STA 250 Lecture 12

Rex Cheung

November 7, 2013

<u>EM Module:</u>

To fit any non-standard statistical model, we need to use numerical techniques (i.e. Metropolis-Hastings, Gibbs, etc...).
For Bayes, we use MCMC methods.
For Maximum Likelihood, we often need to maximize a non-standard function.
This module is all about maximizing "difficult" likelihoods (or posteriors).

First, we will start with some common optimization algorithms:

- Bisection
- Newton-Raphson
- Scoring

<u>Note:</u> We will actually look at root finding algorithms, i.e. finding x such that $g(x) = 0$.
To maximize f (assuming f is continuous), we can solve $g(x) = f'(x) = 0$.

1 Bisection

This is used for one-dimensional continuous functions.
Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function on $[a, b]$, we want to find x_* such that $g(x_*) = 0$.

<u>The algorithm:</u>

1. Find l and u such that $g(l)g(u) < 0$ (by IVT, $\exists$ a root between l and u) .
2. Set $c = \frac{l+u}{2}$, compute $g(c)$.
3. If $|g(c)| < \epsilon$ for some small $\epsilon > 0$, stop.
4. Otherwise, if $g(l)g(c) < 0$, set $u = c$, else set $l = c$.
5. Repeat step 2-4.

Pros	Cons
Easy to code + understand Only requires continuity, not differentiability Linear convergence	There could be multiple roots Limits to 1D only Doesn't use information about fn. beyond side

2 Newton-Raphson

An iterative algorithm to solve for $g(x) = 0$.
Idea: Update x_t to x_{t+1}, where $x_{t+1} = x_t + \eta_t$.
Suppose $g : \mathbb{R} \to \mathbb{R}$, how do we choose η_t?

$$g(x_{t+1}) = g(x_t + \eta_t) \approx g(x_t) + \eta_t g'(x_t) + O(\eta_t^2)$$

So we can get $g(x_t) + \eta_t g'(x_t) = 0 \Rightarrow \eta_t = \frac{-g(x_t)}{g'(x_t)}$.
Algorithm:

- Pick x_0, set $t = 0$.
- Update $x_{t+1} = x_t - \frac{g(x_t)}{g'(x_t)}$.
- If $|g(x_{t+1})| < \epsilon$, stop. Else increment $t \to t + 1$.
- Repeat.

If $g : \mathbb{R}^m \to \mathbb{R}^m$, then the update is

$$\overrightarrow{x_{t+1}} = \overrightarrow{x_t} - [\nabla g(\overrightarrow{x_t})]^{-1} g(\overrightarrow{x_t})$$

To maximize $l(\theta)$, we want to solve $l'(\theta) = 0$, i.e.

$$\theta_{t+1} = \theta_t - [l''(\theta_t)]^{-1} l'(\theta_t)$$

Pros	Cons
Typically fast (quadratic convergence)	Sensitive to choice of x_0
Works in multiple dimensions	Could exists multiple roots
Only need one (or two) derivatives	Need derivatives

3 Scoring

This algorithm is a small modification of the Newton-Raphson algorithm that's specifically for maximizing likelihoods.

$$\begin{aligned} NR : \theta_{t+1} &= \theta_t - [l''(\theta_t)]^{-1} l'(\theta_t) \\ Scoring : \theta_{t+1} &= \theta_t + I^{-1}(\theta_t) l'(\theta_t) \end{aligned}$$

where $I(\theta) = E[-l''(\theta)]$ (the expected Fisher Information). We may prefer scoring if the expected information is easier to compute than $l''(.)$ (i.e. in exponential families). Scoring converges linearly.

4 Rate of convergence of a sequence

Let $x_1, x_2, \ldots$ be a sequence that converges to some value x_*, then we say that the sequence converges with quadratic rate if

$$\lim_{t \to \infty} \frac{|x_{t+1} - x_*|}{|x_t - x_*|^2} = c, \quad 0 < c < \infty$$

The sequence converges with linear rate if

$$\lim_{t\to\infty} \frac{|x_{t+1} - x_*|}{|x_t - x_*|} = c, \quad 0 < c < 1$$

If $c = 1$, it is called super-linear rate of convergence.

