On the convergence and precision of a process of successive differential corrections

P. E. ZADUNAISKY - V. PEREYRA

Publicación Nº 5 / Marzo 1965

ic
PERU 272 - BUENOS AIRES - ARGENTINA

ON THE CONVERGENCE AND PRECISION OF A PROCESS OF SUCCESSIVE DIFFERENTIAL CORRECTIONS

by P.E. Zadunaisky and V. Pereyra.

ABSTRACT

Let $f_i\ (a_1, a_2, \ldots a_k) = y_i (i = 1,2,\ldots n)(n \quad k)$ be a syster of equations, where f_i are given functions, and y_i are observed quantities. Given an initial set of values $a_1^o, a_2^o, \ldots a_k^o$, there are methods of differential corrections to obtain the most plausible set $a_1, a_2 \ldots a_k$ which satisfies the system of equations. When it is used the "least squares" method it can be reduced to an equivalent iterative process of the form $A_r = \Phi (A_{r-1})$. Using some results from functional analysis, concerning the "fixed points" of an operator Φ, we obtain some conclusions on the convergence and precision of a process of successive differential corrections. A practical example is given.

1.- Introduction. The standard method of differential correction, as used in practical applications, can be described as follows (WHITTAKER and ROBINSON- 1929):

Let

$$f_i\ (a_1, a_2, \ldots, a_k) = y_i \qquad (1.1)$$

be a system of "equations of condition", where f_i are given functions of certain parameters $a_1, a_2 \ldots a_k$ and y_i are observed quantities, which means that they are liable to accidental errors. It is proposed to solve the system (1.1) for the unknowns $a_1, a_2, \ldots a_k$. This system usually is "overdetermined" (n k) and to solve it one usually applies the "least squares" criterium, where the parameters are so determined that they reduce to a minimum the sum

$$S^2 = \sum_{i=1}^{n} (y_i - f_i)^2 \qquad (1.2)$$

Consider first the case in which the system (1.1) is linear, of the form

$$MA = B \tag{1.3}$$

where M is an (nxk) matrix and A and B are the vectors.

$$\begin{aligned} A &= (a_1, a_2, \ldots, a_k) \\ B &= (y_1, y_2, \ldots, y_n) \end{aligned} \tag{1.4}$$

It is well known that in this case the vector A which satisfies the least squares condition is obtained by solving the "normal system"

$$M^T MA = M^T B \tag{1.5}$$

We are particularly interested in the case in which the system (1.1) is not linear. We can linearize it approximately by putting

$$A = A_o + \Delta A_o \tag{1.6}$$

where A_o is a first approximation for A and ΔA_o is a certain correction vector. Putting then

$$F(A) = (f_1, f_2, \ldots, f_n), \tag{1.7}$$

$$M = Df(A) = \left(\frac{\partial f_i}{\partial a_j}\right) \quad \begin{matrix} i = 1,2,\ldots,n \\ j = 1,2,\ldots,k \end{matrix} \tag{1.8}$$

and assuming that the square of any component of ΔA_o is negligible we obtain the linear system

$$M \Delta A_o = B - F(A) \tag{1.9}$$

for the unknown ΔA_o. The corresponding normal system is

$$N \Delta A_o = M^T \hat{B} \tag{1.10}$$

where

$$N = M^T M, \quad \hat{B} = B - F(A) \tag{1.11}$$

Solving (1.10) and applying (1.6) we obtain a new approximation for A. The procedure can be repeated and we have an iteration formula

$$A_r = \underset{\sim}{\Phi}(A_{r-1}) \qquad (1.12)$$

where

$$\underset{\sim}{\Phi}(A) = A + N^{-1}(A) M^T(A) \hat{B}(A) \qquad (1.13)$$

In the language of functional analysis our process of successive differential correction is represented by an equivalent problem of finding a fixed point of the function $\underset{\sim}{\Phi}(A)$. There exists an ample literature on this problem (J. TODD (1962) ch. 14). In the present paper we apply some known results of the theory to derive some conclusions on the convergence and precision of the process of successive differential corrections.

It is worth noting that our analysis reffers only to the numerical process defined by (1.12) and (1.13); it does not concern any probabilistic theory which relates the least squares criterium with the possible distribution of the accidental errors in the observations. However we shall show that these accidental errors can affect the convergence of the process.

2.- Some results from functional analysis.

For the proofs of the results that we are going to mention, the reader should refer to J. TODD (loc. cit.) and J. DIEUDONNÉ (1960) ch. 8.

