STA250 Lecture 15 Notes

Chia-Tung Kuo

November 20, 2013

Recap

Last time we saw strategies to deal with "complicated" EM applications (i.e. when the E-step and/or M-step are hard).

Speeding up EM

Example:

$$y_{obs}|y_{mis} \sim N(y_{mis}, 1)$$
$$y_{mis}|\theta \sim N(\theta, v)$$

"sufficient augmentation" (SA): $\theta^{(t+1)} = \frac{\theta^{(t)}+vy_{obs}}{v+1}$, rate of convergence $= \frac{1}{v+1}$. We also saw the ancillary augmentation (AA)

$$y_{obs}|y_{mis} \sim N(y_{mis}+\theta, 1)$$
$$y_{mis}|\theta \sim N(0, v)$$

AAEM: $\theta^{(t+1)} = \frac{\theta^{(t)}v+y_{obs}}{v+1}$. Rate of convergence $\frac{v}{v+1}$.

So the two algorithms have "opposite" performance oas v chanages. If v is unknown we can derive EM's for SA & AA and have similar performance to when v is known. For a given problem how do we decide whether to use the SA or AA?

- could code both and see which converges faster.

One idea would be to "alternate" updates according to the SA & AA. i.e. compute

$$\theta^{(t+0.5)} = \frac{\theta^{(t)}+vy_{obs}}{v+1} \qquad \text{(SA)}$$
$$\theta^{(t+1)} = \frac{\theta^{(t+0.5)}v+y_{obs}}{v+1} \qquad \text{(AA)}$$

i.e. $\theta^{(t+1)} = M_{AA}(M_{SA}(\theta^{(t)}))$. Note: computation time of the two algorithms may not be equal.

Pros:

- Avoids the need to select one of the algorithms

Cons:

- Do no better than the best of the two algorithms, no worse than the worst of the tow algorithms
- Need to implement both algorithms

It turns out there is a way to "combine" two EM's into a single improved update that utilizes "joint information" contained in the 2 EM's.

Consider E-step in AA: $\tilde{y}_{mis}^{(t)} = \mathbb{E}[\tilde{y}_{mis}|y_{obs}, \theta^{(t)}]$
M-step in AA: $\theta^{(t+0.5)} = y_{obs} - \tilde{y}_{mis}$ $\quad (\frac{\theta^{(t)}v+\tilde{y}_{obs}}{v+1})$
$y_{mis} = H(\tilde{y}_{mis}, \theta) = \tilde{y}_{mis} + \theta$

Mappings between SA & AA: $y_{mis} = \tilde{y}_{mis} + \theta$ (or $\tilde{y}_{mis} = y_{mis} - \theta$)

"E-step in SA":

$$y_{mis}^{(t+0.5)} = \mathbb{E}[\mathbb{E}[y_{mis}|y_{obs}, \theta^{(0.5)}, \tilde{y}_{mis}]|y_{obs}, \theta^{(t)}]$$

where the expectation is with respect to $f(\tilde{y}_{mis}, \theta^{(t+0.5)}) = p(y_{mis}|y_{obs}, \theta^{(t+0.5)}, \tilde{y}_{mis})$. Hence

$$\begin{aligned} y_{mis}^{(t+0.5)} &= \mathbb{E}[\tilde{y}_{mis} + \theta^{(t+0.5)}|y_{obs}, \theta^{(t)}] \\ &= \theta^{(t+0.5)} + \underbrace{\mathbb{E}[\tilde{y}_{mis}|y_{obs}, \theta^{(t)}]}_{\text{E-step in AA}} \end{aligned}$$

"M-step in SA":

$$\begin{aligned} &\theta^{(t+1)} = y_{mis}^{(t+0.5)} = \theta^{(t+0.5)} + \tilde{y}_{mis}^{(t)} = y_{obs} - \tilde{y}_{mis}^{(t)} + \tilde{y}_{mis}^{(t)} = y_{obs} \\ \implies &\theta^{(t+1)} = y_{obs} \qquad \text{converge in one iteration!} \end{aligned}$$

We can formalize this as follows:

Define $Q_I = \mathbb{E}_{A2}[\mathbb{E}_{A1}[\log p_{A2}(y_{obs}, y_{mis}|\theta)|y_{obs}, \tilde{y}_{mis}, \theta = G_{A2}(\theta^{(t)})]|y_{obs}, \theta^{(t)}]$.

Set $\theta^{(t+1)} = \arg\min Q_I(\theta|\theta^{(t)}$ where $A1$ is an augmentation scheme with missing data y_{mis}, $A2$ and $G_{A2}(\theta^{(t)}$ is the value from running one iteration of EM in the $A2$ regime.

