Annales Univ. Sci. Budapest., Sect. Comp. 50 (2020) 151–156

A NOTE ON THE SOLUTION OF NARROW-BANDED TOEPLITZ SYSTEMS

Csaba J. Hegedüs (Budapest, Hungary)

Dedicated to the memory of Professor Gisbert Stoyan

Communicated by Ferenc Schipp

(Received January 5, 2020; accepted February 25, 2020)

Abstract. This note observes that the LU-decomposition of narrow-banded Toeplitz matrices can be modified so that L and U are lower and upper band Toeplitz matrices plus a low rank correction matrix is added. The resulting method cannot be considered generally applicable however, it may be useful in special cases. For solving a system, the operation count is about half of Dickinson's algorithm, but the work of data preparation needs less operations only if the half bandwidth is below 6. For the special matrix tridiag $(-1, 2, -1)$ the suggested method of solution needs $4n$ additions and one division.

1. Introduction

Matrices whose entries are constant along each diagonal arise in many applications and are called Toeplitz matrices. A general element of such a matrix can be given by the relation $a_{ij} = c_{j-i}$, where the $2n-1$ scalars $c_{-n+1}, \dots c_0, \dots c_{n-1}$ determine the matrix of order n. Toeplitz matrices attracted a wide interest of research in the past years. There are methods available for general Toeplitz matrices and also, there are efficient procedures for the symmetric positive definite or band cases. The interested reader may consult the book of Heinig and Rost [4] for an overview of the theory. Practical

Key words and phrases: Band Toeplitz matrices, matrix inversion, linear equations.
2010 Mathematics Subject Classification: 15-04, 15B05, 15A23.

numerical algorithms are given in the textbook of Golub and Van Loan [3], or in Russian, a good source is the book of Voevodin and Tyrtyshnikov [7].

Here a special class, the band Toeplitz matrices will be considered. One can find a stability analysis of three algorithms for solving band Toeplitz systems in [5] and further references therein. However, our approach here for the inversion of band systems is more resembling that of Trench [6]. There it is observed that a band matrix is close to a lower (or upper) triangular matrix. Thus the inverse is approximated by the inverse of the lower (or upper) part of the matrix and formulae are given for the corrections in terms of the roots of the polynomial associated with the Toeplitz matrix.

This note will make use of the fact that the LU-decomposition of narrow-banded Toeplitz matrices can be changed to the sum of two LU products of lower and upper band Toeplitz matrices, where one of the products may be considered as a correction of low rank.

With this arrangement there is a similarity to the *Gohberg–Semencul* formula for the inverse of a Toeplitz matrix. It expresses the inverse by the difference of two products, where lower and upper triangular Toeplitz matrices are multiplied.

An n-by-n Toeplitz matrix of total bandwidth $2k+1$, $k < n$ can be given by

$$C = t_n(c_{-k}, c_{-k+1}, \ldots, c_0, \ldots, c_{k-1}, c_k), \tag{1.1}$$

where element c_0 belongs to the main diagonal. If the lower or upper bandwidth is smaller than k, then it is indicated by zeros. When specifying a band Toeplitz matrix with (1.1), always odd number of the c_i-s will be given, where the middle element refers to c_0. With this notation we have tridiag $(-1, \delta, -1) = t_n(-1, \delta, -1)$ and the elementary nilpotent matrix having 1's in the subdiagonal, 0 otherwise is given by $N = t_n(1, 0, 0)$. The polynomial of order $2k$ associated with the Toeplitz matrix (1.1) is

$$c(t) = t^k \sum_{i=-k}^{k} c_i t^i. \tag{1.2}$$

2. The method of inversion

Assume that the lower bandwidth is l and the upper bandwith is r, such that $l, r \leq k$. The polynomial $c(t)$ in (1.2) can then be factored into two polynomials of order l and r. Such polynomials are easily found if the roots of $c(t)$ are known:

$$c(t) = a(t)b(t). \tag{2.1}$$

Now associate with polynomial $a(t)$ the lower triangular matrix

$$A = \begin{bmatrix} a_l & & & & & \\ a_{l-1} & a_l & & & & \\ \vdots & \ddots & \ddots & & & \\ a_0 & \ddots & \ddots & a_l & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & a_0 & \dots & \dots & a_l \end{bmatrix} \tag{2.2}$$

and with polynomial $b(t)$ the upper triangular matrix:

$$B = \begin{bmatrix} b_0 & b_1 & \dots & b_r & & & \\ & b_0 & \dots & b_{r-1} & b_r & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & b_0 & \dots & \dots & b_r \\ & & & & b_0 & \dots & b_{r-1} \\ & & & & & \ddots & \vdots \\ & & & & & & b_0 \end{bmatrix} \tag{2.3}$$

