A survey on numerical methods of unconstrained optimization (I)

د.صلاح حمزة عبد *

1) Introduction

When we estimate statistical model parameters, he needs to optimize some kind of objective function . As an example , least square estimates are obtained by minimizing a sum of squares , but in many situations it is not possible to obtain a closed form for the estimates as a function of the sample values , this occurs , when sum of squares can not be transformed so that the normal equations are linear .

Now , let $G(\underline{\alpha})$ **be an objective function , given that is to be minimized with respect to the** $r \times 1$ **parameter** $\underline{\alpha}$ **. Generally, the minimization methods are iterative and follow the following steps,**

(i) We try to find a sequence $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_s$ **of vectors such that** $\underline{\alpha}_s$ **minimizes** $G(\underline{\alpha})$

approximately.

(ii) Starting with initial vector $\underline{\alpha}_1$ **and a vector step** $\underline{\Im}$ **to** $\underline{\alpha}_n$ **in order to obtain** $\underline{\alpha}_{n+1}$ **,**

that is,

$$\underline{\alpha}_{n+1} = \underline{\alpha}_n + \underline{\Im} \quad ----(1)$$

(iii) A step that meets the condition $G(\underline{\alpha}_n) > G(\underline{\alpha}_{n+1})$ **is called acceptable.**

(iv) If no further reduction of the objective function can be obtained, the procedure should terminate .

There are some common troubles may occur when we apply one of numerical optimization method,

* استاذ / الجامعة المستنصرية / كلية الادارة والاقتصاد / قسم الاحصاء

مقبول النشر بتاريخ 2006/12/26

1. **The minimum is not unique.**
2. **No minimum exists.**
3. **Search in a region of the parameter space for a way from a minimum.**
4. **The algorithm used does not approach the existing minimum for some reason.**

In this research we enlarge to <u>study gradient methods as a survey</u>, and then compare among them according to simulation experiment results.

2) Gradient methods

Given a point $\underline{\alpha}_n$, we have to choose an appropriate step length t and a step direction $\underline{\eta}$ such that ,

$$G(\underline{\alpha}_n + tn) < G(\underline{\alpha}_n) \quad ----(2)$$

In fact , we are looking for $\underline{\eta}$ such that $G(\underline{\alpha}_n + tn)$ is a decreasing function of t for t close to zero . Consequently , for $\underline{\eta}$ [11] ,

$$\left.\frac{\partial G(\underline{\alpha}_n + t\underline{\eta})}{\partial t}\right|_{t=0} = \left(\left.\frac{\partial G}{\partial \underline{\alpha}}\right|_{\underline{\alpha}_n}\right)' \left(\left.\frac{\partial G(\underline{\alpha}_n + t\underline{\eta})}{\partial t}\right|_{t=0}\right) \left(\left.\frac{\partial G}{\partial \underline{\alpha}}\right|_{\underline{\alpha}_n}\right)' \underline{\eta} \quad ----(3)$$

Has to be less than zero . If we substitute the gradient of $\left.\frac{\partial G}{\partial \underline{\alpha}}\right|_{\underline{\alpha}_n}$ by $\underline{\beta}_n$,

we can choose ,

$$\underline{\eta} = -Q_n \underline{\beta}_n \quad ----(4)$$

Where is any positive definite matrix . we can [11] get from (1) , $\underline{\alpha}_{n+1} = \underline{\alpha}_n - t_n Q_n \underline{\beta}_n$ where t_n is the step length in the nth iteration .

Since there are many of positive definite matrices , there are many downhill directions ,

(a) If we choose [5] ,

$$Q_n = I_r \quad ----(5)$$

In all iterations , where r is the dimension of the parameter space , then this method is called steepest descent (SD) method . It may converge very slowly if the minimum is in a long and narrow valley [5]

(b) If we choose Q_n to be the inverse of the Hessian matrix ,

$$Q_n = \left[\frac{\partial^2 G}{\partial \underline{\alpha} \, \partial \underline{\alpha}^{/}} \Big|_{\underline{\alpha}_n} \right]^{-1} \qquad ----(6)$$

Then , the method is called Newton-Raphson (NR) method [8] .

(c) If we choose Q_n to be

$$Q_{n+1} = Q_n + P_n \qquad ----(7)$$

Where P_n is the correction matrix to approximate Q_n to the inverse of the Hessian matrix of the objective function in each iteration. This method called variable metric , and if we choose [3] ,

$$P_n = \frac{\underline{\gamma}_n \underline{\gamma}_n^{/}}{\underline{\gamma}_n^{/} \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)} \qquad ----(8)$$

Where $\underline{\gamma}_n = \left(\underline{\alpha}_{n+1} - \underline{\alpha}_n \right) - Q_n \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)$, then the resulting algorithm is called the rank one correction (ROC) method .

