STA 250 Lecture 15: EM IV – Efficient algorithms

Yuanzhe(Roger) Li

November 20th, 2013

Logistics

- More guidelines on final project have been posted.
- Homework 3 will be posted this evening and is due on class next Wednesday.

Recap

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

Today

Example (V known)
Sufficient augmentation (SA)

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

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

SAEM: $\theta^{(t+1)} = \dfrac{\theta^{(t)} + V y_{obs}}{V+1}$, The rate of convergence: $\dfrac{1}{V+1}$

Ancillary augmentation (AA)

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

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

AAEM: $\theta^{(t+1)} = \dfrac{\theta^{(t)} V + y_{obs}}{V+1}$, The rate of convergence: $\dfrac{V}{V+1}$

So the two algorithms have "opposite" performance as V changes.
If V is unkonwn we can derive EM's for the SA & AA and they 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 just see which converges faster.
One idea could be to "alternate" updates according to the SA&AA, i.e., compute:

$\theta^{(t+0.5)} = \dfrac{\theta^{(t)} + V y_{obs}}{V+1}$ (SA)

$\theta^{(t+1)} = \dfrac{\theta^{(t+0.5)} V + y_{obs}}{V+1}$ (AA)

i.e., $\theta^{(t+1)} = M_{AA}(M_{SA}(\theta^{(t)}))$.

Pros

$\longrightarrow$ Avoids need to select one of the algorithms

Cons

$\longrightarrow$ Do no better than the best of the two algorithms, no worse than the worst of the two algorithms.

$\longrightarrow$ Need to implement two algorithms.

[Note: computation time of thetwo algorithms may not be equal.]

Interwoven EM (IEM)

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

Consider:

E-step in AA: $\tilde{y}_{mis}^{(t)} = \mathbb{E}[\tilde{y}_{mis}|y_{obs}, \theta^{(t)}]$

M-step in AA: $\theta^{(t+0.5)} = y_{obs} - \tilde{y}_{mis}^{(t)} (= \dfrac{\theta^{(t)}V + y_{obs}}{V+1})$

$y_{mis} = H(\tilde{y}_{mis}, \theta) = \tilde{y}_{mis} + \theta$

Mappings betwenn SA & AA.

$y_{mis} = \tilde{y}_{mis} + \theta, \quad \tilde{y}_{mis} = y_{mis} - \theta$

E-step in SA:

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

Expectation w.r.t. $p(y_{mis}|y_{obs}, \theta^{(t+0.5)}, \tilde{y}_{mis}) = f(\tilde{y}_{mis}, \theta^{(t+0.5)})$

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

M-step in SA:

$$\begin{aligned} \theta^{(t+1)} &= y_{mis}^{(t+0.5)} \\ &= \theta^{(t+0.5)} + \tilde{y}_{mis}^{(t)} \\ &= y_{obs} - \tilde{y}_{mis}^{(t)} + \tilde{y}_{mis}^{(t)} \\ \Longrightarrow \theta^{(t+1)} &= y_{obs} \end{aligned}$$

i.e., converges in one iteration! ☺

We can formalize this as follows: Define

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

Then set

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q_I(\theta|\theta^{(t)})$$

Where $A1$ is the augmentation scheme with missing data y_{mis}, $A2$ is the augmentation scheme with missing data $\tilde{y}_{mis}$.
And $G_{A2}(\theta^{(t)})$ is the value from running one iteration of EM in the A_2 regime.

The algorithm can be summarized as follows:

1. Run one iteration of A2-EM to obtain $\theta^{(t+0.5)} = G_{A2}(\theta^{(t)})$;
2. Write down Q-function of the A1-EM;
3. Third item 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

This can be formalized using the Q-function:

$$Q_I(\theta|\theta^{(t)}) = \mathbb{E}_{AA}[\mathbb{E}_{SA}[\log P_{SA}(y_{obs}, y_{mis}|\theta)|y_{obs}, \tilde{y}_{mis}, \theta = G_{AA}(\theta^{(t)})]|y_{obs}, \theta^{(t)}]$$

