Numerical Methods and Approximation Theary
Novi Sad, September, 4–6, 1985

GAUSSIAN ELIMINATION FOR POSITIVE DEFINITE MATRICES

Zvonimir Bohte, Marko Petkovšek

ABSTRACT:

In this paper a sufficient condition for the non-failure of Gaussian elimination for positive definite matrices in standard floating-point arithmetic is proved. The result is similar to Wilkinson's for the Cholesky factorization.

GAUSOVA ELIMINACIJA ZA POZITIVNO DEFINITNE MATRICE. *U ovom radu je dokazan dovoljan uslov za izvodljivost Gausove eliminacije za pozitivno definitne matrice za aritmetiku u pomičnom zarezu. Rezultat je sličan Wilkinsonovom za faktorizaciju Choleskoga.*

1. INTRODUCTION

Let A be a real symmetric positive definite matrix of order n. Gaussian elimination for the solution of the system of linear equations

$$Ax = d$$

yields a set of equivalent systems

$$A^{(r)}x = d^{(r)}, \quad r = 1, \dots, n$$

where $A^{(1)} = A$, $d^{(1)} = d$ and $A^{(n)}$ is an upper triangular matrix. The matrix $A^{(r)}$ has the following block structure

$$(1) \qquad A^{(r)} = \begin{bmatrix} U_r & X_r \\ \emptyset & A_r \end{bmatrix}$$

where U_r is an upper triangular matrix of order r-1 and A_r a square matrix of order n-r+1.

The matrix A_r is symmetric by construction as we calculate only elements in the upper triangle, for instance. Wilkinson [2] proved that the matrix A_r is also positive definite and that there is no growth of the elements. This is true only for the exact arithmetic.

The presence of rounding errors may destroy the original positive definiteness. Therefore, to ensure the non-failure of the method it is necessary to require more than just that the smallest eigenvalue of the ori-

ginal matrix is positive.

In the analysis of rounding errors we shall use the equation

(2) $$fl(x*y) = (x*y)(1 + e), \quad |e| \leq u$$

where x and y are any standard floating point numbers and fl(x*y) denotes the computed result of any of the four arithmetic operations. We shall suppose that the relative error of an arithmetic operation is bounded by the unit rounding error which is normally

$$u = b^{1-t}/2 \quad \text{(for rounding)}$$
$$= b^{1-t} \quad \text{(for chopping)}$$

where t is the length of the mantissa in the base b (usually 2 or 10). It is of course assumed also that during the computation no overflow or underflow occurs.

In the following we shall leave out all the work with the right-hand sides.

2. THE ALGORITHM AND ERROR ANALYSIS

We denote the current calculated matrix at the r-th step by $B^{(r)}$. It has the same block structure as the matrix (1)

$$B^{(r)} = \begin{bmatrix} V_r & Y_r \\ \emptyset & B_r \end{bmatrix}$$

It is assumed that the matrix $A = B^{(1)}$ is the matrix stored in the computer. The matrix B_r is symmetric by construction. We shall calculate only the elements in its upper triangle.

The algorithm for the calculation of the upper triangular matrix $B^{(n)}$ is as follows:

$r = 1, \ldots, n-1$:

$i = r+1, \ldots, n$:

(3) $$m_{ir} = fl(b_{ri}^{(r)}/b_{rr}^{(r)})$$

$k = i, \ldots, n$:

(4) $$b_{ik}^{(r+1)} = fl(b_{ik}^{(r)} - fl(m_{ir}b_{rk}^{(r)}))$$

All essential ideas of the following error analysis are due to Wilkinson [2].

Using (2) in (3) and (4) we have

(5) $$m_{ir} = q_{ir}(1 + x_{ir}), \quad i = r+1, \ldots, n$$

and

(6) $$b_{ik}^{(r+1)} = (b_{ik}^{(r)} - m_{ir}b_{rk}^{(r)}(1 + y_{ik}^{(r)}))(1 + z_{ik}^{(r)}), \quad i = r+1, \ldots, n; \quad k = i, \ldots, n$$

