An algorithm based on orthogonal polynomial vectors for Toeplitz least squares problems

Marc Van Barel[1]*, Georg Heinig[2], and Peter Kravanja[1]

[1] Department of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001 Heverlee (Belgium)
Marc.VanBarel@cs.kuleuven.ac.be, Peter.Kravanja@na-net.ornl.gov
[2] Department of Mathematics, Kuwait University,
POB 5969, Safat 13060 (Kuwait)
georg@mcs.sci.kuniv.edu.kw

Abstract. We develop a new algorithm for solving Toeplitz linear least squares problems. The Toeplitz matrix is first embedded into a circulant matrix. The linear least squares problem is then transformed into a discrete least squares approximation problem for polynomial vectors. Our implementation shows that the normwise backward stability is independent of the condition number of the Toeplitz matrix.

1 Toeplitz linear least squares problems

Let $m \geq n \geq 1$, $t_{-n+1}, \ldots, t_{m-1} \in \mathbb{C}$ and

$$T := [\, t_{j-k}]_{j=0,\ldots,m-1}^{k=0,\ldots,n-1}$$

a $m \times n$ Toeplitz matrix that has full column-rank. Let $b \in \mathbb{C}^m$. We want to solve the corresponding Toeplitz linear least squares problem (LS-problem), i.e., we want to determine the (unique) vector $x \in \mathbb{C}^n$ such that

$$\|Tx - b\| \text{ is minimal} \tag{1}$$

where $\| \cdot \|$ denotes the Euclidean norm.

Standard algorithms for least squares problems require $\mathcal{O}(mn^2)$ floating point operations (flops) for solving (1). The arithmetic complexity can be reduced by taking into account the Toeplitz structure of T. Several algorithms that require only $\mathcal{O}(mn)$ flops have been developed. Such algorithms are called *fast*. One of the first fast algorithms was introduced by Sweet in his PhD thesis [10]. This method is not numerically stable, though. Other approaches include those by Bojanczyk, Brent and de Hoog [1], Chun, Kailath and Lev-Ari [3], Qiao [9], Cybenko [4, 5], Sweet [11] and many more. None of these algorithms has yet been

* The work of the first and the third author is supported by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with the authors.

shown to be numerically stable and for several approaches there exist examples indicating that the method is actually unstable.

Recently, Ming Gu [7] has developed fast algorithms for solving Toeplitz and Toeplitz-plus-Hankel linear least squares problems. In his approach, the matrix is first transformed into a Cauchy-like matrix by using the Fast Fourier Transform or trigonometric transformations. Then the corresponding Cauchy-like linear least squares problem is solved. Numerical experiments show that this approach is not only efficient but also numerically stable, even if the coefficient matrix is very ill-conditioned.

In this paper we will also develop a numerically stable method that works for ill-conditioned problems—in other words, for problems that cannot be solved via the normal equations approach. We proceed as follows. The original LS-problem is first embedded into a larger LS-problem. The coefficient matrix of the latter problem has additional structure: it is a circulant block matrix. This LS-problem is then (unitarily) transformed into a LS-problem whose coefficient matrix is a coupled Vandermonde matrix. The latter LS-problem is then solved by using the framework of orthogonal polynomial vectors.

2 Embedding of the original LS-problem

We embed the original LS-problem (1) in the following way. Let A and B be matrices and let a and y be vectors. The extended LS-problem is formulated as follows: determine the vectors x and y such that the norm of the vector

$$r := \begin{bmatrix} A & B \\ T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} a \\ b \end{bmatrix}$$

is minimal. (We assume, of course, that A, B, a and y have appropriate sizes.) If the matrix B is nonsingular, then the first ‘component’ x of the solution $\begin{bmatrix} x \\ y \end{bmatrix}$ of the extended LS-problem coincides with the solution x of the original LS-problem for any choice of A, B and a. We can always choose A and B such that the two block columns

$$C_1 := \begin{bmatrix} A \\ T \end{bmatrix} \qquad \text{and} \qquad C_2 := \begin{bmatrix} B \\ 0 \end{bmatrix}$$

are circulant matrices. For example, we can choose B equal to the identity matrix of order $n-1$ and we can choose A as the $(n-1) \times n$ Toeplitz matrix

$$A := [\, t_{\,-n+1+j-k}]_{j=0,\dots,n-2}^{k=0,\dots,n-1}$$

with $t_{-n-k} = t_{m-k-1}$ for $k = 0, 1, \dots, n-1$. We take a to be the zero vector. However, we can also choose the size of B larger to obtain a number of rows M for the two circulant matrices C_1 and C_2 such that the discrete Fourier transform of size M can be computed efficiently. For example, we could choose M as the smallest power of two larger than or equal to $m + n - 1$. The matrices A and B are now chosen to have sizes $(M - m) \times n$ and $(M - m) \times (M - m)$, respectively. Note that B is square and assumed to be nonsingular.

