Aplikace matematiky

Pavol Poliak; Mária Postulková; Romana Vyhnanská

Solution of linear algebraic equations with 3-diagonal ill-conditioned system matrix

Aplikace matematiky, Vol. 15 (1970), No. 4, 227–234

Persistent URL: http://dml.cz/dmlcz/103291

SOLUTION OF LINEAR ALGEBRAIC EQUATIONS WITH 3-DIAGONAL ILL-CONDITIONED SYSTEM MATRIX

PAVOL POLIAK, MÁRIA POSTULKOVÁ, ROMANA VYHNANSKÁ

(Received January 26, 1968)

The paper introduces certain methods for the solution of linear algebraic equations with ill-conditioned system matrices and compars them with the Gaussian method. We assume the system matrices to be of a special type, the so called *band matrices*, 3-diagonal in our case. Band matrices are those having non-zero elements only in the main diagonal and in a few ones around it. The algorithms for the solution of linear equation by classical iterative methods and by the Gaussian method for such class of matrices are introduced, together with the attained results, in [1].

Dealing with the solution of systems in which the matrix of the system is diversely conditioned, we come to the conclusion that:

a) Even for systems with ill-conditioned matrices the Gaussian method appears to be the most advantageous;

b) The iterative methods for ill-conditioned systems require unproportionally longer time for the calculation of the solution than does the Gaussian method, while the accuracy of the obtained solution is not better.

When solving systems of linear algebraic equations with ill-conditioned system matrices by means of the Gaussian method it was found that the solution differs greatly from the accurate solution. When solving, for instance, a system of ten equations, for the condition (1*) $P = 10^7$, the maximal absolute error of solution was up to $3{\cdot}1 \,.\, 10^2$. This led to the suppression of these methods for ill-conditioned linear systems.

Let us have a system of linear equations

$$AX = F \tag{1}$$

where A – is the matrix of the system, 3-diagonal,

X – the solution vector,

F – the right-hand side vector.

The individual iterations are marked $X^{(k)}$ for $k = 0, 1, \ldots, X^{(0)}$ being the initial approximation of the solution. We shall further use the following notation:

(1*) $P = \max_i |\lambda_i| / \min_i |\lambda_i|$
(where λ_i is the i^{th} eigenvalue of matrix A) being the state of condition of matrix A, defined by Todd's number

$m = \min_i \lambda_i$

$M = \max_i \lambda_i$

ε = the prescribed difference of two subsequent iterations.

One of the methods of solving the system (1) requiring that matrix A be symmetric and positive definite consists in calculating the individual approximations [4] according to the relation

$$(2) \qquad X^{(k)} = X^{(k-1)} + \left(\frac{1 - \sqrt{B}}{1 + \sqrt{B}}\right)^{10} (X^{(k-1)} - X^{(k-2)}) + \\ + \frac{4}{M(1 + \sqrt{B})^{10}} \sum_{r=0}^{4} c_r A^{4-r}(AX^{(k-1)} - F)$$

where $X^{(1)}$ and $X^{(2)}$ are arbitrary initial approximations of the solution of system (1), $B = 1/P$ and coefficients c_r are calculated according to relations

$$c_0 = -256/M^4, \quad c_1 = 640(1 + B)/M^3, \quad c_2 = -(560B^2 + 1440B + 560)/M^2$$

$$c_3 = (200B^3 + 1080B^2 + 1080B + 200)/M,$$

$$c_4 = -(25B^4 + 300B^3 + 630B^2 + 300B + 25).$$

To carry out the calculation according to (2) by an automatic computer it is necessary to have $25n - 35$ storage units.

Another way of obtaining the solution of the system (1) is that of succesive approximations according to the relation [3]

$$(3) \qquad X^{(k)} = (a_1 E + b_1 A + b_2 A^2) X^{(k-1)} + a_2 X^{(k-2)} - (b_2 A - b_1 E) F$$

where $X^{(0)}, X^{(1)}$ are arbitrary initial approximations of the solution of the system and

$$a_1 = 1 + \left(\frac{1 - \sqrt{B}}{1 + \sqrt{B}}\right)^4, \quad a_2 = -\left(\frac{1 - \sqrt{B}}{1 + \sqrt{B}}\right)^4,$$

$$b_1 = -\frac{16(1 + B)}{M(1 + \sqrt{B})}, \quad b_2 = \frac{16}{M^2(1 + \sqrt{B})^4}.$$

A sufficient condition for the convergence of the method is

$$\sum_{j=1}^{n} |a_{ij}| \leqq \mu < 1 \quad \text{for} \quad i = 1, 2, \ldots, n ,$$

$$\sum_{i=1}^{n} |a_{ij}| \leqq \nu < 1 \quad \text{for} \quad j = 1, 2, \ldots, n .$$

The calculation on an automatic computer required $15n$ storage units.

