STA 250 *Adv Statistical Computing*

Lecture Notes Wednesday 2013/11/20

Annoucements:

- Final Projects - 10 pages max, suggested topics posted on the course website, extensions of our homework assignments, implementations of new algorithms, deepen skills learned such as Hadoop or Hive, etc.
- Homework #3 has will be posted and due 1 week later on 11/27.

EM Algo:

- Last time we saw strategies to deal with complicated EM applications (i.e. E-Step and/or M-Step are difficult)
- Today we see how to speed up EM
- "Sufficient Augmentation" = "SA"
  $Y_{obs}|Y_{mis} \sim N(Y_{mis}, 1)$
  $Y_{mis}|\theta \sim N(\theta, v)$
  EM: $\theta^{(t+1)} = \frac{\theta^{(t)} + vY_{obs}}{v+1}$
  Rate of Convergence is $\frac{1}{v+1}$ which is fast if v is large

  We can reparameterize as:
  "Ancillary Augmentation" = "AA"
  $Y_{obs}|\widetilde{Y}_{mis}, \theta \sim N(\widetilde{Y}_{mis} + \theta, 1)$
  $\widetilde{Y}_{mis} \sim N(0, v)$
  EM: $\theta^{(t+1)} = \frac{\theta^{(t)}v + Y_{obs}}{v+1}$
  Rate of Convergence is $\frac{v}{v+1}$ which is fast if v is small

  Note: smaller rate of convergence is faster
  Note: "SA" and "AA" have "opposite" performance in convergence speed
  Note: If v is unknown, we can still derive EMs for SA and AA and they have similar performance to when v is known.

  Question: How do we pick which algorithm version to run, AA or SA, if v is unknown?
  - Code both, run a few iterations of each, and see which appears to converge faster
  - Try AEM or IEM (see below)
- Alternating EM - AEM
  - EM has monotone convergence so use both algorithms, alternate between each algorithm for two "half steps" to do one update/iteration
  $\theta^{(t+0.5)} = \frac{\theta^{(t)} + vY_{obs}}{v+1}$ "SA"
  $\theta^{(t+1)} = \frac{\theta^{(t+0.5)}v + Y_{obs}}{v+1}$ "AA"
  $\theta^{(t+1)} = M_{AA}(M_{SA}(\theta^{(t)}))$ in terms of update mappings

  or could be the other order of AA and SA and will get different updates in general (but equivalent rates of convergence).

  Pros/Cons
  - If v is extreme (small or large), then one of SA and AA converges very slowly and the other very quickly so basically one step gets nowhere and is wasted and the other step is doing all the work
  - Update step could be very difficult/expensive to compute

- Performance/convergence is somewhere between the worst convergence and the best convergence of the two algorithms
- Computation of each step may or may not have the same computation time (i.e. one may be in closed form and the other requires MC)

- <u>Interwoven EM - IEM</u>
  Turns out there is a way to combine both algorithms into a single, improved update using their joint info so consider:
  AA E-step $\widetilde{Y}_{mis}^{(t)} = \mathbb{E}\left[\widetilde{Y}_{mis}|Y_{obs}, \theta^{(t)}\right]$
  AA M-step $\theta^{(t+0.5)} = Y_{obs} - \widetilde{Y}_{mis}^{(t)} = \frac{\theta^{(t)}v+Y_{obs}}{v+1}$

  Recall:
  "Sufficient Augmentation" = "SA"
  $Y_{obs}|Y_{mis} \sim N(Y_{mis}, 1)$
  $Y_{mis}|\theta \sim N(\theta, v)$

  "Ancillary Augmentation" = "AA"
  $Y_{obs}|\widetilde{Y}_{mis}, \theta \sim N(\widetilde{Y}_{mis} + \theta, 1)$
  $\widetilde{Y}_{mis} \sim N(0, v)$

  Then $Y_{mis} = H(\widetilde{Y}_{mis}, \theta) = \widetilde{Y}_{mis} + \theta \implies \widetilde{Y}_{mis} = Y_{mis} - \theta$
  i.e. map between SA and AA

  SA E-step:

$$
\begin{aligned}
Y_{mis}^{(t+0.5)} &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[Y_{mis}|Y_{obs}, \theta^{(t+0.5)}, \widetilde{Y}_{mis}\right]}_{\textbf{w.r.t. } P(Y_{mis}|Y_{obs}, \theta^{(t+0.5)}, \widetilde{Y}_{mis})} \middle| Y_{obs}, \theta^{(t)}\right] \\
&= \mathbb{E}\left[\widetilde{Y}_{mis} + \theta^{(t+0.5)} \middle| Y_{obs}, \theta^{(t)}\right] \\
&= \theta^{(t+0.5)} + \underbrace{\mathbb{E}\left[\widetilde{Y}_{mis}|Y_{obs}, \theta^{(t)}\right]}_{\textbf{E-step in AA}}
\end{aligned}
$$

