ISSN 1995-4239, Numerical Analysis and Applications, 2019, Vol. 12, No. 4, pp. 388–394. © Pleiades Publishing, Ltd., 2019.
Russian Text © The Author(s), 2019, published in Sibirskii Zhurnal Vychislitel'noi Matematiki, 2019, Vol. 22, No. 4, pp. 465–473.

Improving the Stability of Triangular Decomposition of Ill-Conditioned Matrices

V. N. Lutay[1]*

[1] *Southern Federal University, ul. Bolshaya Sadovaya 105/42, Rostov-on-Don, 344006 Russia*

Received March 27, 2018; in final form, September 23, 2018; accepted July 25, 2019

Abstract—An approach to improving the stability of triangular decomposition of a dense positive definite matrix with a large condition number by using the Gauss and Cholesky methods is considered. It is proposed to introduce additions to standard computational schemes with an incomplete inner product of two vectors which is formed by truncating the lower digits of the sum of the products of two numbers. The truncation in the process of decomposition increases the diagonal elements of the triangular matrices by a random number and prevents the appearance of very small numbers during the Gauss decomposition and a negative radical expression in the Cholesky method. The number of additional operations required for obtaining an exact solution is estimated. The results of computational experiments are presented.

DOI: 10.1134/S1995423919040062

Keywords: *ill-conditioned matrix, triangular decomposition, stability improving, discarding least significant digits, incomplete inner product.*

INTRODUCTION

The instability of solving a SLAE with a positive definite square matrix A of order n,

$$Ax = b, \tag{1}$$

(the matrix A is assumed to be ill-conditioned owing to the fact that its rows are almost linearly dependent) can lead, due to an increase in the rounding errors, to the appearance of very small diagonal elements of the upper triangular matrix, up to zero ones, in the Gauss method and to a negative radicand in the Cholesky method [1]. The generally accepted method of choosing pivot elements for positive definite, and much less for ill-conditioned matrices, has no effect [2]. One way of increasing the stability of solving a SLAE is an explicit or implicit method of preconditioning for the original matrix. In the former case the matrix A is multiplied by a matrix that is inverse to some matrix M [3]. In the latter, some operations that differ from the standard Gauss and Cholesky procedures are performed in the calculation process. The result of this decomposition (called incomplete decomposition) are triangular matrices whose product is the matrix M (called a split matrix):

$$M = A + N, \tag{2}$$

where N is the matrix of errors of the decomposition. Incomplete decomposition is widely used for sparse matrices to avoid the appearance in the decomposition of nonzero values of those elements that are zero in the original matrix. In the Intel Math Kernel Library [4], implicit preconditioning consists in replacing the diagonal elements that are zero or close to zero in the matrices obtained in the calculations by a given small number.

In the present paper, a method for improving the stability of triangular decomposition is considered. It consists in increasing the diagonal elements of triangular matrices by some number determined during the decomposition.

*E-mail: vnlutay@sfedu.ru

Complete triangular Gaussian and Cholesky decomposition forms triangular matrices L_A, U_A, and H_A such that

$$A = L_A U_A, \qquad A = H_A H_A^\top.$$

The formulas for calculating their elements are as follows [5] (the diagonal terms have an additional index A):

$$u_{Aii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik} u_{ki}, \tag{3}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \;\; l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \right) \Big/ u_{Aii},$$

$$i = 2, \ldots, n, \;\; j = i+1, \ldots, n;$$

$$h_{Aii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} h_{ik}^2}, \;\; i = 2, \ldots, n; \tag{4}$$

$$h_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} h_{ik} h_{jk} \right) \Big/ h_{Aii}, \;\; j > i.$$

Once all elements of the matrices are obtained, we solve systems with triangular matrices:

$$U_A x = L_A^{-1} b, \tag{5}$$

$$H_A y = b, \;\; H_A^\top x = y. \tag{6}$$

The number of operations is estimated as follows: for LU-decomposition, $\frac{2}{3}n^3$ to obtain a triangular matrix and $\frac{n^2}{2}$ for (5); for Cholesky decomposition, $\frac{1}{3}\, n^3$ and n^2 according to (6).

