Experimental Study on Round-off Error in Matrix Inversion

Takashi Yoshimura

(Received on 31 October, 1978)

1 *A priori* Error Estimate

We consider the effect of the rounding errors in the computed inverses. Because the j th column of the inverse of A is the solution of $Ax=e_j$, we consider first the bounds for the errors made in the solution of the equations

(1) $Ax=b.$

The method we discuss in this paper depends on the successive transformation of the original matrix $A^{(1)}$ into matrces $A^{(2)}, A^{(3)},\ldots,A^{(n)}$ such that each $A^{(k)}$ is equivalent to $A^{(1)}$ and the final $A^{(n)}$ is triangular. The error bounds are most conveniently expressed in terms of vector and matrix norms, throughout we shall use the maximum norms.

Suppose that the data A in (1) are perturbed by the quantity δA. Then if the perturbation in the solution x of (1) is δx we have

(2) $(A+\delta A)\ (x+\delta x)=b.$

An estimate of the relative change in the solution can be given in terms of the relative changes in A as follows:

Let A be non-singular and the perturbation δA be so small that

$\| \delta A \| < 1/\| A^{-1} \|.$

Then if x and δx satisfy (1) and (2) , we have

(3) $$\frac{\delta x}{x} \leqq \frac{\mu}{1-\mu \| \delta A \| / \| A \|} \frac{\| \delta A \|}{\| A \|}$$

where the condition number μ is defined as

$$\mu=\mu\ (A)=\| A \| \cdot \| A^{-1} \|.$$

The basic problem now is to determine the magnitude of the perturbations δA.
It is clear that δA depends upon the round-off errors and method of computation.

We consider the reduction to triangular form by Gausian elimination using a partial pivoting for size. This strategy mearly determines a re-ordering of the row of A, we can assume that, without any loss of generality, the system has been ordered so that the natural order of pivots is used.

We denote the computed elements of the k th matrix $A^{(k)}$ by $a_{ij}^{(k)}$ and the computed multipliers by m_{ij}. Then we have

(4) $$\sum_{k=1}^{\min(i,j)} m_{ik} a_{kj}^{(k)} = a_{ij}^{(1)} + \sum_{k=2}^{\min(i,j+1)} \varepsilon_{ij}^{(k)}$$

where $\varepsilon_{ij}^{(k)}$ is the error made in computing $a_{ij}^{(k)}$ and m_{ij}.The element $a_{ij}^{(i)}$ is an element of the i th pivotal row and undergoing no further change.

Writing L for lower triangular matrix formed by the m_{ij} augmented by a unit diagonal, and U for the upper triangular matrix formed by the pivotal row, (4) gives

(5) $$LU=A^{(1)}+E^{(2)}+\dots+E^{(n)}=A+E$$

where $E^{(k)}$ is the matrix formed by $\varepsilon_{ij}^{(k)}$. Note that this has null rows 1 to $k-1$ and null columns 1 to $k-2$.

The solution of the equations $Ax=b$ is now obtained by solving

$$LUx=b$$

which is performed in the two steps

$$Ly=b, \quad Ux=y.$$

The vectors actually obtained are the exact solutions of, say,

(6) $$(L+\delta L)\ y=b$$

(7) $$(U+\delta U)\ x=y.$$

The perturbations δL and δU arise from the finite precision arithmetic performed in solving the triangular systems with the coefficients L and U. Upon multiplying (7) by $L+\delta L$ and using (6) we have

$$(A+\delta A)=(L+\delta L)\ (U+\delta U)$$

From (5) , it follows that

$$\delta A=E+L(\delta U)+(\delta L)U+(\delta L)(\delta U).$$

Since L and U are explicitly determined by the computations, their norms can also, in principle, be obtained, we must estimate E, δU and δL. We shall assume that floating-point arithmetic operations are performed with a t-digit mantissa, and let $\rho=\max_{i,jk}|a_{ij}^{(k)}|$. If A is non-singular and t sufficiently large, then we have

$$E=(e_{ij}), \qquad |e_{ij}|\leqq\begin{cases}2(i-1)\rho u & (i\leqq j)\\ (2j-1)\rho u & (i>j)\end{cases}$$

where $u=\beta^{1-t}$.