  SA M-step:

$$
\begin{aligned}
\theta^{(t+1)} &= Y_{mis}^{(t+0.5)} \\
&= \theta^{(t+0.5)} + \widetilde{Y}_{mis}^{(t)} \\
&= Y_{obs} - \widetilde{Y}_{mis}^{(t)} - \widetilde{Y}_{mis}^{(t)} \\
&= Y_{obs}
\end{aligned}
$$

  So $\theta^{(t+1)} = Y_{obs} \implies$ Convergence in 1-iteration to the MLE $= Y_{obs}$ using the joint info

- <u>Formalize as follows:</u>
  Define $Q_I(\theta|\theta^{(t)}) = \mathbb{E}_{A2}\left[\mathbb{E}_{A1}\left[logP_{A1}\left(Y_{obs}, Y_{mis}|\theta\right) \middle| Y_{obs}, \widetilde{Y}_{mis}, \theta = G_{A2}\left(\theta^{(t)}\right)\right] \middle| Y_{obs}, \theta^{(t)}\right]$
  $\theta^{(t+1)} = \arg\max_\theta Q_I\left(\theta|\theta^{(t)}\right)$

  A1 is the first augmentation scheme with missing data Y_{mis}
  A2 is the second augmentation scheme with missing data $\widetilde{Y}_{mis}$
  $G_{A2}\left(\theta^{(t)}\right)$ is the value from one iteration of EM in A2 scheme

  The Interwoven EM Algorithm (IEM):

1) Run 1 iteration in A2 scheme to plug into Q and get $G_{A2}\left(\theta^{(t)}\right) = \theta^{(t+0.5)}$
2) Write down Q function of A1 EM
3) Replace $Y_{mis} = H(\widetilde{Y}_{mis}, \theta^{(t+0.5)})$
4) Now Q fct has expecation with respect to $\widetilde{Y}_{mis}$ so compute it, i.e. do A2 E-step
5) Find maximizer

Example A2 = AA, A1 = SA

$$Q_I\left(\theta|\theta^{(t)}\right) = \mathbb{E}_{AA}\left[\mathbb{E}_{SA}\left[logP_{SA}\left(Y_{obs}, Y_{mis}|\theta\right)\left|Y_{obs}, \widetilde{Y}_{mis}, \theta = G_{AA}\left(\theta^{(t)}\right)\right.\right]\left|\, Y_{obs}, \theta^{(t)}\right.\right]$$

$$P_{SA}\left(Y_{obs}, Y_{mis}|\theta\right) = P(Y_{obs}|Y_{mis})P(Y_{mis}|\theta)$$

$$\Longrightarrow \quad log\, P_{SA}\left(Y_{obs}, Y_{mis}|\theta\right) = -\frac{1}{2}(Y_{obs} - Y_{mis})^2 - \frac{1}{2v}(Y_{mis} - \theta)^2$$

$$Q_I\left(\theta|\theta^{(t)}\right) = \mathbb{E}_{AA}\left[\mathbb{E}_{SA}\left[-\frac{1}{2v}(Y_{mis} - \theta)^2\left|Y_{obs}, \widetilde{Y}_{mis}, \theta = G_{AA}\left(\theta^{(t)}\right)\right.\right]\left|\, Y_{obs}, \theta^{(t)}\right.\right] + constant$$

$$\Rightarrow \qquad \text{Maximized at } \theta = \mathbb{E}\left[Y_{mis}\right].$$

$$\begin{aligned}
\theta^{(t+1)} &= \mathbb{E}_{AA}\left[\mathbb{E}_{SA}\left[Y_{mis}\left|Y_{obs}, \widetilde{Y}_{mis}, G_{AA}\left(\theta^{(t)}\right)\right.\right]\left|\, Y_{obs}, \theta^{(t)}\right.\right] \\
&= \mathbb{E}_{AA}\left[\widetilde{Y}_{mis} + G_{AA}\left(\theta^{(t)}\right)\left|\, Y_{obs}, \theta^{(t)}\right.\right] \qquad \text{(Mapping from SA to AA)} \\
&= G_{AA}\left(\theta^{(t)}\right) + \mathbb{E}_{AA}\left[\widetilde{Y}_{mis}\left|\, Y_{obs}, \theta^{(t)}\right.\right] \\
&= \frac{\theta^{(t)}v + Y_{obs}}{v+1} + \underbrace{\mathbb{E}_{AA}\left[\widetilde{Y}_{mis}\left|\, Y_{obs}, \theta^{(t)}\right.\right]}_{\mathbb{E}\text{ wrt. } P(Y_{obs}, \widetilde{Y}_{mis}|\theta)}
\end{aligned}$$