The iterative calculation of the inverse matrix by Hotteling's method with respect to the given matrix forms the basis of a further method [2]. The resulting solution of the system is obtained from the relation

$$X = GF + HX^{(0)} \tag{4}$$

where $G \sim A^{-1}$, $H = E - GA$.

The matrix G, which is very close to the inverse matrix, is calculated by an iterative process according to the relation

$$G_r = G_{r-1} + H_{r-1}G_{r-1} ,$$

$$H_r = E - G_r A .$$

The end of the iterative process is given by a parameter δ indicating the accuracy necessary for calculating the inverse matrix. (Let $\|H_r\|$ be can arbitrary norm of matrix $H_r = (h_{ij})$, then the iterative process is terminated, i.e. $G_r = G \sim A^{-1}$, if $\|H_r\| < \delta$. We considered the norm $\|H_r\| = n \max |h_{ij}|$.)

The iterative process just described converges if the eigenvalues of matrix H_0, where $H_0 = E - G_0 A$, G_0 – the first approximation of the inverse matrix are smaller than one. This gives rise to the problem of finding such a matrix G_0 that the eigenvalues of matrix H_0 would comply with the abovementioned requirement. It can be shown [2] that the eigenvalues of matrix H_0 fulfil the requirement, hence the iterative process converges, if, assuming A to be 3-diagonal matrix, the matrix elements $G_r = (g_{ij})$ have the following form:

for $j > i + 1$, $g_{ij} = 0$

$j \leqq i + 1$, $g_{ij} = g_{ij}^{(i+1)}$

where the following relations hold for $j = 1$:

$$g_{ij}^{(k)} = g_{ij}^{(k-1)} - a_{ki}\left(a_{k,k-1}g_{k-1,j}^{(k-1)} + a_{kk}g_{kj}^{(k-1)}\right)$$

for $k = 2, 3, \ldots, n$ $(i = k - 1, k, k + 1) \wedge (i \leqq n)$,

$$g_{ij}^{(k)} = 0$$

for $(i = k + 2) \wedge (i \leqq n)$, while $g_{11} = a_{11}$; $g_{21} = a_{12}$; $g_{31} = 0$,

for $i = n$ it holds $g_{nj} = g_{nj}^{(n)}$.

For $j > 1$ the relations

$$g_{ij}^{(k)} = a_{ij}$$

hold for $j = 2, 3, \ldots, n$, $k = j$, $(i = k - 1, k, k + 1) \wedge (i \leqq n)$,

$$g_{ij}^{(k)} = 0$$

for $(i = k + 2) \wedge (i \leqq n)$,

$$g_{ij}^{(k)} = g_{ij}^{(k-1)} - a_{ki}\left(a_{k,k-1}g_{k-1,j}^{(k-1)} + a_{kk}g_{kj}^{(k-1)}\right)$$

for $k = j + 1, j + 2, \ldots, n$, $(i = k - 1, k, k + 1) \wedge (i \leqq n)$,

$$g_{ij}^{(k)} = 0$$

for $(i = k + 2) \wedge (i \leqq n)$ and

$$g_{nj} = g_{nj}^{(n)}$$

The calculation on an automatic computer requires $3n^2 + 6n - 2$ storage units.

In applying the methods referred to, experiments were carried out on testing matrices [1] of the form

$$A = \begin{bmatrix} a, & -1, & & & \\ -1, & a, & -1, & & \\ \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \\ & -1, & a, & -1 & \\ & & -1, & a & \end{bmatrix}$$

in which the state of condition P of matrix A ranged over the basis values of the parameter and

$$|a| = 2\cos\frac{\pi}{n+1}\frac{P+1}{P-1}.$$

The calculations were carried out for $n = 50, 100, 150$ and $P = 10^3, 10^7, 10^8$ on the GIER computer and are shown in Tables 1–3.