1. TRUNCATION AND INCOMPLETE INNER PRODUCT

In both methods of triangular decomposition, the most widely used operation is the inner product of two vectors. A generally accepted method of calculating this product is to accumulate the products of two numbers which have $2t$ digits the number of digits for a t-digit mantissa. This technique allows decreasing the rounding errors in comparison with the accumulation of t-digit products. In this case, the operations of assigning, dividing, and taking square roots are performed with the most significant t digits of the resulting sum.

To use the discarded digits of the accumulated sum of products in the computation process, let us introduce the following operation: divide a $2t$-digit number by two and place these numbers in two memory cells, each containing t digits. The first number consists of $(t - \tau)$ significant digits of the initial sum and is supplemented to t by zeroes on the right (τ is a positive integer smaller than t). The second number consists of the τ digits truncated from the first number, and it is supplemented on the right by the sum digits beginning with the $(t+1)$th digit. Let $[\cdot]_b$ and $[\cdot]_r$ denote the first and second numbers, respectively. Then for the inner product of two n-dimensional vectors we have

$$s = \left[\sum_{i=1}^{n} a_i b_i \right]_b + \left[\sum_{i=1}^{n} a_i b_i \right]_r . \tag{7}$$

If t-digit products are accumulated instead of $2t$-digit ones (no double accumulation mode is available), the first number contains $t-\tau$ significant digits of the inner product, and the second one, the remaining τ digits.

The inner product s is called complete for any τ in the range $t > \tau \geq 0$. Let the first term in (7) for $\tau > 0$ be called an incomplete inner product and the second one, its complement, and let the operation for obtaining them be called truncation. The signs and orders of both numbers are the same as those of s. In magnitude, s is larger than the incomplete product which, in turn, is larger than the complement. As τ increases, the incomplete product decreases and the complement increases.

We use the incomplete inner product to calculate one diagonal element of the triangular matrices U and H:

$$u_{Aii} = a_{ii} - \left[\sum_{k=1}^{i-1} l_{ik}u_{kj}\right]_b - \left[\sum_{k=1}^{i-1} l_{ik}u_{ki}\right]_r, \qquad h_{Aii} = \sqrt{a_{ii} - \left[\sum_{k=1}^{i-1} h_{ik}^2\right]_b - \left[\sum_{k=1}^{i-1} h_{ik}^2\right]_r}.$$

If we take the following new values for the diagonal elements:

$$u_{ii} = a_{ii} - \left[\sum_{k=1}^{i-1} l_{ik}u_{ki}\right]_b, \qquad h_{ii} = \sqrt{a_{ii} - \left[\sum_{k=1}^{i-1} h_{ik}^2\right]_b},$$

we have

$$u_{ii} = u_{Aii} + \left[\sum_{k=1}^{i-1} l_{ik}u_{kj}\right]_r, \qquad h_{ii}^2 = h_{Aii}^2 + \left[\sum_{k=1}^{i-1} h_{ik}^2\right]_r.$$

The off-diagonal elements of the triangular matrices are calculated, according to (3) and (4), with the new values, u_{ii} and h_{ii}.

In positive definite matrices the diagonal elements of the corresponding triangular matrices are positive. Therefore, u_{ii} and h_{ii} are larger than 0, and larger than u_{Aii} and h_{Aii}, respectively.

To determine the structure of the matrix N, we use a method of inverse analysis of LU-decomposition errors [2]. Its main result is the following expression:

$$LU = A + N,$$

where N is a dense matrix whose elements are the errors of rounding of the results of calculation of the elements of the triangular matrices to t digits.