where

(7) $$q_{ir} = b_{ri}^{(r)}/b_{rr}^{(r)}$$

and

(8) $$|x_{ir}|, \ |y_{ik}^{(r)}|, \ |z_{ik}^{(r)}| \leq u$$

We can write the equation (6) in the form

(9) $$b_{ik}^{(r+1)} = c_{ik}^{(r+1)} + d_{ik}^{(r)}$$

where

(10) $$c_{ik}^{(r+1)} = b_{ik}^{(r)} - q_{ir}b_{rk}^{(r)}$$

and

(11) $$d_{ik}^{(r)} = -q_{ir}b_{rk}^{(r)}(x_{ir} + y_{ik}^{(r)} + x_{ir}y_{ik}^{(r)})(1 + z_{ik}^{(r)}) + c_{ik}^{(r+1)}z_{ik}^{(r)}$$

The bound for this error will be given below.

3. THE THEOREM

Let us denote the smallest eigenvalue of a symmetric matrix A by $\lambda(A)$ and define

(12) $$h_r = \max_{r \leq i,k \leq n} |b_{ik}^{(r)}|, \quad r = 1, \ldots, n$$

Then the following theorem holds.

THEOREM. Let A be a symmetric positive definite matrix of order n with its smallest eigenvalue satisfying

(13) $$\lambda(A) > cn(n-1)h_1u$$

where $c = (4 + 3u + u^2)/2$.

Then the following is true for $r = 1, \ldots, n$:

(i) the matrix B_r is symmetric positive definite and

$$\lambda(B_r) > c(n-r+1)(n-r)h_1u$$

(ii) $h_r \leq h_1$

PROOF. We shall prove the theorem by the mathematical induction with respect to r.

Let r = 1. Then (i) coincides with (13) as $A = B^{(1)} = B_1$ and (ii) is trivial.

Let $1 \leq r \leq n-1$ and suppose that (i) and (ii) hold for this r. Denote by C_{r+1} the square matrix of order n-r with the elements (10) for $r+1 \leq i,k \leq n$. Then in view of (9)

$$B_{r+1} = C_{r+1} + D_r$$

where the elements of D_r are defined by (11) for $r+1 \leq i,k \leq n$. Since B_r is symmetric by (i), C_{r+1} is also symmetric (see [2]). The matrix B_{r+1} is symmetric by construction, hence D_r is symmetric as well. By a corollary of the minimax theorem (see [3])

$$\lambda(B_{r+1}) \geq \lambda(C_{r+1}) + \lambda(D_r) \geq \lambda(C_{r+1}) - \|D_r\| \tag{14}$$

The matrix B_r can be written in a block form

$$B_r = \begin{bmatrix} b_{rr}^{(r)} & b_r^T \\ b_r & E_r \end{bmatrix} \tag{15}$$

where $b_r^T = [b_{r,r+1}^{(r)}, \ldots, b_{rn}^{(r)}]$ and E_r is a square matrix of order n-r. Then the matrix C_{r+1} with the elements (10) can be written as

$$C_{r+1} = E_r - q_r b_r^T \tag{16}$$

where $q_r = (1/b_{rr}^{(r)})b_r$ from (7). The matrix B_r is positive definite by hypothesis (i) and therefore $b_{rr}^{(r)} > 0$. Let x_r be any vector of dimension n-r. Then from (16) it follows

$$x_r^T C_{r+1} x_r = x_r^T E_r x_r - (b_r^T x_r)^2 / b_{rr}^{(r)} \tag{17}$$

Let $y_r^T = [a, x_r^T]$ where a is any real number. Then by block multiplication we obtain from (15)

$$y_r^T B_r y_r = a^2 b_{rr}^{(r)} + 2a(b_r^T x_r) + x_r^T E_r x_r$$

which can be expressed by means of (17) as

$$y_r^T B_r y_r = x_r^T C_{r+1} x_r + (a b_{rr}^{(r)} + b_r^T x_r)^2 / b_{rr}^{(r)} \tag{18}$$

If we now take x_r to be any normalized eigenvector corresponding to the smallest eigenvalue $\lambda(C_{r+1})$ and $a = -(b_r^T x_r)/b_{rr}^{(r)}$ then it follows from (18)

$$y_r^T B_r y_r = \lambda(C_{r+1})$$

But

$$y_r^T B_r y_r \geq \lambda(B_r)(y_r^T y_r) = \lambda(B_r)(a^2 + 1) \geq \lambda(B_r)$$

and therefore

$$\lambda(C_{r+1}) \geq \lambda(B_r)$$

Thus it follows from (14)