The elements in δL and δU can be estimated from a single analysis of the error in solving any triangular system with the same arithmetic. Assuming that scalar products are accumulated in a double precision accumulator, we have

$$\delta L=\mathrm{diag}\ (-\varepsilon_i), \qquad |\varepsilon_i|<u$$

and

$$\delta U=\mathrm{diag}\ (-u_{ii}w_i), \qquad |w_i|<u.$$

We are now able to obtain estimates of the elements in δA. Let t be so large that $nu<1$. Then the computed solution x satisfies

$$(A+\delta A)\ x=b$$

where

(8) $$|\delta a_{ij}|\leqq\begin{cases}\rho(2i-1)u & (i<j)\\ 2\rho\ ju & (i\geqq j)\end{cases}$$

From (8) we easily find that

(9) $$\|\delta A\|\leqq\rho n(n+1)u$$

and this can be employed in (3) to obtain maximum norm bounds on the relative error.

Above results can be applied to inversion of a matrix A. Since the j th column x_j of the inverse matrx is the solution of the equation

$$LUx=e_j \quad (j=1,2...,n),$$

the each computed x_j satsfies the realtion

$$(A+\delta A_j)\ x_j=e_j\ .$$

Although the perturbation δA_j depends on e_j, but the bound of $\| \delta A_j \|$ is independent of each j.

Thus, if A is non-singular and $\| A^{-1}\delta A \| <1$, then $A+\delta A$ is non-singular and we have

$$(10) \quad \frac{\| (A+\delta A)^{-1}-A^{-1} \|}{\| A^{-1} \|} \leqq \frac{\| A^{-1}\delta A \|}{1-\| A^{-1}\delta A \|} \leqq \frac{\mu}{1-\mu \| \delta A \| / \| A \|} \frac{\| \delta A \|}{\| A \|}$$

where

$$\| \delta A \| \leqq n(n+1)\rho u.$$

2 *A Posteriori* Error Estimate

As shown in the following numerical experiments *a priori* error bound (10) is, in general, a tremendous overestimate for large n. Thus we consider now the *a posteriori* error bounds for computed inverse.

Let A be the matrix to be inverted and let C be the computed inverse. We use a measure of error called the residual matrix

$$R=AC-I.$$

If $\| R \| <1$, then we have

$$(11) \quad \| C-A^{-1} \| \leqq \| C \| \| R \| / (1-\| R \|).$$

Since A and C are presumed known, we could actually compute $\| C \|$, $\| A \|$ and $\| R \|$ in the estimate (11). This, of course, is what is meant by *a posteriori* estimate.

3 Numerical examples

We consider the numerical inversion of the following symmetric matrices.

$$A_1=(a_{ij}), \quad a_{ij}=\begin{cases} d=1.001 & (i=j) \\ 1 & (i \neq j) \end{cases}$$

$$A_2=(a_{ij}), \quad a_{ij}=n-|i-j|$$

$$A_3=(a_{ij}), \quad a_{ij}=\left(\frac{2}{n+1}\right)^{1/2} \sin\left(\frac{ij\pi}{n+1}\right)$$

$$A_4=(a_{ij}), \quad a_{ij}=\begin{cases} -2 & (i=j) \\ 1 & (|i-j|=1) \\ 0 & (|i-j|\gneq 1) \end{cases}$$

Numerical results are given in the following table.

For simplicity, we have denoted say, 4.45×10^{-5}by4.45 (-5). These numerical experiments were performed with the HITAC 8250 computer. Since for this computer, $\beta=16$, $t_s=6$, and $t_d=14$, so we have used $u=2^{-20}$and $u=2^{-52}$for single and double precision arithmetic respectively. Moreover, we have evaluated the relative error in the computed inverse by $\| R \|$, assuming that, in (11), $\| C \|$ is

A_1 : positive definite

n / s.p. / d.p.	$\|A\|$ / $\frac{\|C-A^{-1}\|}{\|A^{-1}\|}$	$\|A^{-1}\|$ / $n\rho u\|A^{-1}\|$	μ / $\|R\|$	ρ / $\frac{\|R\|}{\|C-A^{-1}\| / \|A^{-1}\|}$
5	5.00	1.60(3)	8.00(3)	1.00
	4.45(−5)	7.64(−3)	3.05(−4)	6.86
	2.20(−14)	1.78(−12)	5.68(−13)	25.9
10	10.0	1.80(3)	1.80(4)	1.00
	4.47(−5)	1.72(−2)	2.90(−4)	6.48
	1.46(−14)	4.00(−12)	9.77(−13)	66.9
15	15.0	1.87(3)	2.80(4)	1.00
	4.50(−5)	2.67(−2)	5.04(−4)	11.2
	1.22(−14)	6.22(−12)	1.47(−12)	120.
20	20.0	1.90(3)	3.80(4)	1.00
	8.55(−4)	3.63(−2)	3.97(−4)	0.464
	2.22(−13)	8.45(−12)	2.06(−12)	9.28
25	25.0	1.92(3)	4.80(4)	1.00
	3.88(−4)	4.59(−2)	8.24(−4)	2.13
	1.96(−13)	1.07(−11)	2.48(−12)	12.6