Truncation will be considered as rounding of a number to $(t-\tau)$ digits without the commonly used division of the resulting error by two. Assume that the only truncated element is u_{Aii}. In this case the error is $\left[\sum_{k=1}^{i-1} l_{ik}u_{ki}\right]_r$, whereas all other elements of the triangular matrices are calculated exactly (the ordinary rounding errors are not considered).

Let N^1 denote the error matrix for one truncation. It has only one nonzero element n_{ii}^1 whose value is equal to the number consisting of the truncated digits. In this case the elements of the triangular matrices below the ith row will change in comparison to the elements of the matrix U_A. Specifically, the elements of the matrix L, according to (3), will decrease, whereas the diagonal elements participating in the calculation will increase.

Assume that after the truncation for the element u_{ii} the truncation for the jth diagonal element $(j > i)$ was made in the process of decomposition. This means that the matrix $A + N^1$ has incomplete decomposition, whereas the matrix of decomposition errros N^2 has one nonzero element n_{jj}^2 calculated for the matrix $A + N^1$ and equal to $\left[\sum_{k=1}^{j-1} l_{jk}u_{kj}\right]_r$.

After k truncations we have

$$LU = A + N^1 + N^2 + \cdots + N^k.$$

Similar reasoning can be used for the Cholesky method, with the only difference that an error appears in calculating the square of a diagonal term [5, pp. 176, 177]. Thus, Eq. (2), where the decomposition error matrix N has nonzero terms only in the main diagonal, holds for decompositions with the truncation operation.

The largest number of elements in N is $(n-1)$; n_{11} is always 0. The quantity n_{ii} is part of the corresponding inner product and ultimately depends on the elements of the matrix A.

If there is no double accumulation, all $n_{ii} = 0$ with $\tau = 0$, and M, L, U, and H coincide with the matrices A, L_A, U_A, and H_A, respectively. If truncation is implemented as a software operation, τ can be changed in the calculations by increasing its value for small diagonal elements of the triangular matrix.

The triangular matrices L, U, and H obtained in the process of incomplete decomposition make it possible to calculate a vector $\tilde{x}$ which can be considered an approximate solution of the system (1):

$$M\tilde{x} = b. \tag{8}$$

To further discuss the truncation effect, we will use a condition number of a matrix calculated as the product of its norm by the norm of the inverse matrix, and as the ratio of the maximum eigenvalue of the matrix to the minimum one. Decreasing the condition number of a matrix is one of the major reasons for preconditioning.

It is generally assumed that in decomposing an ill-conditioned matrix into triangular factors the upper triangular matrix is ill-conditioned and the lower triangular one is well-conditioned [1]. The condition number of the matrix M under LU-decomposition is as follows:

$$\operatorname{cond}(M) \le \operatorname{cond}(L)\operatorname{cond}(U). \tag{9}$$

Since the eigenvalues of the upper triangular positive definite matrix U are the diagonal terms, increasing all diagonal terms, or for least the smallest ones, will decrease the condition number of U in comparison to that of U_A, and that of M in comparison to that of A. This resoning can be used for the Cholesky decomposition as well.

2. EXACT SOLUTION

To obtain an exact solution by using truncated numbers, in (1) we use left preconditioning:

$$M^{-1}(M - N)x = M^{-1}b. \tag{10}$$

Denoting $Z = (I - M^{-1}N)$ and using (8), we write (10) in the following form:

$$Zx = \tilde{x}, \tag{11}$$

where I is a unit matrix. Z is not singular and has the following form:

$$Z = \begin{vmatrix} 1 & 0 & -y_{11} & \cdots -y_{1k} \\ 0 & 1 & -y_{21} & \cdots -y_{2k} \\ 0 & 0 & 1 - y_{31} & \cdots -y_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -y_{n1} & \cdots 1 - y_{nk} \end{vmatrix}.$$

The columns y_l form a dense rectangular matrix, which can be obtained by multiplying M^{-1} by N and rearranging the columns. First, the system (8) is solved. This has an operation count typical for triangular decompositions. The second stage in finding a solution to (1) is solving the system (11). Its implementation requires additional operations. They are used to form the matrix Z and determine the vector x.

