Cowles Commission Discussion Papers: Statistics No. 312A

GRADIENT METHODS OF MAXIMIZATION IN ESTIMATING ECONOMIC PARAMETERS

Herman Chernoff

(Summary and Notations Prepared for the Econometric Society Meeting at Madison, September, 1948.)

The Maximum Likelihood method of estimation is often applied in estimating economic parameters. Mathematically the problem often reduces to maximizing a function of many independent variables. Gradient Methods consist of "climbing uphill" in the direction of steepest ascent with respect to a certain measure of distance. Suppose

$f(x_1, x_2, \ldots, x_n)$ is the function to be maximized and

$(\sum B_{ij} u_i u_j)^{\frac{1}{2}}$ has relevance as the distance from the point

$(x_1, x_2, \ldots, x_n)$ to the point $(x_1 + u_1, x_2 + u_2, \ldots x_n + u_n)$.

If $(x_1, x_2, \ldots, x_n)$ is an initial approximation to the maximum likelihood estimates $(c_1, c_2, \ldots c_n)$ and if our next approximation is $(x_1 + hd_1, x_2 + hd_2, \ldots x_n + hd_n)$, then the d_i which give the direction of steepest ascent satisfy

$$\sum_{j=1}^{n} B_{ij} d_j = \frac{\partial f}{\partial x_i} \qquad \text{or} \qquad d_i = \sum_{j=1}^{n} B^{ij} \frac{\partial f}{\partial x_j}$$

where $\|B^{ij}\| = \|B_{ij}\|^{-1}$. The length of step is then determined by the choice of h.

Suppose that $A_{ij} = \dfrac{-\partial^2 f}{\partial x_i \partial x_j}$ evaluated at the maximum (A_{ij} is not known but can be approximated when we are in the neighborhood of the maximum. Such an approximation may require considerable computing labor.) Suppose that $\|B_{ij}\|$ (which is not necessarily constant) approaches a certain constant matrix as we approach the maximum. Then the convergence rate is greatly dependent upon

$\|C_{ij}\| = \|B_{ij}\|^{-1}\|A_{ij}\|$. If the characteristic roots of $\|C_{ij}\|$ are $0 < \lambda_1 \leq \lambda_2 \leq \lambda_n$ and the corresponding characteristic vectors are $v^1, v^2 \ldots, v^n$, the discrepancy between the maximum likelihood estimates and the mth approximation is given by

$$e^{(m)} = k_1 \prod_{j=1}^{m} (1 - h_j \lambda_1) v^1 + k_2 \prod_{j=1}^{m} (1 - h_j \lambda_2) v^2 + \ldots + k_n \prod_{j=1}^{m} (1 - h_j \lambda_n) v^n$$

Methods of estimating λ_1 and λ_n from the successive iterations are considered. This helps in the choice of those h's which will cause $\prod_{j=1}^{m} (1 - h_j \lambda_i)$ to approach zero rapidly for each i. If λ_1/λ_n is close to one we should have comparatively good convergence. This is the case when $\|B_{ij}\|$ is close to $\|A_{ij}\|$. If $\|B_{ij}\|$ varies during successive iterations the costly inversion may be abbreviated by modifying $\|B_{ij}\|$ so that all but small diagonal blocks of elements vanish. Various compromises in the choice of $\|B_{ij}\|$ are considered. These compromises must take into account the time per iteration for each $\|B_{ij}\|$ and the corresponding rate of convergence.

TERMS:

1. $f(x_1, x_2, \ldots, x_n)$ = likelihood functions which attains its maximum at the Maximum Likelihood estimates $x_1 = c_1$, $x_2 = c_2, \ldots x_n = c_n$.

2. $(\sum B_{ij} u_i u_j)^{\frac{1}{2}}$ = a measure of "distance" from $(x_1, x_2, \ldots x_n)$ to $(x_1 + u_1, x_2 + u_2, \ldots, x_n + u_n)$.

3. $\| B^{ij} \| = \| B_{ij} \|^{-1}$

4. Matrix C has characteristic roots $0 < \lambda_1 \leq \lambda_2 \leq \lambda_n$ and corresponding characteristic vectors $v^{(1)}, v^{(2)}, \ldots v^{(n)}$.

5. x_i^m = i^{th} component of the m^{th} approximation.

   $e_i^m = x_i^m - c_i$ = deviation of the i^{th} component of the m^{th} approximation from the Maximum Likelihood estimates.

   $d_i^{(m)}$ = i^{th} component of a vector in the direction of steepest ascent.

6. $x_i^{m+1} = x_i^m + h_m d_i^m$

   where $\sum B_{ij} d_j^{(m)} = \frac{\partial f}{\partial x_i} (x_1^m, x_2^m, \ldots x_n^m)$

   $$d_i^m = \sum B^{ij} \frac{\partial f}{\partial x_j} (x_1^m, x_2^m, \ldots, x_n^m)$$

7. $$e_i^m \cong (1 - h_0 \lambda_1)(1 - h_1 \lambda_1) \ldots (1 - h_{m-1} \lambda_1) v^{(1)} + (1 - h_0 \lambda_2)(1 - h_1 \lambda_2) \ldots (1 - h_{m-1} \lambda_2) v^{(2)} + \ldots + (1 - h_0 \lambda_n)(1 - h_1 \lambda_n) \ldots (1 - h_{m-1} \lambda_n) v^{(n)}$$