2.1.- Differential and derivative of a continuous operator.

Let $\mathcal{A}$ and $\mathcal{B}$ be normed linear spaces and let $\mathcal{D}$ be an open subset of $\mathcal{A}$. We consider now a continuous operator T (not necessarily linear) of $\mathcal{D}$ into $\mathcal{B}$. Such an operator is differentiable on $\mathcal{D}$ if, for each $x_o \in \mathcal{D}$, the following conditions are fulfilled:

i) for each $h \in \mathcal{A}$ exists the limit

$$\delta T(x_0, h) = \lim_{t \to 0} \frac{1}{t} [T(x_0 + t h) - T(x_0)]$$

where t is a real variable.

ii) $\delta T(x_0, h)$ is linear and continuous in h.

iii) $$\lim_{h \to 0} \frac{1}{\|h\|} \| T(x_0 + h) - T(x_0) - \delta T(x_0, h) \| = 0$$

$\delta T(x_0, h)$ is called the (Frechet) differential of T in x_0 with increment h. The correspondence between $h \in \mathcal{A}$ and $\delta T(x_0, h)$ is a bounded linear operator A_{x_0}. The (Frechet) derivative DT or T', of T, is an operator from $\mathcal{D}$ into $\mathcal{L}(\mathcal{A}, \mathcal{B})$, the normed linear space of the linear and continuous operators of $\mathcal{A}$ into $\mathcal{B}$, which establishes a correspondence between $x_0 \in \mathcal{D}$ and A_{x_0}.

2.2 Lemma.

Let E^ℓ and E^p be euclidean spaces of dimensions ℓ and p respectively.

Let $T = (t_1(x_1, x_2, \ldots, x_\ell) \ldots, t_p(x_1, x_2, \ldots x_\ell)$ be an operator of an open subset $\mathcal{D} \subset E^\ell$ into E^p. If we assume the real functions $t_i(x_1, x_2, \ldots x_\ell)$ (i=1,2,...p) to be continuously differentiable in the usual sense, then T is a differentiable operator and its derivative is

$$DT(x_1, x_2, \ldots, x_\ell) = \left(\frac{\partial t_i}{\partial x_j} \right)$$

which is the usual Jacobian matrix (i=1,2,...,p; j=1,2,..., ℓ)

2.3. Properties.

1) Let T_1, T_2 be two continuous operators of the open subset $\mathcal{D}$ of E into F (Banach spaces) and let λ_1, λ_2 be any two real numbers; if T_1, T_2 are differentiable on $\mathcal{D}$ then the operator $(\lambda_1 T_1 + \lambda_2 T_2)$ is also differentiable and it is valid the relation $D(\lambda_1 T_1 + \lambda_2 T_2) = \lambda_1 D T_1 + \lambda_2 D T_2$

6) Let A and E be two Banach spaces. Let $\mathcal{H}$ be the space of the linear automorphisms of E. Let u: A $\to \mathcal{H}$ be differentiable. Then applying properties 2 and 4 we have

$$D u^{-1}(a) = - u^{-1}(a) \circ D u(a) \circ u^{-1}(a)$$

2.5 Fixed point theorem:

If:

i) The function $\underset{\sim}{\phi}$ (A) is defined in an open subset $D \subset \mathcal{A}$ (Banach space) and it is differentiable there.

ii) $$\alpha = \sup_{A \in D} \| D \underset{\sim}{\phi}(A) \| < 1$$

iii) It does exist $A_0 \in D$ and a closed sphere $\bar{S}(A_0, \rho)$ of center A_0 and radius

$$\rho \geq \frac{\| A_0 - \underset{\sim}{\phi}(A_0) \|}{1 - \alpha}$$

contained in $\mathcal{D}$

Then the problem A = $\underset{\sim}{\phi}$ (A) has a unique solution $A^* \in \bar{S}(A_0, \rho)$ and the iteration defined by

$$A_r = \underset{\sim}{\phi}(A_{r-1})$$

converges to A^*. Furthermore

$$\| A^* - A_r \| \leq \frac{\alpha}{1-\alpha} \| A_r - A_{r-1} \| \qquad (2.1)$$

2.6 Remark.

If convergence were established we could write

$$\alpha^* = \| D \underset{\sim}{\phi}(A^*) \|$$

Due to our hypothesis of continuity, for any positive real number ε, there exists always a sphere $S^*(A^*, \delta) \subset \mathcal{D}$ in which

$$\sup \| D \underset{\sim}{\phi}(A) \| \leq \alpha^* + \varepsilon \qquad (A \in S^*)$$

and such that for $i > N, A_i \in S^*$. In that case we can write $\alpha = \alpha^* + \varepsilon$ which means that for the estimation of the error through formula (2.1) what essentially counts is the value of α^*. We shall give in section 4 a practical rule to find approximate values of α^*.