When finding the vectors y_k, instead of inverting the matrix M (since it does not exist in explicit form), it is sufficient to solve k systems of LAEs with triangular matrices obtained for the first stage and columns with a single nonzero element in the right-hand side:

$$My_l = n_l, \quad l = 1, \ldots, k.$$

For this we need $k\frac{n^2}{2}$ operations in the LU-decomposition and twice as many operations in the Cholesky decomposition.

In finding the vector x we use a standard LU-decomposition. When (10) is solved by an iterative method for large cond (A), one cannot guarantee that the condition $||M^{-1}N|| < 1$, which is necessary for convergence of the process, is satisfied. We need $\frac{2}{3}k^3$ operations to bring the matrix Z to triangular form, and additional $\frac{k^2}{2}$ operations to solve the thus obtained triangular system. On the whole, the operation count for the second stage is $\frac{2}{3}k^3 + O(n^2) + O(k^2)$ operations, and for $k < n$ it is smaller than that for the triangular decomposition.

Let us estimate the ratio of the condition numbers of the matrices M and Z. Since it follows from (10) that

$$A^{-1} = Z^{-1}M^{-1}, \tag{12}$$

we have

$$\mathrm{cond}\,(Z) \geq \frac{\mathrm{cond}\,(A)}{\mathrm{cond}\,(M)}. \tag{13}$$

Since cond (A) does not depend on truncations, a decrease in cond (M) leads to an increase in cond (Z).

There are several methods of truncation: for instance, the diagonal elements of triangular matrices are controlled as follows: if they successively decrease, a truncation is used for the next element. For the Cholesky method, truncation can be used when a radicand becomes less than zero.

3. RESULTS OF EXPERIMENTS

These computational experiments were performed for LAE systems with well-known ill-conditioned matrices in double format with $t = 17$. An LU-decomposition was made using the Crout–Doolittle algorithm [2]. That is, this $HH^\top$-decomposition was made using a “left to right, top to bottom” scheme. The condition numbers were calculated as the products of the Euclidean norms of corresponding matrices, and the norms of vectors, as ∞-norms.

The following SLAEs were used for the LU-decomposition:

(a) A system of two equations known as the Wilkinson test problem [9]:

$$A = \begin{vmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{vmatrix}, \qquad b = \begin{vmatrix} 0.217 \\ 0.254 \end{vmatrix}. \tag{14}$$

The exact solution of the system: $x_1 = 1$, $x_2 = -1$; the condition number of the matrix: 10^6. The norm of A^{-1} is large because the A-rows are almost linearly dependent. Actually, the ratio a_{11}/a_{21} is equal to a_{12}/a_{22} to the fifth decimal place. The results of the calculations are presented in Table 1.

Table 1. Results of truncation for the Wilkinson problem

τ	n_{22}	u_{22}	cond (M)	cond (Z)	$\|r_x\|$
0	0	$1.3 \cdot 10^{-6}$	10^6	2	$1.2 \cdot 10^{-11}$
13	$9.9 \cdot 10^{-5}$	$9.9 \cdot 10^{-5}$	$2.2 \cdot 10^4$	77	$4.0 \cdot 10^{-13}$
15	$9.0 \cdot 10^{-3}$	$9.0 \cdot 10^{-3}$	241	$7.0 \cdot 10^3$	$1.4 \cdot 10^{-16}$
16	$5.8 \cdot 10^{-2}$	$5.9 \cdot 10^{-2}$	50	$6.0 \cdot 10^5$	$1.4 \cdot 10^{-16}$

Table 2. Results of truncation for LU-decomposition ($n = 8$, $k = 3$)