(d) Another member of the family of variable metric algorithms is the Davidon-Powell (DFP) method [4,6] , for which ,

$$P_n = \frac{\underline{\mathfrak{I}}_n \underline{\mathfrak{I}}_n^{/}}{\underline{\mathfrak{I}}_n^{/} \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)} - \frac{Q_n \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right) \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)^{/} Q_n}{\left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)^{/} Q_n \left(\underline{\beta}_{n+1} - \underline{\beta}_n \right)} \qquad ----(9)$$

(e) If we have a statistical model ,

$$\underline{y} = \underline{f} \left(x, \underline{\alpha}^{*} \right) + \underline{\in} \qquad ----(10)$$

Then the objective function is the sum of squared errors ,

$$G(\underline{\alpha}) = \left(\underline{y} - \underline{f}(x, \underline{\alpha}) \right)^{/} \left(\underline{y} - \underline{f}(x, \underline{\alpha}) \right) = \left(\underline{\in}(\underline{\alpha}) \right)^{/} \left(\underline{\in}(\underline{\alpha}) \right) \qquad ----(11)$$

And then after some mathematical operations, Bard 1974 gets [1] ,

$$\underline{\alpha}_{n+1} = \underline{\alpha}_n - \left(\left[\frac{\partial \underline{\in}}{\partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right]' \left[\frac{\partial \underline{\in}}{\partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right] \right)^{-1} \left[\frac{\partial \underline{\in}}{\partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right]' \underline{\in}(\underline{\alpha}_n) \qquad ----(12)$$

The above method is called the Gauss-Newton (GN) method .

(f) Maddala (1977) [9] suggests the method of scoring (MOS) , which can be used for maximum likelihood estimation , with the following direction matrix ,

$$Q_n = -\left[\frac{\partial^2 LnL}{\partial \underline{\alpha} \, \partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right]^{-1} \qquad ----(13)$$

Where L is the likelihood function .

(g) Brown and Dennis (BD) (1971) [2] suggests combining the rank one correction method with a Gauss-Newton method . They approximate the Hessian of $f_t(\underline{\alpha})$ iteratively , that is , they choose ,

$$D_{t,n+1} = D_{t,n} + P_{t,n} \qquad ----(14)$$

Where $D_{t,n}$ is the approximate to $\left. \frac{\partial^2 f_t}{\partial \underline{\alpha} \, \partial \underline{\alpha}'} \right|_{\underline{\alpha}_n}$. The correction matrix of rank one is [2] ,

$$P_{t,n} = \frac{\underline{\theta}_{t,n} (\underline{\alpha}_{n+1} - \underline{\alpha}_n)'}{(\underline{\alpha}_{n+1} - \underline{\alpha}_n)' (\underline{\alpha}_{n+1} - \underline{\alpha}_n)} \qquad ----(15)$$

Where $\underline{\theta}_{t,n} = \left. \frac{\partial f_t}{\partial \underline{\alpha}} \right|_{\underline{\alpha}_{n+1}} - \left. \frac{\partial f_t}{\partial \underline{\alpha}} \right|_{\underline{\alpha}_n} - D_{t,n} (\underline{\alpha}_{n+1} - \underline{\alpha}_n)$

The direction matrix for this algorithm is ,

$$\underline{Q}_n = \left(\left[\frac{\partial \underline{\in}}{\partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right]' \left[\frac{\partial \underline{\in}}{\partial \underline{\alpha}'} \Big|_{\underline{\alpha}_n} \right] - \sum_{t,t^*=1}^{T} (y_t - f_t(\underline{\alpha}_n)) D_{t^*,n} \right)^{-1} \qquad ----(16)$$

(h) Marquardt (M) (1963) [10] , modifies procedures that do not guarantee a positive definite direction matrix Q_n by using the fact that $Q_n + \delta_n \tilde{Q}_n$ is always positive definite if $\tilde{Q}_n$ is positive definite and δ_n is sufficiently large . The direction matrix for this algorithm is ,

$$\underline{Q}_n = \left(\left[\left. \frac{\partial \underline{\in}}{\partial \underline{\alpha}^{/}} \right|_{\underline{\alpha}_n} \right]^{/} \left[\left. \frac{\partial \underline{\in}}{\partial \underline{\alpha}^{/}} \right|_{\underline{\alpha}_n} \right] + \delta_n \tilde{Q}_n \right)^{-1} \qquad ----(17)$$

Where I_r can be used as $\tilde{Q}_n$.

(i) We could modify the Hessian matrix of the objective function and use ,

$$Q_n = \left[\left. \frac{\partial^2 G}{\partial \underline{\alpha} \, \partial \underline{\alpha}^{/}} \right|_{\underline{\alpha}_n} + \delta_n I_r \right]^{-1} \qquad ----(18)$$

As the direction matrix . This algorithm is usually referred to as the quadratic hill-climbing (QHC) method . It is derived by Goldfield , Quandt and Trotter (1966) [7] .

