Mathematica Balkanica

Mathematical Society of South-Eastern Europe
A quarterly published by
the Bulgarian Academy of Sciences – National Committee for Mathematics

Mathematica
Balkanica

New Series Vol.10, 1996, Fasc.4

A Direct Method for Solving Band Systems of Linear Algebraic Equations

M. Salah El-Sayed

Presented by P. Kenderov

A generalization of a known direct method for solving tri-diagonal systems of linear equations is proposed and studied. The $(2m+1)$-diagonal system of linear equations are considered. Comparison in some kinds of linear systems of equations for our method (P-method) and Sweep method (S-method) are discussed.

1. Introduction

As it is well known [see 1 and 3-5] that the S-method for solving tri-diagonal and $(2m+1)$-diagonal systems is effective (stable) if the matrix of the system is diagonally dominant. The p-method solving band systems is effective method if the matrix of the system is non dominant diagonal. In a sense, the applicability of the P-method and S-method complement each other.

In this paper we construct the algorithm for $(2m+1)-$ diagonal (band) matrix of the systems and prove main theorem for choise the initial values for the algorithm. Section 2 describe the method for doing this. Section 3 contains the comparison in some kinds of linear systems of equations for P-method and S-method.

2. Description of the method (Algorithm)

We are describe here the generalization of the method which is given in [4, p.42].

Let we have

(0.1) $$Ax = f$$

be a nxn linear system of equations with a matrix $A = (a_{ij})$, $a_{ij} = 0$ for $|i-j| > m < n, x = \{x_i\}, f = \{f_i\}$ and detA$\neq 0$. Therefore, (1) is band system with width $(2m+1)$ of the band of the matrix A.

The method consists of the following:
In the first equation of (1) we set $x_s = x_s^k \quad (s = 1, 2, \ldots, m)$ and

$$y^k = (x_1^k, x_2^k, \ldots, x_m^k) \qquad K = 0, 1, \ldots, m. \tag{0.2}$$

After this from the same first equation of (1) we can obtain $x_{m+1} = x_{m+1}^k$,from second equation of (1) $x_{m+2} = x_{m+2}^k$ and etc., we stop with the determining of $x_n = x_n^k$ from $(n-m)^{th}$ equation. Thus we obtained $m+1$ solutions $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$ of the system formed by the first $n-m$ equations of (1). Further, we search the solution of (1) in the form

$$x = x^0 + \alpha_1(x^1 - x^0) + \ldots + \alpha_m(x^m - x^{m-1}), \tag{0.3}$$

where $\alpha_s, S = 1, 2, \ldots, m$ are parameters to be found. For this reason we will call this method as parametric method or p-method. It is easily to seen that the independently of the values of the parameters, x from (3) satisfies the first $n-m$ equations of (1). Therefore, parameters α_s and x from (3) must be satisfies the last m equations too. This leads to the following system for $\alpha = (\alpha_1\alpha_2\ldots\alpha_m)^T$

$$\sum_{s=1}^{m} a_i(x^s - x^{s-1})\alpha_s = f_i - a_i x^0, \quad i = n-m+1, n-m+2, \ldots, n, \tag{0.4}$$

where a_i is i^{th} vector row of the matrix A.

Now, arising the question when the system (4) has an unique solution, i.e. when the matrix of the system (4) is non- singular. The answer is given by the following

Theorem. *The linear system (4) has unique solution if the system from vectors*

$$y^1 - y^0, y^2 - y^1, \ldots, y^m - y^{m-1} \tag{0.5}$$

is linear independent.

