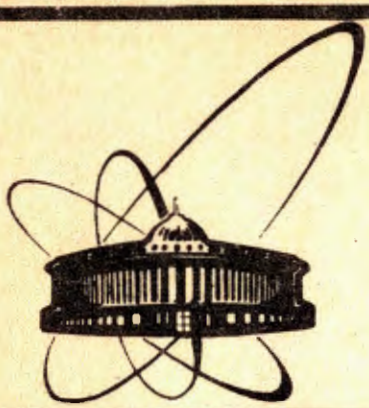

сообщения
объединенного
института
ядерных
исследований
дубна

E10-92-246

V.I.Ilyushchenko

RELATIVE ERROR ENHANCEMENT FACTOR (REEF) VERSUS CONDITION NUMBER (COND) OF SOME TOEPLITZ AND HILBERT TEST MATRICES

1992

1. Introduction

Systems of linear algebraic equations (SLAE) are known to be considered as one of the basic algebraic structures. These systems are treeted as the central research objects in local and global optimization, least squares techniques, linear operator algebra, integral equations etc.

In particular, Fredholm integral equations of the first kind used to describe the smearing effect of experimental detectors by means of an integral convolution transform can be reduced to a discretized SLAE like

$$[A](t) = (f) \tag{1}$$

where the matrix [A] stands for a smearing apparatus function including transformations, acceptance and resolution, the vector (t) is a true spectrum to be found by solving the relevant unstable inverse problem and the vector (f) is an experimental spectrum measured with some hyperadditive, additive and/or multiplicative error (noise) component.

In the most trivial case both [A] and (f) are subject to some additive errors, [dA] and (df), respectively.

2. Condition number (COND) of matrix [A]

An analysis of stability criteria of SLAE results in the following findings [1,2]:

$$[A](dt) = (df) \tag{2}$$

hence

$$|| (dt) || <= || [A]^{-1} || * || (df) || \tag{3}$$

where $|| * ||$ stands for a Holder lp-norm with p =1.0, i.e. a Manhattan (norm, and $[A]^{-1}$ is an inverse matrix of [A].

Then the product of (3) by

$$|| (f) || <= || [A] ||*|| (t) || \qquad (4)$$

yields

$$|| (dt) ||*|| (f) || <=$$

$$<= || [A] ||*|| [A]^{-1} ||*|| (t) ||*|| (df) || \qquad (5)$$

or

$$\frac{|| (dt) ||}{|| (t) ||} <= || [A] ||*|| [A]^{-1} || \frac{|| (df) ||}{|| (f) ||} \qquad (6)$$

where the maximum condition number of the matrix [A] is

$$CMAX(A) = || [A] ||*|| [A]^{-1} || \qquad (7)$$

Here the condition number acts as an enhancement factor of a relative right-hand-side error, $|| (df) ||/|| (f) ||$, to result in a relative true function error, $|| (dt) ||/|| (t) ||$.

The main theoretical properties of COND(A) are as follows:

2.1 COND(A) >= 1.0

2.2 COND(A) = 1.0 for orthogonal or diagonal matrices

2.3 The estimated value of COND(a) depends on the specific norm used in (7), i.e. on the value of index p.

In particular, the minimum theoretical estimate of the COND(A) is provided by the spectral norm, i.e.

$$CMIN(A) = |\lambda_{max}|/|\lambda_{min}| \qquad (8)$$

where λ_{max} and λ_{min} are maximum and minimum eigenvalues of [A], respectively.

On the other hand, if a SLAE with COND(A) = 10^C is solved on a computer with a relative machine precision, MACHEPS = 10^{-E}, then the relevant solution will be correct with N = (E - C) significant figures [3]. With typical values for C =3-6 and E = 6 (IBM computers, single precision mode fll) many well-posed SLAE's will be solved with ill-conditioned final results.

3. Computer tests of COND(A)

However, all of the examples of COND(A) considered in refs. [1,2] and other literature sources pertain to the case when its value is maximum, i.e. COND(A) = CMAX(A) (see (7)).

Moreover, e.g. the generally stable and reliable code DECOMP [2] contains the safeguarding Fortran statement

```
IF(COND.LT.1.0) COND=1.0
```  (9)

which intentionally - but without any feasible mathematical reason - imposes condition 2.2.

