Non-stochastic Optimization

Background: likelihood inference

- let $x_1, \ldots, x_n$ be an iid sample from $f(x \mid \theta^*)$, where true parameter value θ^* is unknown

- the likelihood function is $L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$

- the maximum likelihood estimator (MLE) of θ is the maximizer of $L(\theta)$

- usually it is easier to work with the loglikelihood $l(\theta) = \log L(\theta)$

- typically maximization of $l(\theta)$ is done by solving $l'(\theta) = 0$
  - $l'(\theta)$ is called the score function

- for any θ, $E_\theta\{l'(\theta)\} = 0$

$$E_\theta\{l'(\theta)\, l'(\theta)^T\} = -E_\theta\{l''(\theta)\}$$

  where E_θ is expectation wrt $f(x \mid \theta)$

- Fisher information: $I(\theta) = E_\theta\{l'(\theta)\, l'(\theta)^T\}$

- observed Fisher information: $-l''(\theta)$
  - if $\dim(\theta) = 1$, $I(\theta)$ is a nonnegative number
  - if $\dim(\theta) > 1$, $I(\theta)$ is a nonnegative definite matrix

- Importance of $I(\theta)$: it sets the limit on how accurate an unbiased estimate of θ can be

- as $n \to \infty$, $\sqrt{n}(\hat{\theta}_{MLE} - \theta^*) \Rightarrow N_p(0, I(\theta^*)^{-1})$

Working with Derivatives

- suppose $g(\underset{\sim}{x})$ is a differentiable function, where $\underset{\sim}{x} = (x_1, \dots, x_n)$

- to find its maximum/minimum, one method is to solve the equation $g'(x) = 0$, where $g'(x) = \left(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_n}\right)^T$

- that is, maximization/minimization is equivalent to solving $f(x) = 0$ where $f = g'$

Univariate Case

Newton's method

- a fast approach to solve $f(x) = 0$

- steps: (i) start with an initial estimate x_0
  (ii) for $t = 0, 1, \dots,$ compute

$$x_{t+1} = x_t + h_t \quad \text{with } h_t = -\frac{f(x_t)}{f'(x_t)}$$

  (iii) continue until convergence

- also known as Newton-Raphson

- need to specify x_0

- if $f(x)=0$ has multiple roots, end result will depend on x_0

- iteration cannot continue if $f'(x_e)=0$

Why it works?

- let x^0 be true solution, $\bar{x}$ be an approximation of x^0

- Taylor expansion: $f(x) = f(\bar{x}) + (x-\bar{x}) f'(\bar{x}) + \frac{(x-\bar{x})^2}{2} f''(\breve{x})$

  where $\breve{x}$ lies between x and $\bar{x}$

- since $f(x^0)=0$, we have $0 = f(\bar{x}) + (x^0-\bar{x}) f'(\bar{x}) + \frac{(x^0-\bar{x})^2}{2} f''(\breve{x})$

- since x^0 and $\bar{x}$ are close, the last term can be ignored:

$$0 \approx f(\bar{x}) + (x^0-\bar{x}) f'(\bar{x}) \Rightarrow x^0 \approx \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}$$

Optimization with Newton

- can be applied to optimize g by applying to $f=g'$

- both g' (gradient) and g'' (Hessian) are needed

- many variants of Newton's method avoid the computation of g'', which can be difficult, especially for multivariate functions

Example: To maximize $g(x) = \frac{\log x}{1+x}$

- first find $f(x) = g'(x) = \frac{1+\frac{1}{x} - \log x}{(1+x)^2}$

$$f'(x) = g''(x) = \frac{-(3+4/x + 1/x^2 - 2\log x)}{(1+x)^3}$$

- therefore $h_t = \frac{(x_t+1)(1+\frac{1}{x_t} - \log x_t)}{3+\frac{4}{x_t} + \frac{1}{x_t^2} - 2\log x_t}$

- a simpler formula: note that solving $f(x)=0$ is the same as solving $1+\frac{1}{x} - \log x$.

- treat $1+\frac{1}{x} - \log x$ as a new f function

- then $h_t = x_t - \frac{x_t^2 \log x_t}{1+x_t} \Rightarrow x_{t+1} = 2x_t - \frac{x_t^2 \log x_t}{1+x_t}$ #

Example:

To maximize loglikelihood $l(\theta)$, $\theta_{t+1} = \theta_t - \frac{l'(\theta_t)}{l''(\theta_t)}$

- consider the model with shift $p(x|\theta) = p(x-\theta)$.