Proof. Let (5) linear independent system from vectors. We can write system (4) in the form

$$\sum_{s=1}^{m} \alpha_s(r_i^s - r_i^{s-1}) = -r_i^0 \qquad (i > n-m),$$

where $r_i^s = a_i x^s - f_i$, and we assume that the determinant of matrix of the system is equal zero. In this case there exist a non-zero vector $(t_1, t_2, \ldots, t_m)$, such that

$$\sum_{s=1}^{m} t_s(r_i^s - r_i^{s-1}) = 0 \qquad (i > n - m).$$

But the above equation is true for each $i \leq n - m$ too. Therefore, it will be true and

$$\sum_{s=1}^{m} t_s(r^s - r^{s-1}) = 0, \tag{0.6}$$

where $r^s = (r_1^s, r_2^s, \ldots, r_n^s)^T$. Further, from (6) we find consecutively

$$\sum_{s=1}^{m} A[t_s(x^s - x^{s-1})] = 0$$

$$A\sum_{s=1}^{m} t_s(x^s - x^{s-1}) = 0$$

$$\sum_{s=1}^{m} t_s(x^s - x^{s-1}) = 0$$

$$\sum_{s=1}^{m} t_s(y^s - y^{s-1}) = 0.$$

But the last equation shows the system of vectors (5) is linearly dependent. Then there exist contradiction which proof the theorem. ■

R e m a r k 1. It is easy to see that the system in the above theorem can be replaced by

$$y^1 - y^0, y^2 - y^0, \ldots, y^m - y^0. \tag{0.7}$$

From this, combined with the theorem leads to the conclusion that we can take $y^0 = 0, y^s = e^s (s = 1, 2, \ldots, m)$, where e^s is denoted the s^{th} m-dimensional orthonormal vectors.

R e m a r k 2. In the case $m = 1$ P-method is comparable with the S-method with respect to the total number of the arithmetic operations. For the both methods this number is $O(n)$.

R e m a r k 3. It is evident that for the applicability of the P-method described with $\det A \neq 0$ it is necessary all $a_{i,m+1} \neq 0 \quad (i = 1, 2, \ldots, n - m)$. But, as we will see later, next inequality and moreover the condition

$$|a_{i,m+1}| \geq \sum_{i<m+1} |a_{i,j}| \tag{0.8}$$

is not sufficient for the effectivenesses of the method .

3. Numerical experiments

In each of the following examples we will solve a system of linear equations $Ax = f$, where $A = (a_{ij}), x = (x_i)$ and the vector $f = (f_i)$ is chosen in such a way that the solution of the system to be $x = (1, 1, \ldots, 1)^T$. The error ϵ of the solution computed $\acute{x} = (\acute{x}_1, \acute{x}_2, \ldots, \acute{x}_n)^T$ is measured by the first vector norm

$$\epsilon(\acute{x}) = \|x - \acute{x}\|_1 = \max_i |x_i - \acute{x}_1| \tag{0.9}$$

3.1. $a_{ii} = i \qquad i = 1, 2, \ldots, n$

$a_{i,i+1} = a_{i+1,i} = n \qquad i = 1, 2, \ldots, n - 1$

$a_{ij} = 0 \qquad$ Otherwise.

Experiments with $n = 50$ and $n = 100k$ $(k = 1, 2, \ldots, 10)$ were made. The maximal error ϵ_n for this example
$\max_n \epsilon_n(\acute{x}) = \omega.10^{-5}$
where here ω is a corresponding number in the interval [0.1,1).

3.2. $a_{11} = 1$

$a_{ii} = 2$

$a_{nn} = 1 + \delta$

$a_{i,i+i} = a_{i+1,i} = 1 \qquad i = 1, 2, \ldots, n - 1$

$a_{ij} = 0 \qquad$ Otherwise.

It easy to show that $\det A = \delta$.

In the following table 1 "–" means that the method is ineffective.

Table 1

n	δ	P	S
50	10^{-1}	$\omega.10^{+1}$	$\omega.10^{-5}$
50	10^{-4}	$\omega.10^{+1}$	$\omega.10^{-2}$
50	10^{-7}	$\omega.10^{+1}$	$\omega.10^{+1}$
100	10^{-1}	$\omega.10^{+1}$	$\omega.10^{-5}$
100	10^{-4}	--	$\omega.10^{-3}$
100	10^{-7}	--	$\omega.10^{+1}$
200	10^{-1}	$\omega.10^{+1}$	$\omega.10^{-5}$
200	10^{-4}	$\omega.10^{+1}$	$\omega.10^{-2}$
200	10^{-7}	$\omega.10^{+1}$	$\omega.10^{+1}$
400	10^{-1}	$\omega.10^{+1}$	$\omega.10^{-5}$
400	10^{-5}	$\omega.10^{+1}$	$\omega.10^{-3}$
400	10^{-7}	$\omega.10^{+1}$	$\omega.10^{+1}$

The P and S columns of the Tables 1,2 are the errors of the computed solution arising with P-method and S-method. It is seen that the results by the S-method are much better than P-method, due to the fact that A is a matrix with diagonally dominant which is unfavorable for the P-method.

