27. ON THE DETERMINATION OF NONGRAVITATIONAL FORCES ACTING ON COMETS

P. E. ZADUNAISKY
University of La Plata and Instituto T. Di Tella, Buenos Aires, Argentina

Abstract. Some recent investigations on the existence and nature of nongravitational forces acting on comets have required the application of a good deal of high precision numerical procedures. In this paper these procedures are examined critically; special attention is given to cases where difficulties may appear when a comet makes a close approach to a planet or the Sun.

1. Introduction

In recent times important advances have been made in the study of irregularities in the motions of comets. The results, especially those obtained by Marsden (1968, 1969, 1970), seem to correspond very closely to Whipple's (1950) theory on the physical nature of cometary nuclei and give evidence for the existence of nongravitational forces acting on comets. In Marsden's work the equations of motion for a comet in a system of heliocentric cartesian coordinates are adopted in the following form:

$$\ddot{x} = -\mu x r^{-3} + \partial R/\partial x + F_1 x r^{-1} + F_2(r\dot{x} - \dot{r}x)h^{-1} + F_3(y\dot{z} - z\dot{y})h^{-1},$$

where $x \to y, z$; $h^2 = (y\dot{z} - z\dot{y})^2 + (z\dot{x} - x\dot{z})^2 + (x\dot{y} - y\dot{x})^2$; $r^2 = x^2 + y^2 + z^2$; μ is the gravitational constant, R the planetary disturbing function, and F_1, F_2 and F_3 are orthogonal components of a nongravitational force to be determined. Further, the F's are assumed to be of the form

$$F_i = A_i \exp(-B_i\tau) \exp(-r^2/C) r^{-\alpha}; \qquad i = 1, 2, 3;$$

where C, α, A_i, B_i are constants and τ is the time from the initial osculation epoch. The constants C and α are more or less arbitrarily set at 2 and 3, respectively. On the other hand, a selection of the constants A_i and B_i, together with the six Keplerian elements of the orbit, are determined by a process of successive differential corrections.

To establish the equations of condition it is necessary to integrate the equations of motion given above. In these calculations certain difficulties arise when a comet makes a close approach to Jupiter. Particularly in the cases of P/Schaumasse and P/Perrine-Mrkos it seems to be very difficult to establish a set of Keplerian and nongravitational parameters in order to link several returns without the appearance of systematic trends in the residuals. It has been considered that in these cases the nature of the trouble may be more mathematical than physical. We agree with such an assumption, and the purpose of the present communication is to point out possible sources of error in the numerical calculations and to indicate possible ways of solving this type of problem.

We believe that one of the difficulties stems from truncation errors accumulated in the numerical integration of the equations of motion when the comet makes a

Chebotarev et al. (eds.), The Motion, Evolution of Orbits, and Origin of Comets, 144–151

close approach to a planet or to the Sun. These errors are then reflected and can be magnified in the process of differential corrections of the parameters. In the following sections we examine in order those numerical processes.

2. The Numerical Integration of the Equations of Motion

Let us first consider the following example, which may be a typical case of the motion of a comet that periodically makes close approaches to the Sun. To simplify matters, we have considered that the heliocentric motion of the comet is perturbed only by Jupiter, and that the motion occurs in the orbital plane of the planet – which is assumed in turn to describe a circular orbit around the Sun. The motion can then be referred to synodic coordinates and described by the equations of the restricted problem of three bodies; the unknowns are in this case the coordinates (x, y) of the comet and their derivatives $(\dot{x}, \dot{y})$. We have chosen the initial conditions corresponding to a periodic orbit of a type described by Rabe (1961); it has been demonstrated by Deprit and Palmore (1966) that this type of orbit is stable. The orbit is defined by the following elements: $a=5.20$ AU ($P=11.86$ yr), $e=0.91$ ($q=0.45$ AU).