$$P(Y_{obs}, \widetilde{Y}_{mis}|\theta) \propto \exp\left\{-\frac{1}{2}(Y_{obs} - \widetilde{Y}_{mis} - \theta)^2 - \frac{1}{2v}\widetilde{Y}_{mis}\right\}$$

$$\Longrightarrow \quad P(Y_{obs}, \widetilde{Y}_{mis}|\theta) \propto \exp\left\{-\frac{1}{2}\widetilde{Y}^2_{mis}\left(1 + \frac{1}{v}\right) + \widetilde{Y}_{mis}(Y_{obs} - \theta)\right\}$$

$$\widetilde{Y}_{mis}|Y_{obs}, \theta^{(t)} \sim N\left(\left(1 + \frac{1}{v}\right)^{-1}\left(Y_{obs} - \theta^{(t)}\right), \left(1 + \frac{1}{v}\right)^{-1}\right)$$

$$\widetilde{Y}_{mis}|Y_{obs}, \theta^{(t)} \sim N\left(\frac{v}{v+1}\left(Y_{obs} - \theta^{(t)}\right), \frac{v}{v+1}\right)$$

$$\Longrightarrow \quad \mathbb{E}_{AA}\left[\widetilde{Y}_{mis}\left|\, Y_{obs}, \theta^{(t)}\right.\right] = \frac{v}{v+1}\left(Y_{obs} - \theta^{(t)}\right)$$

$$\begin{aligned}
&= \frac{\theta^{(t)}v + Y_{obs}}{v+1} + \frac{v}{v+1}\left(Y_{obs} - \theta^{(t)}\right) \\
&= Y_{obs}
\end{aligned}$$

- Notes about IEM
  - Generally requires no more computation (and often less, when the mapping between the two augmentations is deterministic) than the two EMs
    - AIM has 4 steps
    - IEM has 3 steps (the fourth is just a deterministic mapping between the two schemes)
  - Preserves monotone convergence and all main convergence properties of EM
  - Convergence rate is generally faster than either of the 2 individual schemes (proofs posted on website)
    - Often ven better than the better of the two individual schemes
    - Key is to minimize the “correlation” between 2 schemes i.e. use SA and AA to get fast convergence

How to construct SA/AA pairs?
Hierarchical models are naturally written as SA
Homework Problem:

$Y_i|\lambda_i \sim Pois(\lambda_i)$
$\lambda_i|\alpha,\beta \sim Gamma(\alpha,\beta)$
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 - separates the layers

How to construct AA?
Transform $\widetilde{Y}_{mis} = H^{-1}(Y_{mis},\theta)$ to remove dependence on $\theta = (\alpha,\beta)$
$\widetilde{Y}_{mis}|\theta \sim N(\theta,v)$
$\widetilde{Y}_{mis} = Y_{mis} - \theta$ where $\widetilde{Y}_{mis} \sim N(0,v)$
$H^{-1}(Y_{mis},\theta)$ so use $\widetilde{Y}_{mis} = Y_{mis} - \theta$ to get $\widetilde{Y}_{mis} \sim N(0,v)$ If v were also a parameter, then transform is $\widetilde{Y}_{mis} = (Y_{mis} - \theta)/v^{1/2}$ to get $\widetilde{Y}_{mis} \sim N(0,1)$

- One recipe for AA of location-scale family is to recenter and rescale as above.
- If not location-scale family, then CDF transform gives an ancillary guaranteed, i.e. $F_X(X) \sim$ Unif[0, 1] if X is univariate

CDF Transform
Set $\widetilde{Y}_{mis} = F(\lambda;\alpha,\beta) \sim Unif(0,1)$ where $F(x;a,b)$ is the CDF corresponding to parameters a and b evaluated at x.
$Y_{obs}|\widetilde{Y}_{mis},\alpha,\beta \sim Pois(F^{-1}(\widetilde{Y}_{mis};\alpha,\beta))$ where F^{-1} is the inverse CDF, i.e. quantile function.