5 The EM Algorithm

For many problems the likelihood itself can be difficult to compute, for example (GLMM):

$$\begin{aligned} \eta_{ij} &= x_{ij}^T\beta + z_{ij}^T\gamma_i \\ y_{ij}|\beta,\gamma_i &\sim Bin(n_{ij}, g^{-1}(\eta_{ij})) \\ \gamma_i &\sim i.i.d. N(0,\Sigma) \end{aligned}$$

- Data: $\{y_{ij}\}$
- Parameters: $\{\beta, \Sigma\ \}$
- Latent Variables: $\{\gamma_i\}$

Suppose we want to find the MLE $\{\beta,\Sigma\}$, we have

$$\begin{aligned} P(\overrightarrow{y}|\beta,\Sigma) &= \int P(\overrightarrow{y}, \{\gamma\}|\beta,\Sigma)d\gamma \\ &= \int \prod_{i,j} \binom{n_{ij}}{y_{ij}} [g^{-1}(\eta_{ij})]^{y_{ij}} [1 - g^{-1}]^{n_{ij}-y_{ij}} \\ &\quad \times \prod_i (2\pi)^{p/2} |\Sigma|^{-1/2} exp\{-\frac{1}{2}\gamma_i^T \Sigma^{-1} \gamma_i\} d\gamma \\ &= \textit{NOTHING NICE!} \end{aligned}$$

Here our likelihood includes integrals that are difficult to compute. It's hard to use NR, or even bisection. However, if we use the EM Algorithm, it turns out we can avoid directly computing the integral!

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. Define:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E[logP(Y_{obs}, Y_{mis}|\theta)|Y_{obs}, \theta^{(t)}] \\ &= \int [logP(Y_{obs}, Y_{mis}|\theta)] P(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis} \end{aligned}$$

Algorithm:

1. Select $\theta^{(0)}$, set $t = 0$.
2. Set $\theta^{(t+1)} = argmax_\theta Q(\theta|\theta^{(t)})$.
3. If $\frac{|\theta^{(t+1)}-\theta^{(t)}|}{|\theta^{(t)}|} < \epsilon$, stop.
4. Increment $t \to t+1$. Repeat 2-4 until converge.

5.1 Example:

Setting:

$$\begin{aligned} y_{obs}|y_{mis} &\sim N(y_{mis}, 1) \\ y_{mis} &\sim N(\theta, V) \end{aligned}$$

Goal: Maximize $P(y_{obs}|\theta)$, where

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

So

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E[log\{P(y_{obs}|y_{mis}, \theta)P(y_{mis}|\theta)\}|y_{obs}, \theta^{(t)}] \\ &= E[-\frac{1}{2}(y_{obs} - y_{mis})^2 - \frac{1}{2}log(2\pi) - \frac{1}{2V}(y_{mis} - \theta)^2 - \frac{1}{2}log(V) - \frac{1}{2\pi}|y_{obs}, \theta^{(t)}] \\ &\Rightarrow E[-\frac{1}{2V}(y_{mis} - \theta)^2|y_{obs}, \theta^{(t)}] \quad (1) \end{aligned}$$

In (1), we ignore the terms that don't involve θ. To compute this, we need to know $P(y_{mis}|y_{obs}, \theta^{(t)})$. From Bayes, we know $P(y_{mis}|y_{obs}, \theta^{(t)}) \propto P(y_{mis}, y_{obs}|\theta^{(t)})$

$$\begin{aligned} \Rightarrow y_{mis}|y_{obs}, \theta^{(t)} &\sim N(\frac{\frac{\theta^{(t)}}{V} + \frac{y_{obs}}{1}}{\frac{1}{V} + \frac{1}{1}}, \frac{1}{\frac{1}{V} + \frac{1}{1}}) \\ &\sim N(\frac{\theta^{(t)} + Vy_{obs}}{V+1}, \frac{V}{V+1}) \end{aligned}$$

Therefore,

$$\begin{aligned} (1) &= -\frac{1}{2V}E[y_{mis}^2 + \theta^2 - 2y_{mis}\theta|y_{obs}, \theta^{(t)}] \\ &= -\frac{1}{2V}(\theta^2 - 2\theta E[y_{mis}|y_{obs}, \theta^{(t)}]) \\ &= -\frac{1}{2V}(\theta^2 - 2\theta(\frac{\theta^{(t)} + Vy_{obs}}{V+1})) \end{aligned}$$

All together, $Q(\theta|\theta^{(t)}) = -\frac{1}{2V}(\theta^2 - 2\theta(\frac{\theta^{(t)}+Vy_{obs}}{V+1}))$ + constant. Maximizing $Q(\theta|\theta^{(t)})$, we have

$$\frac{\partial Q}{\partial \theta} = -\frac{1}{2V}(2\theta - 2(\frac{\theta^{(t)} + Vy_{obs}}{V+1}))$$

$\Rightarrow$ maximized at $\frac{\theta^{(t)}+V y_{obs}}{V+1}$.
Algorithm/Update: $\theta^{(t+1)} = (\frac{1}{V+1})\theta^{(t)} + (\frac{V}{V+1})y_{obs}$.
Remarks:

- If V is big, then the solution converges faster because $\theta^{(t+1)}$ is closer to data.
- If V is small, then the solution converges slower because it's closer to $\theta^{(t)}$, so need to more steps.

As $t \to \infty$, $\theta^{(t+1)} \to y_{obs}$. Also, it is of linear rate convergence: $c = \frac{1}{V+1}$.