A Fast Method for Estimating the Condition Number of a Matrix

MAKOTO NATORI* and ATSUKO TSUKAMOTO**

A method for estimating the condition number of a matrix is given. This method is based on the fact that $||A^{-1}||_1$ can be approximated by $||(A^T)^{-1}\mathbf{e}||_\infty$, if a vector $\mathbf{e}$ is appropriately chosen. Numerical experiments show that this method gives accurate estimates.

1. Introduction

When a system of linear equations

$$A\mathbf{x}=\mathbf{b} \tag{1.1}$$

is numerically solved, the condition number of the matrix A plays an important role. In fact, the condition number is a measure of the sensitivity of the solution $\mathbf{x}$ to changes in A and $\mathbf{b}$, i.e. it can be regarded as a magnification factor of relative error. The condition number of the matrix A is defined by

$$\mathrm{cond}(A)=\|A\|\,\|A^{-1}\| \tag{1.2}$$

where we use for the matrix norm an induced norm given by

$$\|A\|=\max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{1.3}$$

It is well known that if we choose the l_1 vector norm

$$\|\mathbf{x}\|_1=\sum_{i=1}^{n}|x_i|, \tag{1.4}$$

then

$$\|A\|_1=\max_{\mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1}=\max_{1\leq j\leq n}\sum_{i=1}^{n}|a_{ij}|. \tag{1.5}$$

After this, we deal with the condition number with respect to this norm

$$\mathrm{cond}_1(A)=\|A\|_1\|A^{-1}\|_1. \tag{1.6}$$

It is easy to compute $\|A\|_1$, while the computation of $\|A^{-1}\|_1$ takes roughly twice as much time as that required for the Gaussian elimination of A. Therefore, it is desirable to obtain an estimate of the condition number without computing the inverse.

The subroutine DECOMP described in Chapter 3 of [1] estimates the condition number of the matrix A by

$$\mathrm{cond}_1(A)\approx\|A\|_1\frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_1}, \tag{1.7}$$

that is

$$\|A^{-1}\|_1\approx\frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_1}, \tag{1.8}$$

where $\mathbf{y}$ and $\mathbf{z}$ are vectors determined by solving two systems of equations

$$A^T\mathbf{y}=\mathbf{e}, \tag{1.9}$$

$$A\mathbf{z}=\mathbf{y}, \tag{1.10}$$

where A^T is the transpose of A and $\mathbf{e}$ is a vector with components ±1 chosen to maximize the growth in the magnitude of the components of $\mathbf{y}$.

It must be remarked that the 128th and the 130th lines in the subroutine DECOMP should read

```
T=T+A(I, K)*WORK(I)
```

and

```
WORK(K)=T+WORK(K)
```

respectively. These corrections were confirmed by C. B. Moler [2].

In this paper, we propose a new method for estimating the condition number. Our method reduces to one half the computational work required for the estimation compared with the method in DECOMP. Moreover, it is shown by numerical experiments that our method gives more accurate estimates for almost all matrices adopted in our experiments.

2. A New Method for Estimating the Condition Number

We estimate the condition number of a matrix by

$$\mathrm{cond}_1(A)\approx\|A\|_1\|\mathbf{y}\|_\infty, \tag{2.1}$$

that is

$$\|A^{-1}\|_1\approx\|\mathbf{y}\|_\infty, \tag{2.2}$$

where $\mathbf{y}$ is a vector determined by solving the equation (1.9), and $\|\mathbf{y}\|_\infty$ is the l_∞ vector norm defined by

$$\|\mathbf{y}\|_\infty=\max_{1\leq i\leq n}|y_i|. \tag{2.3}$$

Since the vector $\mathbf{z}$ is not necessary in our method, the amount of computational work is reduced to one half compared with that required in DECOMP.

The matrix norm corresponding to the l_∞ vector norm

*Institute of Information Sciences and Electronics, University of Tsukuba, Sakura-Mura, Ibaraki 305, Japan.

**Mechanical Engineering Research Laboratory, Hitachi Ltd., Tsuchiura, Ibaraki 300, Japan.

is given by

$$\|A\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1\leqq i\leqq n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.4}$$

Suppose the maximum of $\sum_{j=1}^{n} |a_{ij}|$ is attained for $i=m$. Then it is clear that the vector $\mathbf{x}$ which maximizes $\|A\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty$ has the components ± 1 whose signs are chosen to be equal to the signs of the elements of the m-th row of A, i.e.

$$x_j = \mathrm{sgn}(a_{mj}), \quad 1\leqq j\leqq n \tag{2.5}$$

where

$$\mathrm{sgn}(t) = \begin{cases} +1 & \text{for } t\geqq 0, \\ -1 & \text{for } t<0. \end{cases} \tag{2.6}$$

Since it is obvious from (1.5) and (2.4) that

$$\|A\|_1 = \|A^T\|_\infty, \tag{2.7}$$

we have

$$\|A^{-1}\|_1 = \|(A^T)^{-1}\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|(A^T)^{-1}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}. \tag{2.8}$$

We now claim that

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\|(A^T)^{-1}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \approx \frac{\|(A^T)^{-1}\mathbf{e}\|_\infty}{\|\mathbf{e}\|_\infty} = \|\mathbf{y}\|_\infty, \tag{2.9}$$

where $\mathbf{e}$ is a vector with components ± 1, the sign of each component being determined by the algorithm described below. The vector $\mathbf{y}$ is given by solving (1.9).

