Lecture note 12

by Qian Li

Nov. 7th, 2013

Part I

Why we use EM Module

To fit any non-standard statistical model, we need to use numerical techniques.
For Bayes $\Rightarrow$use MCMC methods.
For Maximum Likelihood $\Rightarrow$need to maximize a non-standard function.
This module is all about maximizing "difficult" likelihoods (or posteriors).
Before we going to that, just do some common optimization algorithms:
• Bisection • Newton-Raphson • Scoring

Part II

Bisection

Note: We will actually look at root finding algorithms, i.e. finding x such that g(x) = 0. To maximize f (f is continuous), we can solve g(x) = f ′(x) = 0.
For one-dimensional functions(continuous)
Let g : $\boldsymbol{R} \rightarrow \boldsymbol{R}$ be a continuous function on [a, b], we want to find x_*s.t. g(x_*) = 0.
Idea:

1. Find l & u s.t. g(l)g(u) < 0 (one is positive, the other is negative) .
2. Set c =(l+u) / 2 , compute g(c).
3. If g(c) = 0, done. If |g(c)| $<\varepsilon$ for some small one , done.
4. O/W, ex. g(c) = t, reset l and u.
5. O/W, if g(l)g(c) < 0, set u = c, else set l = c.
6. Repeat

Pro and con:

1. Pro: Easy to code + understand ; continuity; not differentiability
2. Con: could be multiple roots ; only 1D; Doesn't use information about function beyond sign.

Part III
Newton-Raphson

An iterative algorithm to solve for g(x) = 0.

Idea :
Update x_t to x_{t+1} where $x_{t+1} = x_t + \eta_t$
How to select η_t?
$g(x_{t+1}) = g(x_t + \eta_t) \approx g(x_t) + \eta_t^{'} g(x_t) + \theta(\eta_t^2)$
If we get $g(x_{t+1}) + \eta_t g^{'}(x_t) = 0 \Rightarrow \eta_t = -\frac{g(x_t)}{g^{'}(x_t)}$
Algorithm:

1. Pick x_0,set t = 0
2. Update, $x_{t+1} = x_t - \frac{g(x_t)}{g^{'}(x_t)}$
3. If $|g(x_{t+1})| < \varepsilon$, stop, else increment t to t+1
4. Repeat.

Pro and con:

1. Pro: Typically fast (quadratic convergence) ; Works in multiple dimensions.
2. Con: Sensitive to choice of x_0 (if start at wrong place); Could exists multiple roots ;Need derivatives; only need one derivative (or two); depend on root finding.

Part IV
Scoring

This is a small modication of the Newton-Raphson method, specically for maximizing likelihoods.

Newton-Raphson method $\theta_{t+1} = \theta_t - [l^{''}(\theta_t)]^{-1} l^{'}(\theta_t)$
Scoring $\theta_{t+1} = \theta_t + I^{-1}(\theta_t) l^{'}(\theta_t)$
where $I(\theta) = E(-l^{''}(\theta)) \Leftarrow$ *expected* fisher information.

We may prefer scoring is the expected information is easier to conjute than l'' (e.g. in exponential families).

Scoring coverges linearly.

Part V
EM Algorithm

- For many problems, the likelihood itself can be dicult to compute. eg:

$\eta_{ij} = x_{ij}^T\beta + z_{ij}^T\gamma_i$

$y_{ij}|\beta, \gamma_i \sim Bin(n_{ij}, g^{-1}(\eta_{ij}))$

$\gamma_i \sim i.i.d.N(0, \Sigma)$ where Data:$\{y_{ij}\}$; Parameters:$\{\beta, \Sigma\}$; Latent variable: $\{r_i\}$

Then we have :

$P(\overrightarrow{y}|\beta, \Sigma) = \int P(\overrightarrow{y}, \{\gamma\}|\beta, \Sigma)d_\gamma$