3 Transformation of the extended LS-problem

Define C_3 as the vector

$$C_3 := -\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{C}^M .$$

The vector C_3 can be interpreted as the first column of a circulant matrix. The extended LS-problem can therefore be formulated as follows: determine the vectors x and y such that the norm of the vector

$$r = \begin{bmatrix} C_1 \; C_2 \; C_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \in \mathbb{C}^M$$

is minimal. Note that the matrix $\begin{bmatrix} C_1 \; C_2 \; C_3 \end{bmatrix}$ is of size $M \times (n + M - m + 1)$. It is well-known that a $p \times p$ circulant matrix C can be factorized as

$$C = \mathcal{F}_p^H \Lambda \mathcal{F}_p$$

where Λ is a $p \times p$ diagonal matrix containing the eigenvalues of C and $\mathcal{F}_p$ denotes the $p \times p$ Discrete Fourier Transform matrix (DFT-matrix)

$$\mathcal{F}_p := \left[\omega_p^{jk} \right]_{j,k=0,\ldots,p-1}$$

where $\omega_p := e^{-2\pi i/p}$ and $i = \sqrt{-1}$. Similarly, if C is of size $p \times q$, where $p \geq q$, then C can be factorized as

$$C = \mathcal{F}_p^H \Lambda \mathcal{F}_{p,q}$$

where Λ is again a $p \times p$ diagonal matrix and where $\mathcal{F}_{p,q}$ denotes the $p \times q$ submatrix of $\mathcal{F}_p$ that contains the first q columns of $\mathcal{F}_p$.

By applying the Discrete Fourier Transform to r, the norm of r remains unchanged: $\|r\| = \|\mathcal{F}_M r\|$. The following holds:

$$\mathcal{F}_M r = \mathcal{F}_M \begin{bmatrix} C_1 \; C_2 \; C_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

$$= \begin{bmatrix} \Lambda_1 \mathcal{F}_{M,n} \; \Lambda_2 \mathcal{F}_{M,s} \; \Lambda_3 \mathcal{F}_{M,1} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3}$$

where $s := M - m$ and where $\Lambda_j =: \operatorname{diag}(\lambda_{j,k})_{k=1}^M$ is a $M \times M$ diagonal matrix for $j = 1, 2, 3$.

We will now translate the extended LS-problem into polynomial language. Define $x(z)$ and $y(z)$ as

$$x(z) := \sum_{k=0}^{n-1} x_k z^k \qquad \text{and} \qquad y(z) := \sum_{k=0}^{s-1} y_k z^k .$$

Here x_k and y_k denote the components of the vectors x and y. The DFT-matrix $\mathcal{F}_M$ can be interpreted as a Vandermonde matrix based on the nodes $z_k = \omega_M^k$, $k = 0, 1, \ldots, M-1$. Equation (3) now implies that the extended LS-problem can be formulated in the following way: determine the polynomials $x(z)$ and $y(z)$, where $\deg x(z) \leq n-1$ and $\deg y(z) \leq s-1$, such that

$$\sum_{k=0}^{M-1} |\lambda_{1,k} x(z_k) + \lambda_{2,k} y(z_k) + \lambda_{3,k} 1|^2 \tag{4}$$

is minimal.

4 Orthogonal polynomial vectors

The minimisation problem (4) can be solved within the framework of orthogonal polynomial vectors developed by Van Barel and Bultheel [2, 12–14]. The following notation will be used: to indicate that the degree of the first component of a polynomial vector $P \in \mathbb{C}[z]^{3\times 1}$ is less than or equal to α, that the degree of the second component of P is less than 0 (hence, this second component is equal to the zero polynomial), and that the degree of the third component is equal to β, we write

$$\deg P = \begin{bmatrix} \alpha \\ -1 \\ \underline{\beta} \end{bmatrix}.$$

We consider the following inner product and norm.