The tables show the results of calculations obtained by separate methods in which we started from the initial vectors obtained by the Gaussian method, the time required for one iteration and the entire number of iterations, the accuracy of solutions in the various systems. Further, the solutions by means of the Gaussian method are introduced as well as the precise solutions obtained by using double accuracy arithmetics.

Comparing the iterative methods referred to with the Gaussian method it can be said, Tab. 1–3, that in solving systems with ill-conditioned 3-diagonal matrices, the iterative methods require an unproportionally longer time for the calculation of the solution than does the Gaussian method. The accuracy of the results obtained is, however, by no means better. In iterating the result obtained by the Gaussian

Table 1

$n = 50$

P	Method	Number of iterations	Time per iteration	Precision	x_1	x_{25}	x_{50}
	relation (2)	8	7·5 sec	10^{-7}	1·01312346	1·39659246	1·97229842
	relation (3)	164	1·6 sec	10^{-7}	1·01675890	1·39658517	1·97956828
10^3	relation (4)	25	16 min	10^{-3}	1·01312296	1·39658370	1·97229655
	Gaussian				1·01312348	1·39659209	1·97229707
	precise solution				1·01312336	1·39659037	1·97229692
	relation (2)	1	7·5 sec	10^{-2}	1164·02049	18884·2056	1165·02060
	relation (3)	15	1·6 sec	10^{-3}	1164·03244	18884·1996	1165·04304
10^7	Gaussian				1164·02057	18884·2067	1165·01914
	precise solution				1166·45819	18923·7813	1167·45629
	relation (2)	1	7·5 sec	10^{-1}	17346·2574	281625·052	17347·2641
	relation (3)	4	1·6 sec	10^{-2}	17346·3042	281625·068	17347·3546
10^8	Gaussian				17346·2588	281625·084	17347·2636
	precise solution				15630·6425	253767·647	15631·6406

Table 2

$n = 100$

P	Method	Number of iterations	Time per iteration	Precision	x_1	x_{50}	x_{100}
	relation (2)	8	14·5 sec	10^{-7}	0·94725323	0·18347321	1·89323823
	relation (3)	158	3·0 sec	10^{-7}	0·95088976	0·18347209	1·90051077
10^3	Gaussian				0·94725327	0·18347370	1·89323786
	precise solution				0·94725323	0·18347295	1·89323780
	relation (2)	1	14·5 sec	10^{-3}	153·341403	4898·83200	154·341024
	relation (3)	14	3·0 sec	10^{-3}	153·353670	4898·83163	154·365609
10^7	Gaussian				153·341415	4898·83250	154·341092
	precise solution				152·391682	4898·29520	153·391189
	relation (2)	1	14·5 sec	10^{-2}	2131·50824	68498·1060	2132·51108
	relation (3)	21	3·0 sec	10^{-3}	2131·52048	68498·1028	2132·53438
10^8	Gaussian				2131·50839	68498·1132	2132·51045
	precise solution				1954·22749	62798·3878	1955·22701

Table 3

$n = 150$

P	Method	Number of iterations	Time per iteration	Precision	x_1	x_{75}	x_{150}
	relation (2)	8	21·5 sec	10^{-7}	0·94203552	0·03265820	1·88402895
	relation (3)	157	5·5 sec	10^{-7}	0·94567210	0·03265790	1·89130086
10^3	Gaussian				0·94203557	0·03265821	1·88402775
	precise solution				0·94203554	0·03265814	1·88402774
	relation (2)	1	21·5 sec	10^{-3}	45·5054990	2140·14903	46·5053504
	relation (3)	14	5·5 sec	10^{-3}	45·5177706	2140·14882	46·5299261
10^7	Gaussian				45·5054998	2140·14920	46·5053821
	precise solution				45·2028812	2125·60086	46·2026503
	relation (2)	1	21·5 sec	10^{-2}	700·680637	33631·6143	701·682735
	relation (3)	14	5·5 sec	10^{-3}	700·692757	33631·6083	701·706501
10^8	Gaussian				700·680629	33631·6169	701·682165
	precise solution				619·855470	28746·6564	620·855252

method, the required accuracy of the solution must also be taken into consideration. Should this accuracy exceed the capabilities of arithmetics in the simple accuracy of the computer, it is necessary to apply double precision arithmetics which implies a prolongation of time required for the solution.