$= \int \prod_{i,j}\left(\begin{array}{c} n_{ij} \\ y_{ij} \end{array}\right)[g^{-1}(\eta_{ij})^{y_{ij}}[1-g^{-1}]^{n_{ij}-y_{ij}}] * \prod_i (2\pi)^{p/2}|\Sigma|^{-1/2}exp\{-\frac{1}{2}\gamma_i^T\Sigma^{-1}\gamma_i\}d_r$

Our likelihood involves integrals that are dicult to compute. Hard to use Bisection or Newton-Raphson. Using EM, we can avoid directly computing the integrals. Suppose we have a model with parameter θ, observed data y_{obs}, and missing data y_{mis} to maximize:

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

we can use the EM algorithm:

$Q(\theta|\theta^t)$

$= E[logP(Y_{obs}, Y_{mis}|\theta)|Y_{obs}, \theta^t]$

$= \int [logP(Y_{obs}, Y_{mis}|\theta)]P(Y_{mis}|Y_{obs}, \theta^t)d_{Y_{mis}}$

Algorithm:

1. Select θ^0, set t=0.
2. Set $\theta^{t+1} = argmaxQ(\theta|\theta^t)$
3. Check convergence. If $\frac{|\theta^{t+1}-\theta^t|}{|\theta^t|} < \varepsilon$, stop
4. Else, increment t to t + 1, repeat step 2-4 until converge.

Part VI
Example

Setting:

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

Goal:maximize $P(y_{obs}|\theta) = \int P(y_{obs}, y_{mis}|\theta)dy_{mis} = \int P(y_{obs}|y_{mis})P(Y_{mis}|\theta)dy_{mis}$

$$Q(\theta|\theta^t) = E[log\{P(y_{obs}|y_{mis}, \theta)P(y_{mis}|\theta)\}|y_{obs}, \theta^t]$$
$$= E[-\tfrac{1}{2}(y_{obs} - y_{mis})^2 - \tfrac{1}{2}log(2\pi) - \tfrac{1}{2V}(y_{mis} - \theta)^2 - \tfrac{1}{2}log(V) - \tfrac{1}{2\pi}|y_{obs}, \theta^t]$$

$$= E[-\tfrac{1}{2V}(y_{mis} - \theta)^2|y_{obs}, \theta^t]$$

we have ignored any term not involving .To compute this expectation, we need to know P($y_{mis}|y_{obs}, \theta^t$)

P($y_{mis}|y_{obs}, \theta^t$)$\propto$P($y_{mis}, y_{obs}|\theta^t$)

$$\Longrightarrow y_{mis}|y_{obs}, \theta^t \sim N\left(\frac{\frac{\theta^t}{V} + \frac{y_{obs}}{1}}{\frac{1}{V} + \frac{1}{1}}, \frac{1}{\frac{1}{V} + \frac{1}{1}}\right) \sim N(\tfrac{\theta^t + V y_{obs}}{V+1}, \tfrac{V}{V+1})$$

$$Q(\theta|\theta^t)$$
$$= E[-\tfrac{1}{2V}(y_{mis} - \theta)^2|y_{obs}, \theta^t]$$
$$= -\tfrac{1}{2V}E[y_{mis}^2 + \theta^2 - 2y_{mis}\theta|y_{obs}, \theta^t]$$
$$= -\tfrac{1}{2V}(\theta^2 - 2\theta E[y_{mis}|y_{obs}, \theta^t])$$
$$= -\tfrac{1}{2V}(\theta^2 - 2\theta\tfrac{\theta^t + V y_{obs}}{V+1}) + constant$$
$$\tfrac{\partial Q}{\partial \theta} = -\tfrac{1}{2V}(2\theta - 2(\tfrac{\theta^t + V y_{obs}}{V+1})) \Rightarrow \text{maximized at } \tfrac{\theta^t + V y_{obs}}{V+1}$$

Algorithm: $\theta^{t+1} = (\frac{1}{V+1}\theta^t + (\frac{V}{V+1})y_{obs})$

As $t \Rightarrow INF, \theta^{t+1} \Rightarrow y_{obs}$

Linear rate convergence : $(\frac{1}{V+1})$, lower rate is better.