In the following examples the matrix A of the system is a tri-diagonal Toeplitz matrix, i.e. the matrix of the form

$$a_{i+1,i} = \text{const } = a \quad i = 1,2,\dots,n-1$$

$$a_{ii} = \text{const } = b \quad i = 1,2,\dots,n$$

$$a_{i,i+1} = \text{const } = c \quad i = 1,2,\dots,n$$

$$a_{ij} = 0 \quad \text{Otherwise.}$$

Further, we can use the denotation $A(a,b,c)$ for such a matrix.

In the case for such a matrix, if x satisfies the first $n-1$ equation of the system, then

$$ax_{s-1} + bx_s + cx_{s+1} = f_s \qquad s = 2,3,\dots,n-1 \tag{0.10}$$

and x_s can be obtained using the formula of the general solution of a recurrent equation with constant coefficients (10). As it is well known, this formula has the form

$$x_s = p\lambda_1^s + q\lambda_2^s + \delta_s \qquad s = 1,2,3,\dots \tag{0.11}$$

or

(0.12) $$x_s = p\lambda^s + qs\lambda^s + \delta_s \qquad s = 1, 2, 3, \dots$$

where p and q are constants, and δ_s is particular solution of (10) and formula (11) is valid if the characteristic equation $c\lambda^2 + b\lambda + a = 0$ has two distinct roots $\lambda_1$and λ_2 and formula (12) is valid if $\lambda_1 = \lambda_2 = \lambda$. Now in the same way for P-method, if $x_1 = x_1^0$ is given which satisfy the first equation $bx_1 + cx_2 = f_1$, and (11) or (12) also, then we can obtain the coefficients $p = p_0, q = q_0$ and $x = x^0(x = (x_1, x_2, \dots, x_n)^T)$ are uniquely determined by the first $n-1$ equations of Ax=f. If we obtain a second particular solution $x^1(p_1, q_1)$ of the first $n-1$ equations in the similar way, then the unique solution of the system is sought of the form

(0.13) $$x = \alpha x^0 + (1 - \alpha)x^1.$$

where α is a numerical parameter. Since x^0 and x^1 satisfy the first $n-1$ equations of the system, then whatever α and x of (13) to be, it will satisfies the same equations. Hence, x from (13) will be a solution of the whole system if it satisfy last n^{th} equation of the system. Thus we can obtain the value

(0.14) $$\alpha = \frac{r_n^0}{r_n^0 - r_n^1}$$

with $\quad r_n^s = ax_{n-1}^s + bx_n^s + f_n, \qquad s = 0, 1.$

These considerations can help to interpret some of the results in the following examples and to characterize the domain of applicability of the P-method for solution of Toeplitz tri-diagonal systems.

3.3. $a_{ii} = \delta \qquad i = 1, 2, \dots, n$

$a_{i,i+1} = a_{i+1,i} + 2 = 10 \qquad i = 1, 2, \dots, n-1$

$a_{ij} = 0 \qquad$ Otherwise.

Table 2

n	δ	P	S
50	1	0	$\omega.10^{-4}$
50	4	0	$\omega.10^{-3}$
50	7	$\omega.10^{-6}$	$\omega.10^{-4}$
50	10	$\omega.10^{-6}$	$\omega.10^{-5}$
100	1	0	$\omega.10^{-2}$
100	4	0	$\omega.10^{-1}$
100	7	$\omega.10^{-6}$	$\omega.10^{-2}$
100	10	0	$\omega.10^{-1}$
200	1	0	$\omega.10^{+4}$
200	4	$\omega.10^{-6}$	$\omega.10^{+3}$
200	7	$\omega.10^{-7}$	$\omega.10^{+4}$
200	10	$\omega.10^{-7}$	$\omega.10^{+3}$
400	1	0	$\omega.10^{+5}$
400	4	0	$\omega.10^{+4}$
400	7	$\omega.10^{-6}$	$\omega.10^{+3}$
400	10	$\omega.10^{-6}$	$\omega.10^{+3}$

It is seen that Table 2 that P-method gives results with much greater accuracy than S-method. Such is the situation for greater n too.