A_2 : positive definite

$\|A\|$ / $\frac{\|C-A^{-1}\|}{\|A^{-1}\|}$	$\|A^{-1}\|$ / $n\rho u\|A^{-1}\|$	μ / $\|R\|$	ρ / $\frac{\|R\|}{\|C-A^{-1}\| / \|A^{-1}\|}$
19.0	2.00	38.0	5.00
1.81(−6)	4.77(−5)	3.44(−6)	1.90
2.83(−16)	1.11(−14)	1.39(−15)	4.94
75.0	2.00	150.	10.0
3.32(−6)	1.91(−4)	1.44(−5)	4.33
7.42(−16)	4.44(−14)	3.35(−15)	4.52
169.	2.00	338.	15.0
7.10(−6)	4.29(−4)	2.77(−5)	3.90
1.81(−15)	9.99(−14)	7.23(−15)	4.00
300.	2.00	600.	20.0
4.07(−5)	7.63(−4)	9.02(−5)	2.21
1.05(−14)	1.78(−13)	2.18(−14)	2.08
469.	2.00	938.	25.0
1.52(−4)	1.19(−3)	2.13(−4)	1.40
1.84(−14)	2.78(−13)	4.42(−14)	2.40

A_3 : orthogonal

5	2.15	2.15	4.64	2.00
	1.88(−6)	2.05(−5)	2.00(−6)	1.06
	9.92(−16)	4.78(−15)	1.10(−15)	1.11
10	2.97	2.97	8.80	3.10
	1.09(−5)	8.78(−5)	4.57(−6)	0.489
	2.39(−15)	2.04(−14)	1.43(−15)	0.599
15	3.59	3.59	12.9	2.88
	8.78(−6)	1.48(−4)	8.26(−6)	0.941
	2.11(−15)	3.45(−14)	2.49(−15)	1.18
20	4.12	4.12	17.0	3.48
	1.22(−5)	2.73(−4)	1.14(−5)	0.935
	3.28(−15)	6.37(−14)	3.41(−15)	1.04
25	4.59	4.59	21.0	4.82
	2.38(−5)	5.27(−4)	1.65(−5)	0.694
	4.21(−15)	1.23(−13)	4.28(−15)	1.02

A_4 : negative definite

4.00	4.50	18.0	2.00
1.21(−6)	4.29(−5)	3.22(−6)	2.67
2.31(−16)	9.99(−15)	8.47(−16)	3.66
4.00	15.0	60.0	2.00
6.76(−6)	2.86(−4)	9.95(−6)	1.47
9.76(−16)	6.66(−14)	2.51(−15)	2.57
4.00	32.0	128.	2.00
1.38(−5)	9.16(−4)	2.45(−5)	1.77
1.45(−15)	2.13(−13)	4.95(−15)	3.42
4.00	55.0	220.	2.00
1.99(−5)	2.10(−3)	4.18(−5)	2.11
2.01(−15)	4.88(−13)	1.00(−14)	4.99
4.00	84.5	338.	2.00
2.68(−5)	4.03(−3)	6.10(−5)	2.28
3.03(−15)	9.38(−13)	1.49(−14)	4.91

approximately equal to ‖ A ‖ , and ‖ R ‖ is far smaller than unity.

From above results, we see that the accuracy of the computed inverse with double precision arithmetic has been improved by 9 or10 decimal places than with single precision arithmetic. For symmetric and positive definite matrix A it can be shown that

$$\rho \leqq \max_{ij} | a_{ij} | \cdot$$

For any real matrix, however, from our experience, it might be expected that

$$\rho = \rho \ (n) \leqq n.$$

References

1) J.H.Wilkinson: Rounding Error in Algebraic Processes. Her Britannic Majesty's Stationery Office (1963) .
2) E.Issacson and H.B.Keller: Analysis of Numerical Methods. John Wiley & Sons, Inc. (1966) .
3) J.R.Westlake: A Handbook of Numerical Matrix Inversion and Solution of Linear Equations. John Wiley & Sons, Inc. (1968) .