It can be checked directly that the matrix product AB will be almost equal to C if the bandwidths are small. We find missing terms only in the left upper $l \times r$ corner of C. They can be identified by complementing the matrices A and B. That is, we have to add some columns in front of A and add the same number of rows on top of B. The necessary number of columns and rows is $\min(l, r)$ and what we have to do is to continue the bands. Assume e.g. that l is the smaller, than we have to prepare the $n \times l$ matrix

$$A_0 = \begin{bmatrix} a_0 & \dots & \dots & a_{l-1} \\ 0 & a_0 & \dots & a_{l-2} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & a_0 \\ \vdots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 \end{bmatrix} \tag{2.4}$$

and the $l \times n$ matrix

$$B_0 = \begin{bmatrix} b_l & \dots & b_r & 0 & \dots & & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & & & \\ b_1 & b_2 & \dots & \dots & b_r & 0 & \dots & 0 \end{bmatrix} \tag{2.5}$$

giving

$$C = \begin{bmatrix} A_0 & A \end{bmatrix} \begin{bmatrix} B_0 \\ B \end{bmatrix} = A_0 B_0 + AB. \tag{2.6}$$

This formula can be identified as the matrix representation of multiplying two polynomials, cf. Sect. 0.3 of [4]. The product $A_0 B_0$ may be thought to be a low rank modification to AB thus the inverse of C can be computed by the Sherman–Morrison–Woodbury formula, which is given for matrices A, U and V by (2.1.4) of [3]:

$$\left(A + UV^T\right)^{-1} = A^{-1} - A^{-1}U\left(I + V^T A^{-1} U\right)^{-1} V^T A^{-1}. \tag{2.7}$$

With a substitution into (2.7), one has for the inverse of (2.6)

$$C^{-1} = B^{-1}A^{-1}\left[I_n - A_0 F^{-1} B_0 B^{-1} A^{-1}\right], \tag{2.8}$$

where

$$F = I_l + B_0 B^{-1} A^{-1} A_0. \tag{2.9}$$

If matrix F is not invertible then it indicates the singularity of C.

When applying (2.8), one has to prepare data at first. That means to calculate the roots of the associated polynomial and the LU-decomposition of F.

The two polynomials $a(t)$ and $b(t)$ have altogether $l + r + 2$ coefficients. As there are $l + r + 1$ nonzero bands, we may choose $a_l = 1$ of $a(t)$. We also have freedom in grouping the roots among the polynomials. The coefficient b_0 is in the diagonal and plays the role of a pivot. It is equal to the product of the roots, thus the largest roots may be chosen into $b(t)$.

The other task at the beginning is to compute the entries of F and perform an LU-decomposition. When computing the necessary number of operations, observe that multiplying with A or A^{-1} to a vector needs ln multiplications and additions and we have the number rn for matrix B^{-1}, $rl - l^2/2$ for B_0 and $l^2/2$ for A_0. Thus the final operation count for F is $nl(l + r) + rl^2 + l^3/3$.

Having these data at hand, the computation of $y = C^{-1}x$ may be done in three steps:

1) $y_1 = B^{-1}A^{-1}x$,
2) $y_2 = x - A_0 F^{-1} B_0 y_1$
3) $y = B^{-1}A^{-1}y_2$.

These operations need altogether $(r+l)(2n+l)$ multiplications and additions assuming that always matrix-by-vector type operations are done and the LU-decomposed form of F is available. Compare this number e. g. to that of

Dickinson [2] for Toeplitz band matrices: $(6(l+1)+4r)n$. To prepare data in Dickinson's version needs $(5(l+r)+6)n+\mathcal{O}((l+r)^2)$ operations. It is seen that preparation of data may need less operations at Dickinson, and computing the solution needs more. Hence, the method suggested here is advantageous if we have to solve the same Toeplitz band system for many right hand sides or if preparation work is simple or it can be done by a paper-and-pencil work.