τ	$\|N\|$	$\max L_{ij}$	u_{88}	cond (M)	cond $V(Z)$	$\|r_x\|$
0	0	5.65	$5.7 \cdot 10^{-9}$	$1.7 \cdot 10^{10}$	8	$6.5 \cdot 10^{-8}$
13	$8.8 \cdot 10^{-5}$	3.33	$1.6 \cdot 10^{-5}$	$7.3 \cdot 10^5$	$1.1 \cdot 10^6$	$9.8 \cdot 10^{-12}$
15	$1.1 \cdot 10^{-3}$	3.71	$6.1 \cdot 10^{-4}$	$1.26 \cdot 10^5$	$7.1 \cdot 10^6$	$2.0 \cdot 10^{-12}$
16	$1.1 \cdot 10^{-2}$	3.71	$6 \cdot 10^{-3}$	$1.2 \cdot 10^5$	$1.6 \cdot 10^8$	$4.3 \cdot 10^{-13}$

It follows from Table 1 that the diagonal element u_{22} of the triangular matrix incereases with increasing τ. Cond (M) decreases, and cond (Z), according to (13), increases. The residual of the solution $r = Ax - b$, for the same number of significant numbers as for $\tau = 0$, decreases.

(b) A SLAE with a Hilbert matrix, which is symmetric positive definite and ill-conditioned, with a condition number increasing abruptly with increasing n [2]. Systems with such a matrix are often used to verify the effeciency of computational algorithms. Its elements are calculated by the formula $a_{ij} = 1/(i + j - 1)$. The ratio of any elements of the kth and lth rows is $(l - 1 + j)/((k - l + j))$: even for small k and l it depends but slightly on j. As a result, the elements of the triangular matrix U become very small with increasing subscripts, and the elements of the inverse matrix, very large. The order of the matrix was taken equal to 8, and the number of digits in the representation of the coefficients of A and b was 10. The vector of free terms was chosen so that the unit vector is a solution to the system. Truncation was used under the condition $a_{ii} < a_{i-1,i-1} < a_{i-2,i-2}$. In our case this takes place for $i = 5, 6, 8$. Table 2 presents the norm of the matrix N, the maximal element of the triangular matrix L, the last diagonal element of the matrix U, the condition numbers of the matrices M and Z, and the norm of the residual vector.

It follows from Table 2 that as τ increases, the elements of the matrix L decrease considerably and the diagonal elements of the matrix U increase beginning from the fifth one (the last diagonal element is the smallest). The condition number of the matrix M decreases, and that of the matrix Z increases. As in the previous experiment, an admissible relation between cond (M) and cond (Z) is reached for $\tau = 15$. The norm of the solution residual decreases with increasing τ.

(c) The same system with a Hilbert matrix and the unit vector as a solution for $n = 8$ was used for the Cholesky decomposition. The number of digits in the representation of the matrix and vector coefficients of the initial system was taken equal to 8. This choice was made because when the number of digits is larger than 8 the computation process terminates normally. Abnormal termination when the number of digits is equal to 16 (almost complete word length) takes place only for $n = 13$. This confirms the statement [10] that the rounding errors for ill-conditioned matrices play a lesser role than the errors in representing the matrix elements. When solving by a standard method, the calculations terminate for h_{88}. The method of truncation used for the LU-decomposition yielded here similar results: an increase in the diagonal elements of the triangular matrix H, a decrease in cond (M), and an increase in cond(H) under normal termination of the calculations. The number of truncations was three.

Table 3 presents the results of solving two SLAEs with a Hilbert matrix for $n = 8$ and $n = 10$ (for $n = 10$ cond$(A) = 10^{13}$) by the Cholesky method. The inner product was truncated for the diagonal element for which the radicand was negative. At $n = 8$ truncation was made once, and for $n = 10$, three times. The condition numbers of the matrices M and Z were smaller than cond (A).