Example

$$P_{SA}(y_{obs}, y_{mis}|\theta) = P(y_{obs}|y_{mis})P(y_{mis}|\theta)$$
$$\implies \log P_{SA}(y_{obs}, y_{mis}|\theta) = -\frac{1}{2}(y_{obs} - y_{mis})^2 - \frac{1}{2V}(y_{mis} - \theta)^2$$

$$Q_I(\theta|\theta^{(t)}) = \mathbb{E}_{AA}[\mathbb{E}_{SA}[-\frac{1}{2}(y_{obs} - y_{mis})^2|y_{obs}, \tilde{y}_{mis},$$
$$\theta = G_{AA}(\theta^{(t)})]|y_{obs}, \theta^{(t)}] + \text{constant (not depending on } \theta)$$

$$\implies \theta^{(t)} = \mathbb{E}_{AA}[\mathbb{E}_{SA}[y_{mis}|y_{obs}, \tilde{y}_{mis}, G_{AA}\theta^{(t)}]|y_{obs}, \theta^{(t)})]$$
$$= \mathbb{E}_{AA}[\tilde{y}_{mis} + G_{AA}(\theta^{(t)})|y_{obs}, \theta^{(t)})]$$
$$= G_{AA}(\theta^{(t)}) + \mathbb{E}_{AA}[\tilde{y}_{miss}|y_{obs}, \theta^{(t)}]$$

$$\implies \theta^{(t+1)} = \frac{\theta^{(t)}V + y_{obs}}{V+1} + \mathbb{E}_{AA}[\tilde{y}_{miss}|y_{obs}, \theta^{(t)}]$$

For AA:

$$P(y_{obs}, \tilde{y}_{miss}|\theta) \propto \exp\{-\frac{1}{2}(y_{obs} - \tilde{y}_{miss} - \theta)^2 - \frac{1}{2V}\tilde{y}_{miss}\}$$
$$\implies P(y_{obs}, \tilde{y}_{miss}|\theta) \propto \exp\{-\frac{1}{2}\tilde{y}_{miss}^2(1 + \frac{1}{V}) + \tilde{y}_{miss}(y_{obs} - \theta)\}$$
$$\implies y_{obs}, \tilde{y}_{miss}|\theta^{(t)} \sim N((1 + \frac{1}{V})^{-1}(y_{obs} - \theta^{(t)}), (1 + \frac{1}{V})^{-1})$$
$$\text{i.e. } y_{obs}, \tilde{y}_{miss}|\theta^{(t)} \sim N((\frac{V}{V+1})(y_{obs} - \theta^{(t)}), \frac{V}{V+1})$$
$$\implies \mathbb{E}[\tilde{y}_{miss}|y_{obs}, \theta^{(t)}] = \frac{V}{V+1}(y_{obs} - \theta^{(t)})$$

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 IEM algorithm:

- Generally requires no more computation (and often less, when the mapping between the two augmentations is deterministic) than the two EMs.
- 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.
Exampls
$Y_i|\lambda_i \sim Pois(\lambda_i)$
$\lambda_i|\alpha, \beta \sim Gamma(\alpha, \beta)$
In this case, to find the maximizer of (α, β), i.e. $(\widehat{\alpha}, \widehat{\beta})$, with
$y_{obs} = \vec{y}, y_{mis} = \vec{\lambda}, \theta = (\alpha, \beta)$. This is an SA.

How to construct an AA?
Transform $\tilde{y_{mis}} = H^{-1}(y_{mis}, \theta)$ so that $\tilde{y_{mis}}$ doesn't depend on θ.
Example

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

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

One recipe to obtain AA's of location-scale family is to recenter and rescale
- What if we don't have a location-scale family?
Apply CDF transform! (for homework) gives an ancillary guaranteed, i.e.
$F_X(X) \sim \text{Unif}[0, 1]$ if X is univariate
CDF Transform
Set $\tilde{y}_{miss} = F(\lambda; \alpha, \beta) \sim Unif(0, 1)$ where $F(x; a, b)$ is the CDF
corresponding to parameters a and b evaluated at x.
Then
$y_{obs}|\tilde{y}_{miss}, \alpha, \beta \sim Pois(F^{-1}(\tilde{y}_{miss}; \alpha, \beta))$ where F^{-1} is the inverse CDF.