- given observations $x_1, \dots, x_n$ iid $\sim p(x|\theta)$,

$$l(\theta) = \sum_{i=1}^n \log p(x_i - \theta), \quad l'(\theta) = -\sum_{i=1}^n \frac{p'(x_i-\theta)}{p(x_i-\theta)},$$

$$l''(\theta) = \sum_{i=1}^n \frac{p''(x_i-\theta)}{p(x_i-\theta)} - \sum_{i=1}^n \left\{\frac{p'(x_i-\theta)}{p(x_i-\theta)}\right\}^2$$

- note that we update θ, not $x_1, \dots, x_n$

- In R, to <u>minimize</u> a function, one can use

z = nlminb (x0, g, gr.g, hess.g)

x0 : initial value

g : function being minimized

gr.g : gradient of g
hess.g : Hessian of g
} have to be analytically calculated

- one can also use

z = nlminb(x0, g, gr.g) or z = nlminb (x0, g),

where gr.g / hess.g will be numerically approximated

Secant Method

- approximating $f'(x_t)$ by $\dfrac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$, the Newton method becomes the secant method:

$$x_{t+1} = x_t - \frac{f(x_t)(x_t - x_{t-1})}{f(x_t) - f(x_{t-1})}$$

- need to specify x_0 and x_1

Fisher Scoring

- another variant of Newton's method
- specific for MLE
- replace the Hessian $l''(\theta)$ by its expectation; i.e., Fisher information $I(\theta)$

$$\theta_{t+1} = \theta_t + \frac{l'(\theta_t)}{I(\theta_t)}$$

- in practice, use Fisher scoring in the beginning to make rapid improvements, then Newton's method for refinement near the end

Example continue with the previous example on $p(x|\theta) = p(x-\theta)$

- to use Fisher scoring, need to compute $I(\theta) = -E_\theta(l''(\theta))$

$$\Rightarrow I(\theta) = -n E_\theta\left[\frac{p''(x-\theta)}{p(x-\theta)} - \left\{\frac{p'(x-\theta)}{p(x-\theta)}\right\}^2\right]$$

$$= -n \int \left[\frac{p''(x-\theta)}{p(x-\theta)} - \left\{\frac{p'(x-\theta)}{p(x-\theta)}\right\}^2\right] p(x-\theta)\, dx$$

$$= -n\int p''(x-\theta)\, dx + n\int \frac{[p'(x-\theta)]^2}{p(x-\theta)}\, dx$$

$$= -n\frac{d^2}{d\theta^2}\int p(x-\theta)\, dx + n\int \frac{\{p'(x)\}^2}{p(x)}\, dx$$

$$= -n\frac{d^2}{d\theta^2} 1 + n\int \frac{p'(x)^2}{p(x)}\, dx = n\int \frac{p'(x)^2}{p(x)}\, dx$$

#

Multivariate Case

now g is a function in $\underset{\sim}{x} = (x_1, \dots, x_p)^T$.

Newton's Method

- generalization is straight-forward
- to maximize / minimize $g(\underset{\sim}{x})$, use

$$\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t - [g''(\underset{\sim}{x}_t)]^{-1} g'(\underset{\sim}{x}_t)$$

  - $g''(\underset{\sim}{x})$: $p \times p$ matrix with (i,j)th element as $\dfrac{\partial^2 g(\underset{\sim}{x})}{\partial x_i \partial x_j}$
  - $g'(x) = \left[\dfrac{\partial g(\underset{\sim}{x})}{\partial x_1}, \dots, \dfrac{\partial g(\underset{\sim}{x})}{\partial x_p}\right]^T$, a $p \times 1$ vector

- note: need to compute the inverse of $g''(x_t)$

Fisher Scoring

- use $\underset{\sim}{\theta}_{t+1} = \underset{\sim}{\theta}_t + I(\underset{\sim}{\theta}_t)^{-1} \ell'(\underset{\sim}{\theta}_t)$

Other Newton-like Methods

- computing $g''(x)$ or $[g''(x)]^{-1}$ could be hard.
- the idea is to replace $g''(x)$ by some easily-computable matrix, say $M(x)$
- $\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t - M_t^{-1} g'(\underset{\sim}{x}_t)$

