STA 250. Fall, 2013.

Lecture 12: Optimization + EM Lecture # 1.

Transcribed by Eliot Paisley. 11/6/13

Introduction:

- We'll spend only a short time on optimization ... this is really an EM module.
- To fit any non-standard statistical models (i.e., outside of just `lm`, `glm`, or `lme`), we need know a little bit about numerical methods. We've already seen one example, the Metropolis-Hastings algorithm.
- For Bayes problems we use Markov-Chain Monte Carlo (MCMC) methods, while for maximum likelihood (ML) problems we need to maximize a non-standard function. This entire module is about maximizing these 'difficult' likelihoods (or posteriors).

To begin, we start by looking at some common optimization algorithms; Bisection, Newton-Raphson, and Scoring.
Note: we're actually looking at root-finding algorithms. i.e. finding x such that $g(x) = 0$. To maximize f (if continuous) we can solve $g(x) = f'(x) = 0$.

Bisection:

- For 1-dimensional continuous functions.
- Let $g: \mathbb{R} \to \mathbb{R}$ be a continuous function on $[a, b]$. We want to find x_* such that $g(x_*) = 0$.
- Idea: Find l and u such that $g(l) \cdot g(u) < 0$, which implies that $g(l)$ and $g(u)$ will have different signs.
- Set $c = \frac{l+u}{2}$, and compute $g(c)$. If $g(l) \cdot g(c) < 0$, then set $u = c$. Otherwise, set $l = c$.
- Repeat the step above.
- Pros: Easy to code, and understand. Converges in linear time. We only need continuity, not differentiability.
- Cons: Could be multiple roots. Only works in 1-dimension, doesn't generalize nicely to higher dimensions. Doesn't use much information about other values. For example, if $g(l) = -0.1$, and $g(u) = 10000$, then we still select c to be in the middle.

Newton-Raphson

- This is 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 choose η_t is the question.
- We can write
$$g(x_{t+1}) = g(x_t + \eta_t) \approx g(x_t) + \eta_t g'(x_t) + \mathcal{O}(\eta_t^2)$$
and ignoring the higher-order terms, if we set $g(x) + \eta_t g'(x_t) = 0$, then we have
$$\eta_t = -\frac{g(x_t)}{g'(x_t)}$$
- Algorithm:
  - Pick x_0. Set $t = 0$.
  - Update $x_{t+1} = x_t - \frac{g(x_t)}{g'(x_t)}$.
  - If $|g(x_{t+1})| < \epsilon$, then stop. Otherwise, set $t \to t + 1$ and update again.
- Pros: Typically fast (quadratic convergence). Works in multiple dimensions. Only needs one (or two) derivatives.
- Cons: Sensitive to the choice of x_0. There could be multiple roots. We need to be able to calculate derivatives.
- If $g : \mathbb{R}^m \to \mathbb{R}^m$, then $\vec{x}_{t+1} = \vec{x}_t - [\nabla g(\vec{x}_t)]^{-1} g(\vec{x}_t)$
- To maximize $l(\theta)$, we want to solve $l'(\theta) = 0$, where $\theta_{t+1} = \theta_t - [l''(\theta_t)]^{-1} l'(\theta_t)$.

Rate of Convergence of a Sequence: Let $x_1, x_2, \dots,$ be a sequence that converges to some value x_*. Then we say that the sequence converges with quadratic rate if

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

Similarly, we say that a sequence converges with a linear rate if

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

where if $c = 1$ we say the sequence has a 'super linear' rate of convergence.
Question: Are there algorithms that converge in cubic time?
Answer: Yes, but only for specific types of problems.

Scoring

- This is a small modification of the Newton-Raphson method, specifically for maximizing likelihoods.
- In Newton-Raphson we had $\theta_{t+1} = \theta_t - [l''(\theta_t)]^{-1} l'(\theta_t)$, where in Scoring we use $\theta_{t+1} = \theta_t - I(\theta_t)^{-1}l'(\theta_t)$.
- $l''(\theta_t)$ is the *observed* Fisher information, while $I(\theta_t)^{-1} = E(-l''(\theta))$ is the *expected* Fisher information.
- Scoring is preferred to Newton-Raphson if the expected information is easier to compute than the observed (e.g. in exponential families).
- Scoring coverges linearly.

The EM Algorithm:

- For many problems, the likelihood itself can be difficult to compute. e.g.

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

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

$$\gamma_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma^{-1})$$

  where $\{y_i\}$ is the data, with parameters $\{\beta, \Sigma\}$, and latent variables $\{\gamma_i\}$.
  The likelihood for this model is then

$$\begin{aligned} p(\vec{y}|\beta, \Sigma) &= \int p(\vec{y}, \{\gamma\}|\beta, \Sigma)\, d\gamma \\ &= \int \prod_{i,j} \binom{n_{ij}}{y_{ij}} \left[g^{-1}(\eta_{ij})\right]^{y_{ij}} \left[1 - g^{-1}(\eta_{ij})\right]^{n_{ij}-y_{ij}} \cdot \prod_j (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\gamma_j^T\Sigma^T\gamma_j\right\}\ d\gamma_1, \dots, d\gamma_j \\ &= \text{nothing nice at all} \end{aligned}$$

- Overall, our likelihood involves integrals that are difficult to compute.
- For these situations it's hard to use Bisection or Newton-Raphson. Using EM we avoid directly computing the integrals.
- Suppose we have a model with parameter θ, observed data y_{obs}, and "missing" data y_{mis} to maximize.

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

  here, we can use the EM algorithm.
- Define $Q(\theta|\theta^{(t)}) = E\left[\log p(y_{obs}, y_{mis}|\theta)|y_{obs}, \theta^{(t)}\right] = \int \log p(y_{obs}, y_{mis}|\theta)p(y_{mis}|y_{obs}, \theta^{(t)})dy_{mis}$.
  Algorithm:
  - Select $\theta^{(0)}$, set $t = 0$.
  - Set $\theta^{(t+1)} = \underset{\theta}{argmax}\ \ Q(\theta|\theta^{(t)})$.
  - Check convergence. If $\frac{|\theta^{(t+1)}|-|\theta^{(t)}|}{|\theta^{(t)}|} < \epsilon$, then stop. Otherwise, increment $t \to t+1$ and go back to the previous step.

- Simple Example:

$$y_{obs}|y_{mis} \sim \mathcal{N}(y_{mis}, 1)$$
$$y_{mis} \sim \mathcal{N}(\theta, V), \quad V \text{ known.}$$

<u>Goal:</u> maximize $p(y_{obs}|\theta)$.

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

Here we have

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

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

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

Thus,

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

So,

$$\frac{dQ}{d\theta} = -\frac{1}{2V}\left(2\theta - 2\frac{\theta^{(t)} + Vy_{obs}}{V+1}\right)$$

and setting equal to 0, we arrive at

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

.

<u>Algorithm:</u>

$$\theta^{(t+1)} = \frac{1}{V+1}\theta^{(t)} + \frac{V}{V+1}y_{obs}$$
$$y_{obs}|y_{mis} \sim \mathcal{N}(y_{mis}, 1)$$
$$y_{mis} \sim \mathcal{N}(\theta, V)$$

and as $t \to \infty$, $\theta^{(t+1)} \to y_{obs}$.
This is a linear rate of convergence: $\frac{1}{V+1}$, with speed inversely proportional to the size of V. Note that low rates indicate fast convergence, rates close to 1 indicate slow convergence.