3.4. $A = A(1/8, 1, 4)$

For the above matrix A we can show that

$$\det A = 2^{\frac{1-n}{2}} \cos \frac{(n-1)\pi}{4} \tag{0.15}$$

where n is the order of the matrix. From (15) we obtain that $detA = 0$ if and only if $n = 4k + 3$; $k = 0, 1, 2, \dots$. It is clear that $detA \to 0$ for $n \to> \infty$. The solutions of the corresponding characteristic equation are

$\lambda_{1,2} = \frac{1 \pm i}{8}$

i.e. $|\lambda_1| = |\lambda_2| = \sqrt{2}/8 < 1$. In this case, according to the considerations made above, if x is the solution of the first $n-1$ equations of $Ax = f$, then we will have

$$x_k = p\lambda_1^k + q\lambda_2^k + 1 \qquad k = 1, 2, \dots \tag{0.16}$$

From the above equation (16) it is clear that $x_k \to 1$ For $k \to \infty$. This implies that for arbitrary chosen $\epsilon > 0$ and x_1, there is a great enough n such that x can be taken as an approximate solution not only of the first $n-1$ equations of the system, but also for the whole system $Ax = f$ too. In this case of n the P-method is unapplicable. The computational practice confirm this.

3.5. Other experiments of tri-diagonal Toeplitz system with the matrix of the form $A(a, b, c; \lambda_1, \lambda_2)$ were made. They concern the cases:

3.5.1. $A = A(1, 1, -2; -1, 2)$

3.5.2. $A = A(4, 1, -4; -0.88, 1.13)$

3.5.3. $A = A(1, 4, 4; -0.5, -0.5)$

3.5.4 $A = A(4, -4, 1; 2, 2)$

3.6. $a_{ii} = 10^{-k}$ $\quad i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots 5$

$a_{i,i+1} = 9.899 \quad i = 1, 2, \ldots, n-1$

$a_{i+1,i} = 9.10^{-3} \quad i = 1, 2, \ldots, n-1$

$a_{i,i+2} = 10 \quad i = 1, 2, \ldots, n-2$

$a_{i+2,i} = 10^{-3} \quad i = 1, 2, \ldots, n-2$

$a_{i,j} = 0$ Otherwise.

In this example of five-diagonal systems the S-method happens to be ineffective. The solution by the P-method of the above examples were obtained by the maximal error of the form
$\max_{\delta,n} \epsilon(\acute{x}) = \omega.10^{-2}$.

Along the fact that P-method solves linear systems for which the S-method is unapplicable, it has following advantages more:

P-method give a good and natural options for parallel treating and for computing of separate components of the solution of the Toeplitz systems only without looking for the whole solution.

References

1. V. P. I l' i n, Yu. I. K u z n e t s o v. Tri-diagonal matrices and their applications. *Naouka*, Moscow, 1985.

2. D. K. F a d d e e v, V. N. F a d d e e v a. Computational methods of linear algebra . Moscow, Fizmatgiz, 1963.
3. V. P. M a i e r. Double Sweep algorithm for $a(2m+1)$ diagonal matrix. *Zh. Vychisl. Math. i Math. Fiz.*, **24**, 1984, 627-632.
4. A. A. S a m a r s k i i. Theory of the difference schemes. *Naouka*, Moscow, 1977.
5. A. A. S a m a r s k i i, E. S. N o k o l a e v. Methods of solution of net's equations. *Naouka*, Moscow, 1978.

Institute of Mathematics,
Acad. G. Bonchev Str., Bl. 8,
1113 Sofia,
BULGARIA

Received 30.12.1994