For the numerical integration we used the Runge-Kutta-Gill method; during the whole computation the step-size was controlled in order to keep the local truncation error under a certain tolerance limit, given as an input parameter. To estimate the local truncation error we used the well-known method of advancing the computation for two steps (of size h), then repeating it in one step of double size ($2h$) and comparing the results. We found also a good estimate of the *total* errors accumulated after n steps of integration by applying a method developed by ourselves that can be outlined as follows:

Let us consider an 'original problem' of ordinary differential equations that, without loss of generality, may be written in the form:

$$\mathrm{d}x/\mathrm{d}t = f(t, x), \qquad x(0) = x_0.$$

By any numerical process we obtain numerical results $\tilde{x}_n$, and we want an estimate of the error $\xi_n = x(t_n) - \tilde{x}_n$ after n steps of integration. For that purpose we may determine first an empirical function $P(t)$, which can be a polynomial or any other simple function involving coefficients that are adjusted to represent, in the best possible manner, the numerical values $\tilde{x}_n$ for a certain interval of t. Now we can establish a 'pseudo-problem' of the form

$$\mathrm{d}z/\mathrm{d}t = f(t, z) + P'(t) - f(t, P(t)), \qquad z(0) = x_0.$$

The exact solution is evidently $z=P(t)$. If we apply the same numerical process to this pseudo-problem we shall obtain numbers $\tilde{z}$, and the error after n steps will be exactly $\zeta_n = P(t_n) - \tilde{z}_n$. Under certain conditions this error ζ_n is also a good estimate of the error ξ_n in the original problem. These conditions are based on the asymptotic theory of error propagation; for a discussion see Zadunaisky (1964), and for practical applications see Zadunaisky (1969).

We first performed the computation to a moderate degree of accuracy by carrying

https://doi.org/10.1017/S0074180900006422 Published online by Cambridge University Press

nine significant digits and controlling the step size so as to keep the local truncation errors under 10^{-8}. The computation was extended to three orbital periods, and the results are shown in Figure 1, where the accumulated error x and $\dot{y}$, obtained by our method outlined above, are plotted on a semilogarithmic scale as a function of time.

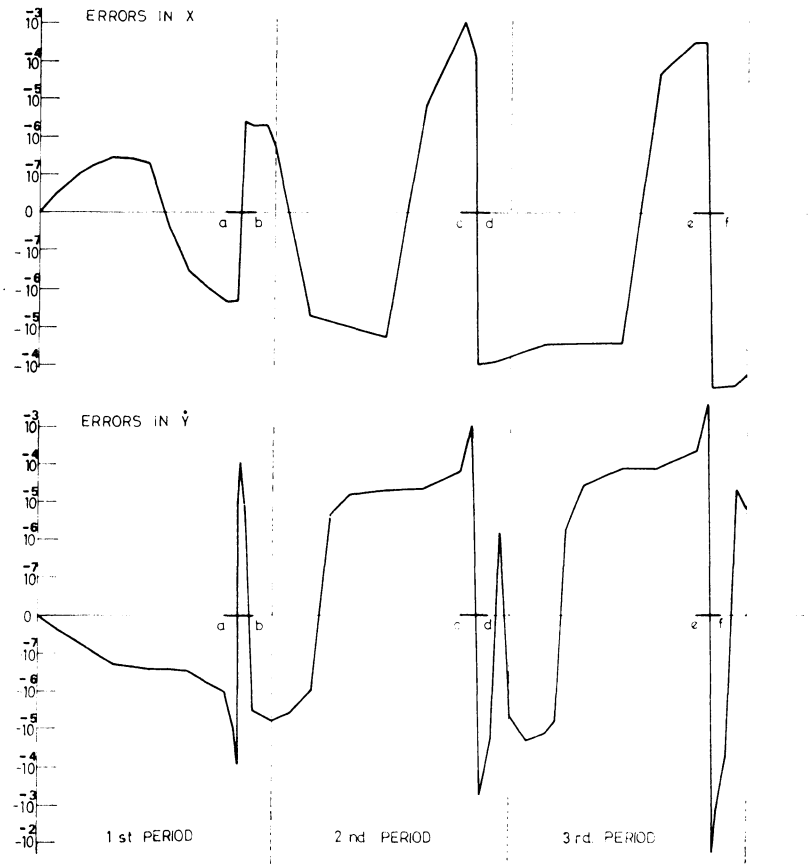

Fig. 1. The accumulated errors in x and $\dot{y}$ during three orbital periods. The local roundoff error is 10^{-9} and the tolerated local truncation error 10^{-8}. In the intervals (a, b), (c, d), and (e, f) the heliocentric distance of the comet is smaller than 3 AU.