3. Convergence of the process of successive differential corrections.

As we may recall, our iteration formula could be written in the simple form

$$A = \underline{\Phi}(A) = A + N^{-1}(A)\, M^{T}(A)\, \hat{B}(A) = A + \Delta A \tag{3.1}$$

Here $\underline{\Phi}$ is an operator (in general not linear) of E^{k} into E^{k}. If the functions f_i are twice continuously differentiable and the matrix M has the rank k then $\underline{\Phi}$ will be differentiable.

In what follows we shall simplify for convenience our notation by writing (3.1) in the form

$$\underline{\Phi} = I + N^{-1} M^{T} \hat{B} \tag{3.2}$$

where I is the identity operator.

In general if X and Y are any operators then it will be understood that $XY(A) = X(A) \circ Y(A)$. Also from (3.1) it is clear that

$$\Delta A = N^{-1} M^{T} \hat{B}(A) \tag{3.3}$$

Now, apply the "fixed point theorem" we are immediately faced with the problem of finding the derivative of the operator $\underline{\Phi}$. By Lemma 2.2 we know that the differentials of operators in the case of finite dimension can be easily expressed by means of the partial derivatives of the components. As we may see from (3.1) these components are not explicitly given and our task becomes apparently a very difficult one.

However we shall show that the derivative of $\underline{\Phi}$ can be obtained explicitly by the expression

$$D\underline{\Phi} = -N^{-1}\left[DM^{T}(M\Delta A - \hat{B}) + M^{T} D M \Delta A\right] \tag{3.4}$$

To that purpose we shall apply all the results stated in the previous section.

We have first

$$D\underset{\sim}{\phi} = D(I + N^{-1}M^T\hat{B}) = I - D(N^{-1}M^T\hat{B}) \qquad (3.5)$$

because DI = I.

Then by Property 5

$$D(N^{-1}M^T\hat{B}) = DN^{-1}M^T\hat{B} + N^{-1}DM^T\hat{B} + N^{-1}M^TD\hat{B} \qquad (3.6)$$

It is easy to see that the last term of (3.6) is equal to $-I$ because $D\hat{B} = -M$.

We have then

$$D\underset{\sim}{\phi} = DN^{-1}M^T\hat{B} + N^{-1}DM^T\hat{B} \qquad (3.7)$$

As, by Property 6,

$$DN^{-1} = -N^{-1}DN\,N^{-1} \qquad (3.8)$$

formula (3.7) reduces to

$$D\underset{\sim}{\phi} = -N^{-1}(DNN^{-1}M^T\hat{B} - DM^T\hat{B}) \qquad (3.9)$$

Furthermore

$$DN = DM^TM + M^TDM \qquad (3.10)$$

and replacing in (3.9) we obtain finally formula (3.4).

Applying this formula we could obtain the value of α to check the condition ii) of the fixed point theorem, which is sufficient for the convergence of the process.

Besides we could apply formula (2.1) to get an estimation of the error.

4.- Practical Application.

Formula (3.4) is too complicated to compute α in practical cases. However one can have reasons other than condition ii) to know if the iteration is convergent, namely the behaviour of S^2, the sum of the squares of the residuals, which can be

then

$$\|DM^T\hat{B}\| \leq \|\Omega\| \sum_{i=1}^{n} |y_i - f_i|$$

where $\Omega = (\omega_{\mu\nu})$

We have finally

$$\alpha^* \leq \|N^{-1}\| \cdot \|\Omega\| \sum_{i=1}^{n} |y_i - f_i| \qquad (4.4)$$

In conclusion:

This upper bound of α^* which through formula (2.1) has a main bearing on the precision and the speed of convergence of the least squares process, depends on three factors, each of them having a special meaning, as follows:

The factor $\|N^{-1}\|$ measures the degree of singularity of the normal matrix N, which in turn depends on the matrix of condition M. If, for instance, among the equations of condition many of them were almost linear dependent then matrix N would be ill-conditioned and $\|N^{-1}\|$ would be large.

The second factor $\|\Omega\|$ reduces evidently to zero in a linear problem. Then it measures in the general case the influence of the non-linear terms neglected in the process.