(19) $\lambda(B_{r+1}) \geq \lambda(B_r) - \|D_r\|$

To get an upper bound for $\|D_r\|$ we observe that $(b_{ik}^{(r)})^2 < b_{ii}^{(r)} b_{kk}^{(r)}$ since B_r is positive definite by hypothesis (i) and therefore it follows from (7) and (12)

$$|q_{ir} b_{rk}^{(r)}| = |b_{ri}^{(r)} b_{rk}^{(r)} / b_{rr}^{(r)}| < (b_{rr}^{(r)} b_{ii}^{(r)} b_{rr}^{(r)} b_{kk}^{(r)})^{1/2} / b_{rr}^{(r)} =$$
$$= (b_{ii}^{(r)} b_{kk}^{(r)})^{1/2} \leq h_r$$

Using this inequality we obtain from (11), (8), (10) and (12) the bound

$$|d_{ik}^{(r)}| \leq (4 + 3u + u^2) h_r u = 2ch_r u \ , \quad r+1 \leq i,k \leq n$$

For any of the standard norms

$$\|D_r\|_p \leq 2c(n-r)h_r u \ , \quad p = 1, 2, \infty, E$$

Therefore for any such norm it follows from (19) and hypotheses (i) and (ii)

$$\lambda(B_{r+1}) \geq c(n-r+1)(n-r)h_1 u - 2c(n-r)h_r u \geq c(n-r)(n-r-1)h_1 u$$

which proves (i).

To prove (ii) note that $h_r = \max b_{ii}^{(r)}$, $r \leq i \leq n$, since B_r is positive definite. From (4), (5) and (7) we have

$$b_{ii}^{(r+1)} = fl(b_{ii}^{(r)} - (b_{ri}^{(r)})^2 (1 + x_{ir})(1 + y_{ii}^{(r)}) / b_{rr}^{(r)})$$

For any base $b \geq 2$ and length $t \geq 1$ the unit rounding error satisfies $u \leq 1$ and therefore $1 + x_{ir}$ and $1 + y_{ii}^{(r)}$ are because of (8) nonnegative. It can be shown (see [1]) that when subtracting a nonnegative number from a positive one the computed result cannot exceed the minuend. Hence

$$b_{ii}^{(r+1)} \leq b_{ii}^{(r)}$$

and because of positive definitness and (ii)

(20) $h_{r+1} \leq h_r \leq h_1$

which proves (ii). The theorem is proved.

The result is similar to the one obtained by Wilkinson in [4] for Cholesky factorization of the matrix A.

4. CONCLUSIONS

The immediate consequence of (20) is the assertion that there is no growth of the computed elements during the Gaussian elimination, provided the assumptions of the theorem hold. Such assertion for exact arithmetic was proved by Wilkinson in [2].

Suppose we have a symmetric positive definite matrix A of order n. In the following table the precision required by the theorem when rounding in base 10 as a function of the quotient $h_1/\lambda(A)$ for some typical n are given.

	minimal length of mantissa		
$h_1/\lambda(A)$	n = 5	n = 10	n = 100
1	3	3	5
1·1	3	3	6
1·3	3	4	6
10	4	4	6
100	5	5	7
1000	6	6	8

The results are rather pessimistic.

We can express the assumptions of the theorem in terms of the condition number. Since $\lambda(A) = 1/\|A^{-1}\|_2$ and $\|A\|_2 \geq h_1$ we have directly

COROLLARY. If A is a symmetric positive definite matrix and

$$\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 < 1/(cn(n-1)u)$$

then the conclusions of the theorem are valid.

REFERENCES:

1. BOHTE Z.: Fundamentals of finite arithmetic (in manuscript).

2. WILKINSON J. H.: *Error analysis of direct methods of matrix inversion.* J. ACM 8 (1961), 281 - 330.

3. WILKINSON J. H.: The algebraic eigenvalue problem. Clarendon Press. Oxford 1965.

4. WILKINSON J. H.: *A priori error analysis of algebraic processes.* International Congress of Mathematicians, Moskva 1966. Izdatel'stvo "MIR". Moskva 1968, 629 - 640.