It is clearly seen how the errors follow an oscillating pattern as a consequence of the fact that both the orbit and the numerical process are stable. However, when the comet comes closer to the Sun and the heliocentric distance becomes smaller than three astronomical units there is a sudden increase in the size of the accumulated errors, which become much larger than the tolerated local error. The reason for this is that the estimation of the local errors is based on the implicit assumption that the higher derivatives of the unknowns do not become too large, which is true only while the comet is not close to one of the primaries.

https://doi.org/10.1017/S0074180900006422 Published online by Cambridge University Press

We have also performed the same calculation with higher standards of accuracy, namely carrying 16 significant digits and setting the tolerance limit for the local error at 10^{-13}. The behaviour of the accumulated error followed more or less the same oscillatory pattern as before but, of course, its size was much smaller, reaching a maximum limit of the order of 10^{-8}.

In the ordinary process for the correction of just the six Keplerian elements of an osculating orbit, errors of that size should not bother us too much. However, if one adds as unknowns the parameters defining nongravitational forces, the order of magnitude may be that of the errors, and their determination can be substantially affected. The point of interest here is that nonnegligible errors in the calculation may occur precisely in that region of the orbit where most of the observations are made. If one considers a linear system of equations of condition $Ax=b$, those errors appear as perturbations δA in the matrix A and δb in the vector b of residuals. As we shall show in the following sections, the effect of those perturbations can be considerably magnified in the solution x due to the instability that characterizes the process of least squares solution.

3. The Process of Successive Differential Corrections

A. CONVERGENCE AND ESTIMATION OF ERRORS

The standard method of differential correction, as used in practical applications, can be described as follows. Let

$$f_i(a_1, a_2, \ldots, a_k) = y_i; \qquad i = 1, 2, \ldots, n, \tag{1}$$

be a system of equations of condition, where f_i are given functions, in general not linear, of certain parameters $a_1, a_2, \ldots, a_k$, and y_i are observed quantities. Assuming $n>k$, it is proposed to solve the system for the unknowns $a_1, a_2, \ldots, a_k$ in the least squares sense. By using the vectorial notation $A=(a_1, a_2, \ldots, a_k)$, $B=(y_1, y_2, \ldots, y_n)$, $F(A)=(f_1, f_2, \ldots, f_n)$, and assuming that an approximate solution A_r is known, one wants to obtain a further approximation $A_{r+1}=A_r+\Delta A_r$, where the correction ΔA_r is so small that its square and higher powers are supposed negligible.

Such a correction is obtained as the least squares solution of the linear system

$$M \times \Delta A_r = \bar{B}(A_r), \tag{2}$$

where M is the Jacobian matrix $(\partial f_i/\partial a_j)$; $i=1, 2, \ldots, n$; $j=1, 2, \ldots, k$; and $\bar{B}(A_r)=B-F(A_r)$ is the vector of residuals. The least squares solution is given by

$$\Delta A_r = [(M^T M)^{-1} M^T]\bar{B}(A_r). \tag{3}$$

The procedure can be repeated, and we have the iteration formula

$$A_{r+1} = \varphi(A_r), \tag{4}$$

where

$$\varphi(A) = A + N^{-1}(A)M^T(A)\bar{B}(A), \tag{5}$$

https://doi.org/10.1017/S0074180900006422 Published online by Cambridge University Press

$N = M^T M$ being the normal matrix of the system.

In the language of functional analysis the process of successive differential corrections is represented by the equivalent problem of finding a 'fixed point' of the function $\varphi(A)$. On the other hand, the operator φ is said to be a contraction operator when there exists a factor α such that

$$\|\varphi(A_{r+1}) - \varphi(A_r)\| \leq \alpha \|A_{r+1} - A_r\|, \qquad 0 < \alpha < 1, \tag{6}$$

for any pair of vectors A_{r+1} and A_r. Under these conditions the set of approximations A_r converges to a unique solution A^*. Furthermore, an upper bound of the error of the rth approximation is given by

$$\|A^* - A_r\| \leq \frac{\alpha}{1 - \alpha} \|A_r - A_{r-1}\|. \tag{7}$$

Applying some known results of the theory (Zadunaisky and Pereyra, 1965; Pereyra, 1967), we have found an upper bound for the factor α to be

$$\alpha \leq \|N^{-1}(A_r)\| \times \|\Omega(A_r)\| \times \|\bar{B}(A_r)\|, \tag{8}$$

where $\Omega(A_r)$ is a matrix whose elements are defined by

$$\Omega_{pq} = \max_i \frac{\partial^2 f_i}{\partial a_p \partial a_q}. \tag{9}$$

This upper bound of α evidently has, by Equation (6), an important effect on the precision and speed of convergence of the iterated least squares process. According to Equation (8), it depends on three factors, each of them having a special meaning, as shown below.