The size of the third factor depends on the quality of the observations y_i and on the adequate choice of the functions f_i used to represent them.

5.- Numerical example.

The following example stems from a simplification of a problem which is well known to astronomers, namely that of improving the parameters which define an elliptical orbit in order to fit a set of observations in the "least squares" sense. The main simplification consists of reducing the number of parameters, which usually is 6, to only 2 of them.

Given a set of points (x_i, y_i) in a plane, we want to find two parameters a_1 and a_2 such that the ellipse, whose equation is

$$y = f(x, a_1^*, a_2^*) = \frac{a_1^*}{a_2^*}\sqrt{a_2^{*2} - x^2}$$

satisfies the condition

$$S^2 = \sum_{i=1}^{n} (y_i - f(x_i; a_1^*, a_2^*))^2 = minimum$$

According to our notation we can write

$$f_i(a_1, a_2) = f(x_i, a_1, a_2), \quad i = 1, 2, \ldots, n$$

$$B = (y_1, y_2, \ldots, y_n)$$

$$B = (B_i) \text{ where } B_i = y_i - f_i$$

Then we have to form a nx2 matrix

$$M(a_1, a_2) = \left(\frac{\partial f_i}{\partial a_j}\right) \qquad \begin{matrix} i = 1, 2, \ldots, n \\ j = 1, 2 \end{matrix}$$

where the i-th row corresponding to the i-th observation will be indicated by $M_i(a_1, a_2)$. We have also to form the tridimensional array

$$DM^T(a_1, a_2) = \left(\!\!\left(\frac{\partial^2 f_i}{\partial a_\mu \, \partial a_\nu} \right)\!\!\right)$$
$$(\mu, \nu) = 1, 2$$

where for each i we have a 2x2 matrix $DM_i(a_1, a_2)$.

In order to form the product in the right hand member of formula (4.1) we observe that both

$$N = \sum_{i=1}^{n} M_i^T M_i$$

and formula (4.3), by which $DM^T \hat{B}$ is calculated, have a form which may avoid the necessity of keeping in the storage of the computer the equations of condition.

In one of the numerical experiences performed on the Ferranti-Mercury computer of the University of Buenos Aires, we used as observations 10 points $(x_i \; y_i)$ distributed along an ellipse of

parameters $a_1 = 4$ $a_2 = 1$. Before starting the computation we added to the values of y_i arbitrary "errors of observation" smaller than 10^{-2}.

We give in a table the results of the iteration process, which was stopped one step after the estimated error norm reached a value smaller than one unit in the last decimal place furnished by the computer.

At the same time S^2, the sum of the squares of the residuals, reached a value of the same order of magnitude as the errors of the observations.

Taking as a solution (a_1^*, a_2^*) of our problem the values of a_1 and a_2 obtained in the last iteration we computed the error norms $\|A_r - A^*\|$ at each step. Evidently formula (2.1) had given rather good estimations of these error norms.

The theoretical results obtained in this paper could help in designing numerical experiments to analize the behaviour of the iterated least squares process in other typical or more complicated cases.

Universidad Nacional de Buenos Aires
Instituto de Cálculo
Buenos Aires-Argentina.

BIBLIOGRAPHY

J. DIEUDONNE (1960) - Foundations of Modern Analysis (Academic Press, New York and London).

J. TODD (Editor) (1962)- A Survey of Numerical Analysis (Mc Graw-Hill Book Co.Inc. New York).

E. WHITTAKER, G. ROBINSON (1929)- The Calculus of Observations (Blackie and Son Ltd.London).

EXEMPLE

Iteration	Parameters		α_r	Error Norm		
r	a_1	a_2	(Formula (4.1))	(Formula (2.1))	$\|A_r - A^*\|$	S
0	3.000 0000	0.800 0000				
1	3.556 5655	1.018 8672	>1			0.693
2	3.755 8302	1.019 3649	>1			0.455
3	3.944 8389	1.002 2778	>1			0.175
4	3.995 1821	1.000 4295	0.586	0.713×10^{-1}	0.451×10^{-2}	0.547×10^{-1}
5	3.999 6613	1.000 0741	0.365×10^{-1}	0.170×10^{-3}	0.324×10^{-4}	0.513×10^{-2}
6	3.999 6937	1.000 0714	0.380×10^{-2}	0.123×10^{-6}	$<10^{-7}$	0.315×10^{-2}
7	3.999 6937	1.000 0714	0.401×10^{-2}	0.137×10^{-10}	$<10^{-7}$	0.315×10^{-2}