The algorithm can be summarized as follows:

1. Run one iteration of $A2$-EM to obtain $G_{A2}(\theta^{(t)})$
2. Write down Q-function of the $A1$-EM
3. Replace y_{mis} with $y_{mis} = H(\tilde{y}_{mis}, \theta^{(t+0.5)})$
4. Now the Q function has expectations w. r. t. $\tilde{y}_{mis}$, so compute them (i.e. E-step in $A2$-EM

5. Find maximizer.

Example: $A2 = AA$ and $A1 = SA$

$$p_{SA}(y_{obs}, y_{mis}|\theta) = p(y_{obs}|y_{mis})p(y_{mis}|\theta)$$
$$\implies \log p_{SA}(y_{obs}, y_{mis}|\theta) = -\frac{1}{2}(y_{obs} - y_{mis})^2 - \frac{1}{2v}(y_{mis} - \theta)^2$$

$$\begin{aligned} Q_I(\theta|\theta^{(t)}) &= \mathbb{E}_{AA}[\mathbb{E}_{SA}[\log p_{SA}(y_{obs}, y_{mis}|\theta)|y_{obs}, \tilde{y}_{mis}, \theta = G_{AA}(\theta^{(t)})]|y_{obs}, \theta^{(t)}] \\ &= \mathbb{E}_{AA}[\mathbb{E}_{SA}[-\frac{1}{2}(y_{mis} - \theta)^2|y_{obs}, \tilde{y}_{mis}, \theta = G_{AA}(\theta^{(t)})]|y_{obs}, \theta^{(t)}] + \text{some constant (not dependent on } \theta) \end{aligned}$$

$$\begin{aligned} \theta^{(t+1)} &= \mathbb{E}_{AA}[\mathbb{E}_{SA}[y_{mis}|y_{obs}, \tilde{y}_{mis}, G_{AA}(\theta^{(t)})]|y_{obs}, \theta^{(t)}] \\ &= \mathbb{E}_{AA}[\tilde{y}_{mis} + G_{AA}(\theta^{(t)})|y_{obs}, \theta^{(t)}] \\ &= G_{AA}(\theta^{(t)}) + \mathbb{E}_{AA}[\tilde{y}_{mis}|y_{obs}, \theta^{(t)}] \\ &= \frac{\theta^{(t)}v + y_{obs}}{v+1} + \mathbb{E}_{AA}[\tilde{y}_{mis}|y_{obs}, \theta^{(t)}] \end{aligned}$$

For AA we have:

$$p(y_{obs}, \tilde{y}_{mis}|\theta) \propto \exp\{-\frac{1}{2}(y_{obs} - \tilde{y}_{mis} - \theta)^2 - \frac{1}{2v}\tilde{y}_{mis}\}$$
$$\implies p(\tilde{y}_{mis}|y_{obs}, \theta) \propto \exp\{-\frac{1}{2}\tilde{y}_{mis}^2(1 + \frac{1}{v}) + \tilde{y}_{mis}(y_{obs} - \theta)\}$$

and

$$\tilde{y}_{mis}|y_{obs}, \theta^{(t)} \sim N((1 + \frac{1}{v})^{-1}(y_{mis} - \theta^{(t)}), (1 + \frac{1}{v})^{-1}) = N(\frac{v}{v+1}(y_{mis} - \theta^{(t)}), \frac{v}{v+1})$$

So $\theta^{(t+1)} = (\frac{v}{v+1})\theta^{(t)} + (\frac{1}{v+1})y_{obs} + (\frac{v}{v+1})y_{obs} - (\frac{v}{v+1})\theta^{(t)} = y_{obs}$

Notes about the interwoven EM algorithm (IEM)

- Generally requires no more computation (often less) than the two separate EM's.
- Convergence rate is generally much better than the best convergence rate of the two EM's. [Key: minimize "correlation" between the two schemes; using an SA & AA turns out to be a great way to do this.]
- IEM preserves monotone convergence and all convergence properties of EM.

How to construct SA/AA pairs?

Hierarchical models are usually written as SA's

Example:

$y_i|\lambda_i \sim Pois(\lambda)$

$\lambda_i|\alpha, \beta \sim Gamma(\alpha, \beta)$

In this case to form $(\hat{\alpha}, \hat{\beta})$ with $y_{obs} = \bar{y}, \bar{y}_{mis} = \bar{\lambda}, \theta = (\alpha, \beta)$ is an SA.

How to construct an AA?

Transform $\tilde{y}_{mis} = H(y_{mis}, \theta)$ so that $\tilde{y}_{mis}$ doesn't depend on θ.

$y_{mis}|\theta \sim N(\theta, v)$

$H^{-1}(y_{mis}, \theta) \implies \tilde{y}_{mis} = (y_{mis} - \theta)/v^{1/2} \implies \tilde{y}_{mis} \sim N(0, 1)$

One recipe to obtain AA's for location-scale families is to recenter and rescale. What if we don't have a location-scale family? Apply CDF transformation! (trickier for multivariate settings)

Set $\tilde{y}_{mis} = F(\lambda, \alpha, \beta)$

$y_{obs}|\tilde{y}_{mis}, \alpha, \beta \sim Pois(F^{-1}(\tilde{y}_{mis}, \alpha, \beta))$