Definition 1 (inner product, norm). *Consider the subspace $\mathcal{P} \subset \mathbb{C}[z]^{3\times 1}$ of polynomial vectors P of degree*

$$\deg P = \begin{bmatrix} n \\ s \\ 0 \end{bmatrix}.$$

Given the points $z_k \in \mathbb{C}$ and the weight vectors

$$F_k = \begin{bmatrix} \lambda_{1,k} & \lambda_{2,k} & \lambda_{3,k} \end{bmatrix} \in \mathbb{C}^{1\times 3}, \qquad k = 1, 2, \ldots, M,$$

we define the discrete inner product $\langle P, Q \rangle$ for two polynomial vectors $P, Q \in \mathcal{P}$ as follows:

$$\langle P, Q \rangle := \sum_{k=1}^{M} P^H(z_k) F_k^H F_k Q(z_k). \tag{5}$$

The norm $\|P\|$ of a polynomial vector $P \in \mathcal{P}$ is defined as:

$$\|P\| := \sqrt{\langle P, P \rangle}.$$

A necessary and sufficient condition for (5) to be an inner product in $\mathcal{P}$, is that $\mathcal{P}$ is a subspace of polynomial vectors such that a nonzero polynomial vector $P \in \mathcal{P}$ for which $\langle P, P\rangle = 0$ (or equivalently: $F_k P(z_k) = 0$, $k = 1, 2, \ldots, M$) does not exist. Our original LS-problem can be now stated as the following discrete least squares approximation problem: determine the polynomial vector $P^\star \in \mathcal{P}'$ such that $\|P^\star\| = \min_{P \in \mathcal{P}'} \|P\|$ where $\mathcal{P}'$ denotes all vectors belonging to $\mathcal{P}$ and having their third component equal to the constant polynomial 1.

In [14], Van Barel and Bultheel formulated a fast algorithm for computing an orthonormal basis for $\mathcal{P}$. The degree sequence of the basis vectors B_j, $j = 1, 2, \ldots, \delta$, is as follows:

$$\begin{bmatrix} \underline{0} & \underline{1} & \cdots & \underline{n-s} & n-s & \underline{n-s+1} & n-s+1 & \cdots & \underline{n} & n & n \\ -1 & -1 & \cdots & -1 & \underline{0} & 0 & \underline{1} & \cdots & s-1 & \underline{s} & s \\ -1 & -1 & \cdots & -1 & -1 & -1 & -1 & \cdots & -1 & -1 & \underline{0} \end{bmatrix}.$$

Every polynomial vector $P \in \mathcal{P}'$ can be written (in a unique way) as:

$$P = \sum_{j=1}^{\delta} a_j B_j$$

where $a_1, \ldots, a_\delta \in \mathbb{C}$. The coordinate a_δ is determined by the fact that the third component polynomial of P has to be monic and of degree 0. The following holds:

$$\begin{aligned} \|P\|^2 &= \langle P, P\rangle \\ &= \left\langle \sum_{j=1}^{\delta} a_j B_j, \sum_{j=1}^{\delta} a_j B_j \right\rangle \\ &= \sum_{j=1}^{\delta} |a_j|^2 \quad (\text{since } \langle B_i, B_j\rangle = \delta_{ij}). \end{aligned}$$

It follows that $\|P\|$ is minimized by setting $a_1, \ldots, a_{\delta-1}$ equal to zero. In other words,

$$P^\star = a_\delta B_\delta \quad \text{and} \quad \|P^\star\| = |a_\delta|.$$

The discrete least squares approximation problem can therefore be solved by computing the orthonormal polynomial vector B_δ. We obtain $P^\star$ by scaling B_δ to make its third component monic.

5 Numerical experiments

We have implemented our approach in Matlab (MATLAB Version 5.3.0.10183 (R11) on LNX86). The numerical experiments that we will present in this section are similar to those done by Ming Gu in [7]. The computations have been done in double precision arithmetic with unit roundoff $u \approx 1.11 \times 10^{-16}$. We have considered two approaches:

- QR: the QR method as implemented in Matlab. This is a classical approach for solving general dense linear least squares problems;
- NEW: the approach that we have described in the previous sections.

We have compared the two approaches QR and NEW for two types of Toeplitz matrices:

- Type 1: the entries t_k are taken uniformly random in the interval $(0, 1)$;
- Type 2: $t_0 := 2\omega$ and $t_k := \frac{sin(2\pi\omega k)}{\pi k}$ for $k \neq 0$ where $\omega := 0.25$. This matrix is called the Prolate matrix and is very ill-conditioned [6, 15].

The right-hand side vector b has been chosen in two ways:

- Its entries are generated uniformly random in $(0, 1)$. This generally leads to large residuals.
- The entries of b are computed such that $b = Tx$ where the entries of x are taken uniformly random in $(0, 1)$. In this case, we obtain small residuals.

To measure the normwise backward error, we have used the following result of Waldén, Karlson and Sun [16]. See also [8, section 19.7].

Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $0 \neq x \in \mathbb{R}^n$, and $r := b - Ax$. Let $\theta \in \mathbb{R}$. The normwise backward error*

$$\eta_F(x) = \min\{ \, \| \, [\Delta A, \theta \Delta b] \, \|_F \; : \; \|(A + \Delta A)x - (b + \Delta b)\|_2 = \min \, \}$$

is given by

$$\eta_F(x) = \min\{ \, \eta_1, \sigma_{\min}\left([A \quad \eta_1 C]\right) \, \}$$

where

$$\eta_1 := \frac{\|r\|_2}{\|x\|_2}\sqrt{\mu}, \quad C := I - \frac{rr^T}{r^T r} \quad \textit{and} \quad \mu = \frac{\theta^2\|x\|_2^2}{1 + \theta^2\|x\|_2^2}.$$

We have computed $\eta_F(x)$ with $\theta := 1$.

The numerical results are shown in Tables 1 and 2 for the two possible choices of the right-hand side vector b.

Table 1. Normwise backward error (small residuals)

Matrix	Order		$\kappa(T)$	$\eta_F(x)/u$	
type	m	n		QR	NEW
1	160	150	5.4×10^2	1.9×10^2	1.7×10^4
	320	300	3.4×10^2	7.5×10^2	9.1×10^4
	640	600	7.7×10^2	5.9×10^2	3.3×10^5
2	160	150	2.1×10^{16}	3.9×10^1	2.7×10^2
	320	300	1.5×10^{16}	2.5×10^0	5.5×10^2
	640	600	1.3×10^{16}	2.8×10^0	1.5×10^3

Table 2. Normwise backward error (large residuals)

Matrix	Order		$\kappa(T)$	$\eta_F(x)/u$	
type	m	n		QR	NEW
1	160	150	5.4×10^2	4.1×10^1	3.0×10^3
	320	300	3.4×10^2	1.3×10^2	2.5×10^4
	640	600	7.7×10^2	1.1×10^2	1.4×10^5
2	160	150	2.1×10^{16}	1.3×10^2	3.9×10^0
	320	300	1.5×10^{16}	1.5×10^0	8.2×10^0
	640	600	1.3×10^{16}	2.7×10^0	2.3×10^1

6 Conclusions

The numerical experiments show that the current implementation is still not accurate enough to be comparable with QR or with the algorithms developed by Ming Gu. However, the results show that the normwise backward error does not depend on the condition number of the Toeplitz matrix. We are currently working on improving the accuracy as well as the speed of the implementation to obtain a viable alternative for the algorithms of Ming Gu where the Toeplitz matrix can range from well-conditioned to very ill-conditioned.

References

1. A. Bojanczyk, R. Brent, and F. de Hoog, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.
2. A. Bultheel and M. Van Barel, *Vector orthogonal polynomials and least squares approximation*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 863–885.
3. J. Chun, T. Kailath, and H. Lev-Ari, *Fast parallel algorithms for QR and triangular factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 899–913.
4. G. Cybenko, *A general orthogonalization technique with applications to time series analysis and signal processing*, Math. Comp., 40 (1983), pp. 323–336.
5. ———, *Fast Toeplitz orthogonalization using inner products*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 734–740.
6. I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.
7. M. Gu, *Stable and efficient algorithms for structured systems of linear equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 279–306.
8. N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, 1996.
9. S. Qiao, *Hybrid algorithm for fast Toeplitz orthogonalization*, Numer. Math., 53 (1988), pp. 351–366.
10. D. Sweet, *Numerical Methods for Toeplitz matrices*, PhD thesis, University of Adelaide, Adelaide, Australia, 1982.
11. ———, *Fast Toeplitz orthogonalization*, Numer. Math., 43 (1984), pp. 1–21.
12. M. Van Barel and A. Bultheel, *A parallel algorithm for discrete least squares rational approximation*, Numer. Math., 63 (1992), pp. 99–121.

13. ———, *Discrete linearized least squares approximation on the unit circle*, J. Comput. Appl. Math., 50 (1994), pp. 545–563.
14. ———, *Orthonormal polynomial vectors and least squares approximation for a discrete inner product*, Electron. Trans. Numer. Anal., 3 (1995), pp. 1–23.
15. J. Varah, *The Prolate matrix*, Linear Algebra Appl., 187 (1993), pp. 269–278.
16. B. Waldén, R. Karlson, and J.-g. Sun, *Optimal backward perturbation bounds for the linear least squares problem*, Numerical Linear Algebra with Applications, 2 (1995), pp. 271–286.