Table 3. Results of truncation for Cholesky decomposition ($\tau = 15$)

n	Termination of calculations	k	Truncation at	n_{ii}	cond (M)	cond (Z)
8	h_{88}	1	h_{88}	$6.7 \cdot 10^{-4}$	$8 \cdot 10^{8}$	$6.7 \cdot 10^{5}$
10	h_{88}	3	h_{88}; h_{99}; $h_{10\,10}$	$6.7 \cdot 10^{-4}$; $8.2 \cdot 10^{-4}$; $6.3 \cdot 10^{-4}$	$6.7 \cdot 10^{8}$	$3.5 \cdot 10^{8}$

Let us find out whether the algorithm being proposed can be simplified if (instead of truncation) small diagonal elements of the triangular matrix are replaced by a prespecified number. In library [4] this number, ε, is equal to 10^{-13} for nonsymmetric matrices and 10^{-8} for symmetric ones. For large sparse matrices, for which this library is designed, the resulting solution error is assumed to be insignificant; otherwise it is proposed to use iterative correction.

In our case ε must be large enough to decrease the condition number of a matrix. However, it is difficult to choose its value a priori. For instance, for the Wilkinson matrix, as is clear from Table 1, we can take $\varepsilon = 10^{-2}$ according to u_{22}. However, the same value for the Hilbert matrix will lead to a too large cond (Z). This is because the diagonal elements of the triangular matrix U in the Wilkinson system are larger than those of the Hilbert matrix and, as a result, the numbers of truncated digits are different for the same τ. In addition, if the true value of a diagonal element of the triangular matrix cannot be calculated (as in the case of a negative radicand in the Cholesky decomposition), it is also impossible to obtain the corresponding value of the matrix N, which is necessary for calculating the exact value.

CONCLUSIONS

In this paper, the following results have been obtained: It has been shown that truncation of the least significant digits of the innear product in the LU-decomposition and Cholesky decomposition increases the diagonal elements of the triangular matrices. This improves the stability of the decomposition process. In particular, it prevents termination in the calculation process of the Cholesky decomposition. Solving the initial equation with an ill-conditioned matrix of coefficients is reduced to successively solving two systems of equations whose matrices are better conditioned than that of the initial system. The size of the second system can be much smaller than that of the initial one. The number of needed additional operations is not as large as the number of operations in the standard triangular decomposition.

REFERENCES

1. Golub, G.H. and Van Loan, C.F., *Matrix Computations*, 3d ed., Johns Hopkins University Press, 1996.
2. Forsythe, G.E. and Moler, C.B., *Computer Solution of Linear Algebraic Systems*, N.J.: Prentice-Hall, Englewood Cliffs, 1967.
3. Il'in, V.P., *Metody nepolnoi faktorizatsii dlya resheniya algebraicheskikh sistem* (Incomplete Factorization Methods for Solving Algebraic Systems), Moscow: Fizmatlit, 1995.
4. mkl-pardiso-pivot, URL:https://software.intel.com/en-us/mkl-developer-reference-fortran-mkl-pardiso-pivot.
5. Voevodin, V.V. and Kuznetsov, Yu.A., *Matritsy i vychisleniya* (Matrices and Calculations), Moscow: Nauka, 1984.
6. Saad, Y., *Iterative Methods for Sparse Linear Systems*, N.Y.: PWS Publ., 1996.
7. Stone, H.L., Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations, *SIAM J. Sci. Num. An.*, 1968, no. 3, pp. 530–558.
8. Draper, N.R. and Smith, H., *Applied Regression Analysis*, 3rd ed., John Wiley and Sons, 1998.
9. Manichev, V.B., Glazkova, V.V., Kozhevnikov, D.Yu., Kiryanov, D.A., and Sakharov, M.K., Solution of Systems of Linear Algebraic Equations with Double Precision in C, *Vestnik BMSTU, Ser. "Instrumentation,"* 2011, no. 4, pp. 25–35.
10. Wilkinson, J.H., *The Algebraic Eigenvalue Problem*, Oxford: Clarendon Press, 1965.