Presently there is no computational evaluation of realistic COND(A) values for any specific SLAE. On the other hand, our computational experience demonstrates the theoretical value of COND(A) to be generally overestimated and, moreover, there are observed many cases with

$$COND(A) < 1.0 \qquad (10)$$

corresponding to a "dumping" of initial errors by matrix [A]. Till now these cases were not analyzed in any proper way.

4. Test matrices [A] and test vector (f)

4.1 All of the matrices studied here possessed either a Toeplitz (shift) structure (TM) characteristic of apparatus smearing functions involved in convolution integrals or a Hilbert structure (HM) specified by abnormally high values of COND(A).

The specific shift property of the TM enables one to represent these square matrices by the first row.

The first rows of the test matrices TM1-TM3 are presented below:

| 8.0 | 7.0 | 6.0 | 5.0 | 4.0 | 3.0 | 2.0 | 1.0 | (11)

| 19 | 17 | 13 | 11 | 7 | 5 | 3 | 2 | (12)

|0.240E0 |0.540E-1 |0.443E-2 |0.149E-5|

|0.607E-8|0.913E-11|0.505E-14| (13)

Hilbert matrix HM was generated by means of the well-known formula:

$$H(I,J) = 1.0/FLOAT(I+J-1) \quad (14)$$

These matrices were normalized to unity and multiplied by a factor FACT=10.0.

The test matrix TM1 is composed of eight real numbers arranged in a descending order, elements of TM2 are eight prime numbers, while the matrix TM3 was generated by means of a Gaussian function with a unit width and a zero bias. The first two Toeplitz matrices can be considered as representatives of the well-known triangular apparatus function.

4.2 The right-hand-side vector (f) was chosen in the form of a unimodal distribution characteristic of, e.g. $F_2(x,Q^2)$ structure functions:

| 1.1 | 4.4 | 8.8 | 7.7 | 6.6 | 5.5 | 3.3 | 2.2 | (15)

5. Analyzing computer codes

To analyze both stability and conditionality of the above test matrices we used a Fortran test program composed of codes FSQRT [4], LUDCMP+LUBKSB (LU-decomposition) [5], MINV from the program library SSPP IBM [6] as well as a few additional user-supplied codes to compute l_1-norms for relative errors, determinants, inverted matrices and some other additional information.

6. Test results

6.1 The computer tests were performed for the four main cases:

(df) = 0.0, [dA] = 0.0;

(df) = RNDM(*)*SQRT(f), [dA] = 0.0;

(df) = 0.0, [dA] = RNDM(*)*SQRT(A);

(df) = RNDM(*)*SQRT(f), [dA] = RNDM(*)*SQRT(A).

the relevant results are presented in Table 1.

The most striking features of these simple tests are due to REEF(A) < CMIN(A). Moreover, for a slightly different right-hand-side terms we detected a few cases with REEF(A) > CMAX(A) and REEF(A) < 1.0.

TABLE 1

#	Test matrix	(df)	[dA]	CMIN(A)	CMAX(A)	REEF(A)	RSQ
1	TM1	-	-	4.4	4892	*)	3E-10
2	TM1	+	-	4.4	4892	2.5	2.1
3	TM1	-	+	35.7	20379	*)	72.0
4	TM1	+	+	35.7	20379	35.2	125.0
5	TM2	-	-	5.4	6676	*)	1.0
6	TM2	+	-	5.4	6676	1.7	2.7
7	TM2	-	+	32.4	37824	*)	138.0
8	TM2	+	+	32.4	37824	77.5	288.0
9	TM3	-	-	1.0	144	*)	261.5
10	TM3	+	-	1.0	144	5.0	452.6
11	TM3	-	+	1.3	161	*)	165.4
12	TM3	+	+	1.3	161	4.7	299.1
13	HM	-	-	1E8	2E11	*)	5E13
14	HM	+	-	1E8	2E11	4E7	5E13
15	HM	-	+	90.4	9E5	*)	2E5
16	HM	+	+	90.4	9E5	87.5	1E5

Notes: *) - corresponds to || (df) || = 0.0

These results contradict the approved version of the conditionality theory developed for SLAE apparatus matrices [1,2]. Moreover, the case of REEF(A) < 1.0 corresponds to a pure smoothing effect of initial errors, (df) and [dA], i.e. to mathematical models exceeding in quality the test known orthogonal algebraic structures.