Steepest Ascent Method

- set $M_t = -\alpha_t^{-1} I_p$ (I_p: identity matrix)

- $\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t + \alpha_t g'(\underset{\sim}{x}_t)$

- $\alpha_t > 0$: step size at t which can shrink to ensure ascent

- if at step t, the original step turns out to be downhill, the updating can be back track by halving α_t

- also known as steepest Descent (for minimization)

Gauss-Newton Method

- want to maximize $g(\underset{\sim}{\theta}) = -\sum_{i=1}^{n} \{y_i - f_i(\underset{\sim}{\theta})\}^2$

  where each $f_i(\theta)$ is differentiable

- first consider linear regression $y_i = \underset{\sim}{x}_i^T \underset{\sim}{\theta} + \varepsilon$, $i = 1, \dots, n$

- the least-squares estimator of θ maximizes $g(\underset{\sim}{\theta})$ with $f_i(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{x}_i^T \hat{\underset{\sim}{\theta}}$

- $\hat{\underset{\sim}{\theta}} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{y}$ where $\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{x}_1^T \\ \underset{\sim}{x}_2 \\ \vdots \\ \underset{\sim}{x}_n^T \end{pmatrix}$, $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

- Gauss Newton uses a similar idea for nonlinear $f_i(\underset{\sim}{\theta})$

- let $\underset{\sim}{\theta}^*$ be the unknown maximizer of $g(\underset{\sim}{\theta})$

- consider $h(\underset{\sim}{u}) = -\sum_{i=1}^{n} \{y_i - f_i(\underset{\sim}{\theta} + \underset{\sim}{u})\}^2$

- $h(\underset{\sim}{u})$ is maximized by $\underset{\sim}{u}^* = \underset{\sim}{\theta}^* - \underset{\sim}{\theta}$ ($\underset{\sim}{u}^*$ unknown)

- if $\underset{\sim}{\theta}$ is near $\underset{\sim}{\theta}^*$, $\underset{\sim}{u}^* \approx 0$ and by Taylor expansion of $h(\underset{\sim}{u})$, $\underset{\sim}{u}^*$ should be close to the maximizer of

$$-\sum_{i=1}^{n} \{y_i - f_i(\underset{\sim}{\theta}) - f_i'(\underset{\sim}{\theta})^T \underset{\sim}{u}\}^2$$

- treat $y_i - f_i(\underset{\sim}{\theta})$ as y_i as in linear regression

- $f_i'(\underset{\sim}{\theta})$ as $\underset{\sim}{x}_i$

- we have $\underset{\sim}{u}^* = \underset{\sim}{\theta}^* - \underset{\sim}{\theta} \approx (\underset{\sim}{A}^T\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{z}$

where $\underset{\sim}{A} = \underset{\sim}{A}(\underset{\sim}{\theta}) = \begin{pmatrix} f_1'(\underset{\sim}{\theta})^T \\ \vdots \\ f_n'(\underset{\sim}{\theta})^T \end{pmatrix}$, $\underset{\sim}{z} = \underset{\sim}{z}(\underset{\sim}{\theta}) = \begin{pmatrix} y_1 - f_1(\underset{\sim}{\theta}) \\ \vdots \\ y_n - f_n(\underset{\sim}{\theta}) \end{pmatrix}$

- the updating formula is

$$\underset{\sim}{\theta}_{t+1} = \underset{\sim}{\theta}_t + (\underset{\sim}{A}_t^T \underset{\sim}{A}_t)^{-1} \underset{\sim}{A}_t^T \underset{\sim}{z}_t$$

where $\underset{\sim}{A}_t = \underset{\sim}{A}(\underset{\sim}{\theta}_t)$, $\underset{\sim}{z}_t = \underset{\sim}{z}(\underset{\sim}{\theta}_t)$

- some tricks for Newton / Fisher Scoring:

- calculate $g''(\underset{\sim}{x}_t)$ / $I(\underset{\sim}{\theta}_t)$ every, say, 3 iteration

- use $M_t = \alpha I + g''(\underset{\sim}{x}_t)$

  or $M_t = \alpha I + I(\underset{\sim}{\theta}_t)$

  if $g''(\underset{\sim}{x}_t)$ / $I(\underset{\sim}{\theta}_t)$ is near singular, where $\alpha > 0$