The factor $\|\Omega(A_r)\|$ evidently reduces to zero in a linear problem; in the general case it measures the influence of the nonlinear terms neglected in the process.

The third factor is the sum of the squares of the residuals (if one adopts the Euclidean norm for vectors), and it depends, of course, on both an adequate choice of the functions f_i and the good quality of the observations.

B. NUMERICAL STABILITY OF THE PROCESS

Now let us turn our attention to the factor $\|N^{-1}(A_r)\|$ in Equation (8). In our calculations we have assumed implicitly that in the linear system, Equation (2), the $n \times k$ matrix M has a rank $r = k$; i.e., all the columns are linearly independent. In that case the normal matrix N is nonsingular and the least squares solution is given by Equation (3), where the expression in brackets is called a pseudo-inverse of the matrix M and is usually indicated by the notation

$$M^+ = (M^T M)^{-1} M^T. \tag{10}$$

When $r < k$ the normal matrix is singular, but it is still possible to obtain two different types of least squares solutions in a way that can be outlined as follows (Rosen, 1964; Pereyra and Rosen, 1964).

To simplify the notation let us write the linear system, Equation (2), in the form

$$Ax = b, \tag{11}$$

where A is an $n \times k$ matrix, and x and b are vectors of k and n dimensions, respectively. In A there are r linearly independent columns, and without loss of generality the matrix A can be partitioned in the form

$$A = (B, \underline{B}),$$

where B is an $n \times r$ matrix of independent columns, and $\underline{B}$ is an $n \times (k-r)$ matrix formed by the rest of the columns of A. The normal matrix $(B^T B)$ is then nonsingular, and the pseudo-inverse of B is

$$B^+ = (B^T B)^{-1} B^T.$$

Then a pseudo-inverse of A is given by the formula

$$A^+ = C^T (C C^T) B^+,$$

where

$$C = B^+ A.$$

It is possible to show that the vector

$$x_m = A^+ b$$

satisfies the least squares condition and x_m has minimum modulus. On the other hand, if we form the matrix

$$A^\# = \begin{pmatrix} B^+ \\ \vdots \\ 0 \end{pmatrix},$$

where the first r rows consist of the matrix B^+ and the remaining rows are zero, the vector

$$x_b = A^\# b$$

also satisfies the least squares condition, and it has at most r nonzero components. x_m is called a *minimum approximate solution* and x_b is a *basic approximate solution.*

The matrices B and $(B^T B)^{-1}$ may be determined by the following recursive algorithm. Let a_q be the qth column of A and B_q a submatrix of A formed by its first q columns. Assuming $(B_q^T B_q)^{-1}$ to be known, one obtains

$$(B_{q+1}^T B_{q+1})^{-1} = \left(\begin{array}{c|c} (B_q^T B_q)^{-1} + \alpha_{q+1}^{-1} u_q u_q^T & -\alpha_{q+1}^{-1} u_q \\ \hline -\alpha_{q+1}^{-1} u_q^T & \alpha_{q+1}^{-1} \end{array} \right),$$

where

$$u_q = B_q^+ a_{q+1}$$

$$\alpha_{q+1} = \| P_q a_{q+1} \|^2,$$

and

$$P_q = (I - B_q (B_q^T B_q)^{-1} B_q^T).$$

https://doi.org/10.1017/S0074180900006422 Published online by Cambridge University Press

The process is initiated with the first column of A, obtaining $(B_1^T B_1)^{-1} = (a_1^T a_1)$ and then adding one column at a time. The column a_{q+1} is linearly independent of those in B_q if $\alpha_{q+1} > 0$, because P_q is a projection matrix that takes the vector a_{q+1} into the space orthogonal to that spanned by B_q. In an actual computation one should never obtain a value of α_{q+1} exactly equal to zero, so that one gives a properly chosen parameter γ, and a column a_{q+1} is considered as linearly independent of those of B_q when $\alpha_{q+1} > \gamma$.

