) is the KL divergence $= \text{KL}\left[P(y_{mis}|y_{obs}, \theta^t||P(y_{mis}|y_{obs}, \theta^{t+1}))\right]$. By properties of KL divergence, we know $\Delta H \geq 0$. $\Delta H = 0$ if and only if $P(y_{mis}|y_{obs}, \theta^t) = P(y_{mis}|y_{obs}, \theta^{t+1})$.
Therefore

$$l(\theta^{t+1}) - l(\theta^t) \geq 0 \tag{19}$$

Aside: We can use EM to find posterior models, not just MLE's.
To maximize $\log(\theta|y_{obs})$, Let

$$\begin{aligned} Q(\theta|\theta^t)_{MAP} &= E\left[P(\theta, y_{mis}|y_{obs})|y_{obs}, \theta^t\right] && (20)\\ &= \int \log\left[P(\theta, y_{mis}|y_{obs})\right] P(y_{mis}|y_{obs}, \theta^t)dy_{mis} && (21)\end{aligned}$$

Example:
Probit Regression:

$$y_i|x_i, \beta \sim Bin(1, g({x_i}^T\beta)) \tag{22}$$

For logistic regression,

$$g(\mu) = \frac{e^\mu}{1+e^\mu} \tag{23}$$

For probit regression,

$$g(\mu) = \Phi(\mu) \tag{24}$$

Here Φ is the CDF of N(0,1).
We form a complete data model:

$$y_i|z_i, \beta = I_{z_i \geq 0} \tag{25}$$

$$z_i|\beta \sim N({x_i}^T\beta, 1) \tag{26}$$

This model could be connnected to the patient's response to some drugs in real application.
Here the complete data is $\{(y_i, z_i), i = 1, ..., n\}$, the parameter is β, the observed data is $\{y_i, i = 1, ..., n\}$, the missing data is $\{z_i, i = 1, ..., n\}$.
First we need to check $\int P(y_i, z_i|\beta)dz_i = P(y_i)$. We need Prof Baines to provide information for this demonstration.
Let's derive the EM algorithm for this model:

$$Q(\beta|\beta^t) = E\left[\log P(y, z|\beta)|y, \beta^t\right] \tag{27}$$

To take the expectations, we need to know the distribution of $z_i|y_i, \beta^t$.

$$z_i|y_i = 0, \beta^t \sim TN({x_i}^T\beta^t, 1, [-\infty, 0]) \tag{28}$$

$$z_i|y_i = 1, \beta^t \sim TN({x_i}^T\beta^t, 1, [0, \infty]) \tag{29}$$

$$Q(\beta|\beta^t) = -E\left[\frac{1}{2}\sum_{i=1}^{n}(z_i - {x_i}^T\beta)^2|y_i, \beta^t\right] \tag{30}$$

Maximize $Q(\beta|\beta^t)$.
Let

$$ {z_i}^{t+1} = \begin{cases} {x_i}^T\beta^t + \frac{\Phi\left({x_i}^T\beta^t\right)}{1-\Phi(-{x_i}^T\beta^t)} & \text{if } y_i = 1 \\ {x_i}^T\beta^t - \frac{\Phi\left({x_i}^T\beta^t\right)}{\Phi(-{x_i}^T\beta^t)} & \text{if } y_i = 0 \end{cases} $$

The maximizer of $Q(\beta|\beta^t)$ with respect to β is seen to be the LS estimator of β when regressing z^{t+1} on x.

$$\beta^{t+1} = (x^T x^{-1})x^T z^{t+1} \tag{31}$$

where

$$z^{t+1} = ({z_1}^{t+1}, {z_2}^{t+1}, \ldots, {z_n}^{t+1})^T \tag{32}$$

- E-step $\rightarrow$ to compute z^{t+1}
- M-step $\rightarrow$ to compute $\beta^{t+1} = (x^T x^{-1})x^T z^{t+1}$