TABLE 2

#	(df) factor	[dA] factor	CMIN(A)	REEF(A)	CMAX(A)	RSQ
1	0.1	0.0	1.05	46.7	144.0	278.2
2	0.5	0.0	1.05	9.5	144.0	350.5
3	1.0	0.0	1.05	5.0	144.0	452.6
4	2.0	0.0	1.05	2.7	144.0	696.4
5	0.0	0.1	1.08	*)	146.2	249.1
6	0.0	0.5	1.20	*)	152.5	206.5
7	0.0	1.0	1.36	*)	161.0	165.4
8	0.0	2.0	1.68	*)	179.4	109.9
9	0.1	0.1	1.08	46.5	146.2	262.5
10	0.5	0.5	1.20	9.3	152.5	280.2
11	1.0	1.0	1.36	4.7	161.0	299.1
12	2.0	2.0	1.68	2.4	179.4	337.3

Notes: *) - corresponds to || (df) || = 0.0

There is an urgent need for finding explicit analytical forms or other alternative explanations of such superorthogonal structures and/or their geometrical interpretation.

6.2 By introducing a multiplying factor like FACT*RNDM(*)*SQRT(f) or FACT*RNDM(*)*SQRT(A) it is possible to study the dependence of REEF(A) and other condition numbers, CMIN(A) and CMAX(A), on relative error levels. The appropriate results are given in Table 2 for the values of FACT = 0.1; 0.5; 1.0; and 2.0 in the case of TM3.

7. Conclusion

The first systematic studies of the dependence of theoretical condition numbers, CMIN(A) and CMAX(A), as well as computed relative error enhancement factor, REEF(A), on relative randomized Poisson errors show many important features which cannot be explained within the framework of the existing linear algebraic models. Especially interesting are the "abnormal" computational results with REEF(A) > CMAX(A), REEF(A) < CMIN(A) and REEF(A) < 1.0, which lack an adequate theoretical analysis. Some theoretical candidates can be found among nonlinear models due to a combined effect of computational and initial errors. To remind the computational error scale, in our case the machine epsilon MACHEPS = 1.0E-6.

REFERENCES

1. G.E.Forsythe and C.B.Moler, Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs (1967).
2. G.E.Forsythe, M.A.Malcolm and C.B.Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs (1977)
3. P.E.Gill and W.Murray, Ch.2 in Numerical Methods for Constrained Optimization (P.E.Gill and W.Murray, Eds.),Academic Press, London (1974).
4. A.A.Kostylev et al., Statistical Processing of Experimental Data on Microcomputers and Programmed Calculators, Energoatomizdat, Leningrad (1991) (in Russian).
5. W.H.Press et al., Numerical Recipes, Cambridge University Press, Cambridge (1986).
6. System/360 Scientific Subroutine Package (360A-CM-03X), Version III (Programmer's Manual), IBM, N.Y. (1960).

Received by Publishing Department
on June 9, 1992.

Илющенко В.И. E10-92-246

Сравнение коэффициента усиления относительной ошибки REEF с числом обусловленности COND для некоторых матриц Теплица и Гильберта

Решение неустойчивых обратных задач определяется алгебраическими свойствами соответствующих систем линейных алгебраических уравнений (СЛАУ), в частности числами обусловленности (COND) аппаратурных матриц (A). Исследованы зависимости минимального и максимального чисел обусловленности, CMIN(A) и CMAX(A) соответственно, а также коэффициента увеличения относительной ошибки, REEF(A), от относительных ошибок матрицы и правой части. Найдено, что в случае тестовых матриц Теплица и Гильберта наблюдаются аномальные значения CMIN(A) > REEF(A) > CMAX(A) и REEF(A) < 1.0. Эти величины не находят адекватного объяснения в рамках стандартных линейных алгебраических моделей.

Работа выполнена в Лаборатории высоких энергий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна 1992

Ilyushchenko V.I. E10-92-246

Relative Error Enhancement Factor (REEF) Versus Condition Number (COND) of Some Toeplitz and Hilbert Test Matrices

The solution of unstable inverse problems is controlled by algebraic properties of the relevant systems of linear algebraic equations (SLAE), e.g. by condition numbers (COND) of the corresponding apparatus matrices (A). The studied relative error dependencies of minimum and maximum condition numbers, CMIN(A) and CMAX(A), respectively, as well as computed relative error enhancement factor, REEF(A), are shown to be specified by the two abnormal features, CMIN(A) > REEF(A) > CMAX(A) and REEF(A) < 1.0. for some Toeplitz and Hilbert test matrices. These new findings cannot be adequately explained within the framework of standard linear algebraic models.

The investigation has been performed at the Laboratory of High Energies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna 1992