Lecture 6

Numerical Optimization Techniques

Want to maximize or minimize the (possibly nonlinear) objective function $F\left(\underset{\sim}{\theta}\right)$, which is usually log likelihood function.

3 parts to numerical search algorithms:
(1) obtaining initial starting values for the parameters, say $\underset{\sim}{\theta}_0$
(2) updating the candidate value for $\underset{\sim}{\theta}$
(3) determining when the optimum has been reached.

If the objective function is globally concave so there is a unique maximum, then any algorithm which improves the parameter vector at each iteration will eventually find the maximum (example: logit likelihood function). If the function $F\left(\underset{\sim}{\theta}\right)$ is not globally concave, then different algorithms may find different local maxima. However, all iterative algorithms will suffer from the same problem of not being able to distinguish between a local and a global maximum.

The main thing that distinguishes different algorithms is how fast they find the maximum. One algorithm may be better in one case but not in another. Performance is often case specific.

Numerical optimization algorithms can be broadly classified into two types: first derivative methods and second derivative methods. First derivative methods form candidates based on using only the first derivative of the objective function. Second derivative methods form candidates based on using the second derivative of the objective function.

Ref.: EVIEWS User's Manual pp. 619-622.
Also see Ch. 4 in A.C. Harvey EATS.

Second Derivative Methods

A) Newton-Raphson Method

$$\underset{\sim}{\theta}_{(i+1)} = \underset{\sim}{\theta}_{(i)} - H_{(i)}^{-1} \underset{\sim}{g}_{(i)}$$

where g is the gradient vector ($\partial F\left(\underset{\sim}{\theta}\right) / \partial \underset{\sim}{\theta}$) and H is the Hessian matrix $\partial^2 F\left(\underset{\sim}{\theta}\right) / \partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'$.

* Motivation: First-Order Taylor Series Expansion linearize about θ:

$$0 = \frac{\partial \ln L\left(\underset{\sim}{\theta}\right)}{\partial \underset{\sim}{\theta}} = \underset{\sim}{g}\left(\underset{\sim}{\theta}\right) = \underset{\sim}{g}\left(\underset{\sim}{\theta}_i\right) + H\left(\underset{\sim}{\theta}_i\right)\left(\underset{\sim}{\theta} - \underset{\sim}{\theta}_i\right) + R_i = 0$$

$$\because \; \underset{\sim}{\theta} - \underset{\sim}{\theta}_i = -H_{(\underset{\sim}{\theta}_i)}^{-1} \underset{\sim}{g}_{(\underset{\sim}{\theta}_i)}$$

$$\therefore \; \underset{\sim}{\theta}_{(i+1)} = \underset{\sim}{\theta}_{(i)} - H_{(\underset{\sim}{\theta}_i)}^{-1} \underset{\sim}{g}_{(\underset{\sim}{\theta}_i)}$$

If the function is quadratic, the Newton-Raphson technique will find the maximum in a single iteration.

Note: The method of scoring is the same as the Newton-Raphson method except that $-H_{(i)}^{-1}$ is replaced with the inverse of the information matrix evaluated at i-th iterate of $\underset{\sim}{\theta}$. The method of scoring is likely to have a slower convergence rate since the information matrix is only an approximation of the Hessian. However, in many applications, the information matrix has a single form and is much easier to compute. Furthermore, provided the model is identifiable, the information matrix is always positive definite. Therefore, some of the convergence problems of the Newton-Raphson method may be avoided.

B) Quadratic Hill-Climbing (Goldfeld-Quandt)

A straightforward variation of the Newton-Raphson method attributed to Goldfeld and Quandt is the following.

$$\underset{\sim}{\theta}_{(i+1)} = \underset{\sim}{\theta}_{(i)} - \tilde{H}_{(i)}^{-1} \underset{\sim}{g}_{(i)}$$

where $-\tilde{H}_{(i)} = -H_{(i)} + \alpha I$, where I is the identity matrix and α is a positive number (chosen by the algorithm).

The effect of the modification is to push the parameter estimates in the direction of the gradient vector.

(In EVIEWS Quadratic Hill Climbing is used as the default. Note, however, that asymptotic standard errors are always computed from the unmodified Hessian once convergence is achieved).

First Derivative Methods

A) Gauss-Newton/ BHHH

This algorithm follows the Newton-Raphson approach but replaces the negative of the Hessian by an approximation formed from the sum of the outer products of the gradient vectors for each observation’s contribution to the objective function. For least squares and log-likelihood functions, this approximation is asymptotically equivalent to the actual Hessian when evaluated at the parameter values which maximize the function. When evaluated away from the maximum, this approximation may be quite poor.

The algorithm is referred to as Gauss-Newton for general nonlinear least squares problems, and Berndt, Hall, Hall, and Hausman (BHHH) for maximum likelihood problems.

The advantages are: (1) you only need 1st derivatives; (2) the outer product is necessarily positive definite. The disadvantage is that, away from the maximum, this approximation may provide a poor guide to the overall shape of the function, so more iterations may be needed for convergence.

B) Marquardt

This algorithm modifies the Gauss-Newton/ BHHH algorithm in the same manner as the quadratic hill climbing modifies the Newton-Raphson method by adding a correction matrix (or ridge factor) to the outer product matrix. The ridge correction handles numerical problems when the algorithm is near singular and may improve the convergence rate. As above, the algorithm pushes the updated parameter values in the direction of the gradient.

In EVIEWS, the Marquardt Algorithm is the default when a first derivative method is chosen.

Note that when calculating the asymptotic standard errors, they are calculated from the unmodified outer product matrix once convergence is achieved.

Derivative Free Methods

Different Types of Grid Searches:

A) Evaluation of a Random Drawing of Points Uniformly over the Parameter Space (and then after choosing best point, doing a Newton-Raphson iteration for example).

B) Simplex Method (Nelder and Mead 1965)
Nelder, J.A. and Mead, R. (1965), "A Simplex Method for Function Minimization," Computer Journal, 7, 308-313.

C) Adaptive Random Search
Prozanto, C. Walter, E., Venof, A., and Hebruchec, J. (1984), "A General-Purpose Global Optimizer: Implementation and Applications," Mathematics and Computers in Simulation, 26, 412-422.

D) Simulated Annealing
Conana, A., Marchesi, M., Martini, C., and Ridella, S. (1987), "Minimizing Multimodal Functions of Continuous Variables with the Simulated Annealing Algorithm," ACM Transactions on Mathematical Softwares, 13, 262-280.

Goffe, W.L., Ferrier, G.D. and Rogers, J. (1944), "Global Optimization of Statistical Functions with Simulated Annealing," Journal of Econometrics, 60, 65-99.

E) Genetic Algorithm

Dorsey, R.E. and Mayer, W.J. (1995), "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features," Journal of Bus. & Eco. Stat., January, vol.13, no.1, 53-66.