The whole process can be performed by rows instead of by columns. If the system $Ax = b$ represents a linear model of a physical process, the analysis by columns may give an indication of how well the parameters have been selected in the sense that strong correlations do not exist among them. On the other hand, the analysis by rows should show the effects of the successive observations on the model.

So far we have assumed that the rank of A can be well determined and x computed without difficulties. But when the rank of A is not well determined and the normal matrix becomes nearly singular or ill-conditioned, serious difficulties may arise.

In the case that A is a square nonsingular matrix it is known that a perturbation δb in the right-hand member of $Ax = b$, or a perturbation δA in A, may produce changes δx in the solution such that

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \times \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

and

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A\| \times \|A^{-1}\| \frac{\|\delta A\|}{\|A\|},$$

respectively.

The product $\|A\| \times \|A^{-1}\|$ is the 'condition number' of A, and when A is nearly singular it becomes large, and the effects of the perturbations δb and δA can be greatly magnified.

When the matrix A is rectangular, similar, although more complicated, results may be obtained by introducing the 'pseudo-condition' number $\|A\| \times \|A^+\|$. We have described before a method of selection of the successive columns of A, in order to form the submatrix B, by checking the degree of correlations among the variables. The procedure may be completed so as to obtain a pseudo-inverse that produces the smallest possible magnification of the errors introduced by the perturbations δb and δA. The details of these procedures fall beyond the limits of this report; see Pereyra (1969).

4. Final Remarks

We have shown in Section 2 how the numerical integration of the equations of motion may introduce in the equations of condition perturbations, which may be small but not negligible. In Section 3 we have shown how these perturbations can be magnified in the least squares solution of the equations of condition. We think that this can be a possible explanation of the anomalies observed in those comets that

make close approaches to Jupiter or the Sun.

These difficulties are not unavoidable; the standard procedures for the numerical integration of the equations of motion can be applied and completed with the method of error estimation described in Section 2. On the other hand, the resolution of the equations of condition can be performed by the methods outlined in Section 3, with all the precautions indicated there for avoiding the troublesome effects that result from their instability.

We intend to perform a series of numerical experiments on typical cometary orbits by applying this type of technique.

Acknowledgments

The participation of the writer in the Symposium has been made possible through a grant from the National Council of Geo-Heliophysics Research of Argentina.

References

Deprit, A. and Palmore, J.: 1966, *Astron. J.* **71**, 94.
Marsden, B. G.: 1968, *Astron. J.* **73**, 367.
Marsden, B. G.: 1969, *Astron. J.* **74**, 720.
Marsden, B. G.: 1970, *Astron. J.* **75**, 75.
Pereyra, V.: 1967, *Soc. Indust. Appl. Math. J. Num. Anal.* **4**, 27.
Pereyra, V.: 1969, *Aequationes Mathematicae* **2**, 194.
Pereyra, V. and Rosen, J. B.: 1964, *Stanford Univ. Tech. Rept.* CS 13.
Rabe, E.: 1961, *Astron. J.* **66**, 500.
Rosen, J. B.: 1964, *J. Soc. Indust. Appl. Math.* **12**, 156.
Whipple, F. L.: 1950, *Astrophys. J.* **111**, 278.
Zadunaisky, P. E.: 1964, in G. Contopoulos (ed.), 'The Theory of Orbits in the Solar System and in Stellar Systems', *IAU Symp.* **25**, 281.
Zadunaisky, P. E.: 1969, in G. E. O. Giacaglia (ed.), *Periodic Orbits, Stability and Resonances*, Reidel, Dordrecht, p. 216.
Zadunaisky, P. E. and Pereyra, V.: 1965, *Proceedings of the Int. Fed. Inf. Processing Congress*, New York, Vol. 2, p. 488.

Discussion

B. G. Marsden: I appreciate the difficulties you have mentioned, but I think Sitarski and Yeomans will agree with me that these strange anomalies in the motions of comets are sometimes very large indeed. Furthermore, they consistently occur for the same comets, and around the same time. In particular, a large anomaly appears in the case of P/Perrine-Mrkos whether one fits the 1955 and 1962 observations and extrapolates forward to 1968, or whether one fits the 1962 and 1968 observations and extrapolates back to 1955; and calculations have been made, by Sitarski and myself, using completely independent procedures.

P. E. Zadunaisky: I should like to make experiments with this method and see what happens. I am in a position now to obtain quantitative results about the dependence and the upper bounds of the errors we may expect.