We describe the algorithm for computing the vector $\mathbf{y}$. If the Gaussian elimination with partial pivoting is carried out, A is decomposed into the product of a permuted version of a lower triangular matrix L and an upper triangular matrix U so that

$$A=(PL)U \tag{2.10}$$

where P is a permutation matrix. Thus the system of equations (1.9) is decomposed into two triangular systems

$$U^T\mathbf{x}=\mathbf{e}, \tag{2.11}$$

$$(PL)^T\mathbf{y}=\mathbf{x}. \tag{2.12}$$

Let $\mathbf{e}=(-e_1, -e_2, \ldots, -e_n)^T$. The algorithm to solve (2.11) is described as follows.

$e_1 = 1$
$x_1 = -e_1/u_{11}$
for $k=2$ to n do
  $t = \sum_{i=1}^{k-1} u_{ik}x_i$
  $e_k = \mathrm{sgn}(t)$
  $x_k = -(t+e_k)/u_{kk}$

The sign of e_k is chosen to be the same as that of t. This gives $|x_k|$ the larger of its two possible values. The systems of equations (2.12) is solved by back substitution with row exchanges. Suppose the record of pivoting be stored in the pivot vector $\mathbf{p}=(p_k)$, where p_k is the index of the k-th pivot row. Then the algorithm to solve (2.12) is written as follows.

$y_n = x_n$
for $k=n-1$ to 1 do
  $t = \sum_{i=k+1}^{n} l_{ik}y_i$
  $y_k = x_k - t$
  $m = p_k$
  if $m\neq k$ then exchange y_m and y_k

It is usually found that the vector $\mathbf{y}$ obtained above has large component [3]. Then, from (2.8) and (2.9) it is seen that $\|\mathbf{y}\|_\infty$ gives a good approximation to $\|A^{-1}\|_1$.

3. Numerical Examples

In this section we give some numerical examples. The computations were carried out on a HITAC M-170 (single precision).

Example 1. As a typical matrix whose condition number is large, we chose the Hilbert matrix $A=(a_{ij})$:

$$a_{ij} = \frac{1}{i+j-1}, \quad 1\leqq i,j\leqq n. \tag{3.1}$$

The condition numbers of the Hilbert matrices for $3\leqq n\leqq 6$ were estimated by DECOMP and by our method. The exact values were computed using the inverse matrices given in [4]. The results are shown in Table 1. It is observed that our estimates are quite accurate. For $n=3$ our estimate is slightly larger than the exact value. This is due to the rounding errors in the triangular decomposition.

Table 1. Results for the Hilbert matrix.

n	DECOMP	Ours	Exact
3	6.808213×10^2	7.480132×10^2	7.48×10^2
4	2.152290×10^4	2.837426×10^4	2.8375×10^4
5	6.886816×10^5	9.363152×10^5	9.43656×10^5
6	1.975518×10^7	2.541669×10^7	2.907029×10^7

Example 2. As a matrix whose condition number is of moderate magnitude, we chose the Frank matrix $A=(a_{ij})$:

$$a_{ij} = n+1-\max(i,j), \quad 1\leqq i,j\leqq n. \tag{3.2}$$

The inverse matrix $A^{-1}=(b_{ij})$ is tridiagonal, whose non-zero elements are given by

$$\begin{aligned} &b_{11}=1, \qquad b_{ii}=2 \quad (2\leqq i\leqq n), \\ &b_{i,i-1}=b_{i-1,i}=-1 \quad (2\leqq i\leqq n). \end{aligned} \tag{3.3}$$

In this case

$$\|A\|_1 = \frac{1}{2}n(n+1) \tag{3.4}$$

and

$$\|A^{-1}\|_1 = 4 \tag{3.5}$$

hold for $n \geqq 3$. The results are shown in Table 2. Our estimates agree well with the exact values. Again our estimate for $n=6$ gives a slightly larger value than the exact value due to the rounding errors.

Table 2. Results for the Frank matrix.

n	DECOMP	Ours	Exact
3	19.33330	23.99995	24
4	34.61534	39.99997	40
5	53.82347	59.99988	60
6	77.00011	84.00017	84

Example 3.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \\ -0.0002 & 1 & 1 \end{bmatrix}$$

In this case $\|A\|_1 = 5$ and $\|A^{-1}\|_1 = 9.990014$, therefore

$$\mathrm{cond}_1(A) = 49.95007.$$

The estimates by DECOMP and our method are 39.40695 and 29.97002, respectively. This is one of the rare cases in our numerical experiments that produced a poorer estimate than DECOMP.

References

1. Forsythe, G. E., Malcolm, M. A. and Moler, C. B. Computer Methods for Mathematical Computations, Prentice-Hall, 1977.
2. Moler, C. B. Private communication.
3. Cline, A. K., Moler, C. B., Stewart G. W. and Wilkinson, J. H. An Estimate for the Condition Number of a Matrix, *SIAM J. Num. Anal.*, **16**, 368–375 (1979).
4. Gregory, R. T. and Karney, D. L. A Collection of Matrices for Testing Computational Algorithms, *John Wiley & Sons*, 1969.

(Received June 14, 1982)