For the other case of $l > r$, one can proceed similarly, so that matrix B_0 will have a full lower triangular block and A_0 will have a truncated upper triangular pattern. It is seen that the suggested method can be used theoretically for any Toeplitz matrices. But it is of practical value only if the matrix is banded with sufficiently narrow bands.

It is still worth mentioning that C may differ from a Toeplitz matrix in the region, where the elements of A_0B_0 are located, for example, because of boundary conditions. Those differences may be taken into account by recalculating A_0B_0 and the method of solution is otherwise the same. Other minor differences can be incorporated similarly.

3. Example

Consider the n-by-n tridiagonal matrix $T(\delta) = t_n(-1, \delta, -1)$. Its associated polynomial is quadratic having discriminant $\sqrt{\delta^2-4}$. The roots are real if $\delta \geq 2$ and their product is equal to -1. Denote by α one of the roots then the polynomial takes the form of $(t-\alpha^{-1})(\alpha - t)$ and with the notation $N = t_n(1,0,0)$, $T(\delta)$ can be written in the form:

$$T(\delta) = \left(I - \frac{1}{\alpha}N\right)\left(\alpha I - N^T\right) + \frac{1}{\alpha}e_1e_1^T, \tag{3.1}$$

where e_1 is the first Cartesian unit vector. Introduce vector e by

$$e = \left(I - \frac{1}{\alpha}N\right)^{-1}e_1 \tag{3.2}$$

then the inverse of $T(\delta)$ is expressible as

$$T^{-1}(\delta) = \left(\alpha I - N^T\right)^{-1}\left(I - \frac{ee^T}{\alpha^2 + e^Te}\right)\left(I - \frac{1}{\alpha}N\right)^{-1}. \tag{3.3}$$

Although the entries of $T^{-1}(\delta)$ are available explicitly – e.g. in [8], this formula is more adequate for computational purposes. If $\delta = 2$ then $\alpha = 1$ and the solution of $T(2)x = b$ needs only $4n$ additions and one division. It is known, cyclic reduction (see Sect 4.5.4 of [3] or [1]) or block cyclic reduction with block size q has a better complexity $q^3 \log n$. We have done running time

comparison for $T(2)$ between the method given here and that of [1]. To our surprise, the running time was the same. An explanation may be that cyclic reduction needs more organizational work that is not taken into account when giving the complexity number.

$T(2)$ comes from a difference scheme for the second derivative. Observe that the coefficients in an mth order difference scheme are proportional to those of the polynomial $(t-1)^m$ if equal spacing is used. If we have a difference scheme that can be associated with a known polynomial then the resulting Toeplitz matrix is easily invertible with the above suggested method. It is also possible to use such easily invertible Toeplitz matrices for the preconditioning of more complicated schemes.

References

[1] **Bini, D. and B. Meini,** Effective methods for solving banded Toeplitz systems, *SIAM J. Matrix Anal. Appl.*, **20(3)** (1999), 700–719.

[2] **Dickinson, B.W.,** Efficient solution of linear equations with banded Toeplitz matrices, *IEEE Trans. Acoust. Speech Sig. Proc.*, **27** (1979), 421–423.

[3] **Golub, G.H. and C.E. Van Loan,** *Matrix Computations*, 2nd Ed., Sect. 4.7., The Johns Hopkins Univ. P., Baltimore, 1989.

[4] **Heinig, G. and K. Rost,** *Algebraic Methods for Toeplitz-like Matrices and Operators*, Akademie Verlag, Berlin, 1984.

[5] **Linzer, E.,** On the stability of solution methods for band Toeplitz systems, *Linear Algebra Appl.*, **170** (1992), 1–32.

[6] **Trench, N.F.,** Explicit inversion formulas for Toeplitz band matrices, *SIAM J. Algebraic Discrete Methods*, **6** (1985), 546–554.

[7] **Voevodin, V.V. and E.E. Tyrtyshnikov,** *Computational Algorithms for Toeplitz matrices*, (in Russian), Nauka, Moscow, 1987.

[8] **Windisch, G.,** Die Inverse der Matrix $A = \text{tridiag}(-1, c, -1), c \geq 2$, (in German), *Wiss. Z. Techn. Univ. Chemnitz*, **34** (1992), 123–129.

C. J. Hegedüs
Department of Numerical Analysis
Faculty of Informatics
Eötvös Loránd University
H-1117 Budapest
Pázmány Péter sétány 1/C
Hungary
`hegedus@numanal.inf.elte.hu`