In addition to refining the results obtained by the Gaussian method, calculations were carried out for the zero initial vectors. We get a solution of the system for the zero initial vector but its quality is not proportionate to the required machine time. In case of condition 10^3 the solution of the system was obtained approximately in 9–20 minutes and $\varepsilon = 10^{-6}$. Under condition 10^7 and 10^8 the solution time increased out of proportion. To get the solution it would be necessary to calculate from tens to hundreds of hours. For example for the condition $P = 10^7, 10^8$ according to relation (3), with 10 000 iterations made, the solution did not equal the accurate one in none of the valid places, although the sequence of iterations was converging.

It ensues from the above considerations that the application of relation (3) for the calculation of the solution of ill-conditioned systems with a 3-diagonal matrix is inadequate both from the viewpoint of precision and from that of the required machine time. The situation is similar when calculating according to (2).

The calculation of the solution according to (4) has a special position because in this case it is not the solution which is directly calculated but only the inverse matrix.

Besides the systems shown in Tables 1–3, relation (4) was also used to solve

systems with matrices of the dimensions 10×10, 25×25 under various conditions. The following was found:

1. Calculation time increases out of proportion along with the increase in size of the matrix and the accuracy of calculation of the inverse matrix.

2. If the inverse matrix is calculated with precision δ ($\|H\| < \delta$), then the obtained solution is calculated with precision $\delta' \geqq 10\delta$.

3. In case of an ill-conditioned matrix ($P = 10^7$, 10^8) no solution was obtained for small values δ, although the conditions of convergence were fulfilled. In this case the norm of matrix H converges to δ up to a certain $\delta_1 > \delta$ and then begins to oscillate in the neighbourhood of the point δ_1 not reaching δ.

4. Since the solution is calculated on the basis of calculating the inverse matrix which is full also in the case that A is 3-diagonal, it is meaningless to make calculations specially for 3-diagonal systems because the saving of storage capacity is $n^2 - 3n + 2$ which is out of proportion with respect to the complexity of the programming.

With reference to the results obtained it can be stated that this method is not suitable for the calculation of systems with ill-conditioned matrices because, compared with Gaussian method, it not only fails to attain the proper accuracy but even lacks the capacity to solve some ill-conditioned systems (see point 3).

Recent experiences gained in calculating various systems show, as can be seen from *Tables* 1–3, *that the Gaussian method appears to be the most appropriate even if systems with ill-conditioned matrices are concerned.*

Note: The adjusted algorithms can be used e.g. in solving partial differential equations, or in solving tasks of construction mechanics where the type of diagonal matrices is most frequently encountered.

References

[1] *J. Gruska, V. Chmurná, M. Kasmanová, P. Poliak, M. Postulková, R. Vyhnanská:* Riešenie systémov lineárnych rovníc s pásovými maticami na samočinných počítačoch, Výskumná zpráva Z 24/65—66, ÚTK SAV, 1966.

[2] *М. П. Симою:* Итеративный метод обращения матриц, Журнал вычислительной математики и физики, Том 5, 1965.

[3] *А. В. Буледза:* О построении многошаговых итеративных процессов решения линейных операторных уравнений, Вестник Киевского Университета, серия математики и механики, 1961.

[4] *А. В. Буледза:* О разностных методах решения краевых задач для дифференциальных уравнений эллиптического типа, Доклады и сообщения Ужгородского Университета, серия физ.-мат. и истор., наук, Том 1, 1965.

[5] *Zurmühl R.:* Matrizen und ihre technischen Anwendungen, Berlin 1961.

Súhrn

RIEŠENIE LINEÁRNACH ALGEBRAICKÝCH ROVNÍC
S 3-DIAGONÁLNOU, ZLE PODMIENENOU MATICOU SYSTÉMU

PAVOL POLIAK, MÁRIA POSTULKOVÁ, ROMANA VYHNANSKÁ

V článku sú porovnané niektoré metódy na riešenie systému lineárnych algebraických rovníc so zle podmienenou maticou systému s Gaussovou metódou. Uvažované metódy boli modifikované pre trojdiagonálne systémy.

Authors' addresses: Pavol Poliak prom. mat., *Mária Postulková* prom. mat., Ústav technickej kybernetiky, SAV, Dúbravská cesta, Bratislava, *Romana Vyhnanská* prom. mat., Výskumné výpočtové stredisko, Program OSN, Bratislava.