3) An empirical study

A simulation experiment was conducted to compare among the gradient methods performance from the speed and the number of iterations points of view , according to the following assumptions ,

1. **we consider the non linear statistical model ,**

$$y_t = \alpha_1^* + \alpha_1^{*2} x_{t1} + \alpha_2^* x_{t2} + \alpha_2^{*3} x_{t2} + \in_t \qquad ----(19)$$
$$t = 1,2,...,t$$

2. **three different sample sizes were selected ,small one (T=10) , moderate (T=30) and large (T=100) .**
3. **the run size R is equal to 1000 .**
4. **the error term $\in_t$ random variable chosen to distribute as standard normal .**
5. **two sets of parameter values $\alpha_1^* = -1$, $\alpha_2^* = 2$ and $\alpha_1^* = 1$, $\alpha_2^* = -1$ have been considered .**

6. **we allow of 0.00001 as absolute error between α_i^* and α_i $(i=1,2)$.**
7. **the values of initial points is equal to zero for any of α_1^* and α_2^* .**
8. **the following two criterions of comparison are considered ,**
   (i) the speed of reach to the real parameter value from the initial point .
   (ii) the number of iterations needed to reach to the real parameter value, starting from the initial point .

The results of simulation was recorded in table (1) and table (2) . Our conclusions are as follows ,

1. **The reach time to the real parameter value, starting from the initial point considered does not effect by the change of sample sizes or the change of real parameter values . It is appear that the sample size value cover the precision and vice versa .**
2. **The performance of Marquardt (M) , Quadratic hill-climbing (QHC) and Gauss-Newton (GN) methods respectively , better than the other methods .**
3. **The number of iterations needed to reach to the real parameter value, starting from the initial point ,increases if one increase the sample size , but it is not effect by the change of real parameter values .**

Table (1) : The average reach time (in minutes) to the real parameter value from the initial point .

method	$\alpha_1^* = -1$, $\alpha_2^* = 2$			$\alpha_1^* = 1$, $\alpha_2^* = -1$		
	T=10	T=30	T=100	T=10	T=30	T=100
SD	18.336	18.822	19.091	17.667	17.917	17.928
NR	6.323	6.947	7.092	6.113	6.121	6.221
ROC	7.839	7.621	7.934	8.093	8.111	8.211
DFP	12.118	12.001	12.139	11.814	11.805	11.827
GN	3.617	3.814	3.615	3.922	3.927	3.842
MOS	4.120	4.321	4.444	4.091	4.004	4.118
BD	4.008	4.128	4.521	4.001	3.512	3.969
M	2.817	2.714	2.776	2.869	2.819	2.886
QHC	2.989	2.817	3.000	2.949	2.958	3.018

Table (2) : The number of iterations needed to reach to the real parameter value , starting from the initial point .

method	$\alpha_1^* = -1$, $\alpha_2^* = 2$			$\alpha_1^* = 1$, $\alpha_2^* = -1$		
	T=10	T=30	T=100	T=10	T=30	T=100
SD	239	286	417	246	269	406
NR	131	158	226	130	155	228
ROC	161	187	285	159	191	293
DFP	203	217	391	200	219	401
GN	106	120	164	109	118	169
MOS	125	146	198	127	152	222
BD	118	136	200	112	130	217
M	85	96	127	81	97	126
QHC	91	109	139	90	112	146

References

1) Bard , Y. (1974) , *Non linear parameter estimation* , Academic press , New York .
2) Brown , K. and Dennis , J. (1971) , *A new algorithm for non linear least square curve fitting* , J. of Math. Software , No.6 , Vol.15 , p.518-528 .
3) Broyden , C. (1967) , *Quasi-Newton methods and their application to function minimization* , Math. Of Comp. , Vol.21 , p.368-381 .
4) Davidon , W. (1959) , *Variable metric method for minimization* , research and development report , purdo university .
5) Flanagan , P. & Vitale , P. and Mendelsohn , J. (1969) , *A numerical investigation of several one-dimensional search procedures in nonlinear regression problems* , Technometrics , Vol.11 , p.265-284 .
6) Fletcher , R. and Powell , M. (1963) , *A rapidly convergent descent method for minimization* , The computer journal , Vol.6 , p.163-168 .
7) Goldfield , S. & Quandt , R. and Trotter , H. (1966) , *Maximization by Quadratic Hill-Climbing* , Econometrica , Vol.34 , p.541-551 .
8) Judge , G. & Hill , R. & Griffiths , W. & Lutkepohl , H. and Lee , T. (1982) , *Introduction to the theory and practice of econometrics* , wiley series , New York .
9) Maddala , G. (1977) , *Econometrics* , McGraw Hill , New York .
10) Marquardt , D. (1963) , *An algorithm for least square estimation of nonlinear parameters* , Journal of applied mathematics , Vol.11 , p.341-441 .
11) Seber , G. (1989) , *Nonlinear estimation* , wiley series , New York .