Solving Toeplitz least squares problems via discrete polynomial least squares approximation at roots of unity

Marc Van Barel[a], Georg Heinig[b] and Peter Kravanja[a]

[a]Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee (Belgium)

[b]Department of Mathematics, Kuwait University
POB 5969, Safat 13060 (Kuwait)

ABSTRACT

We present an algorithm for solving Toeplitz least squares problems. By embedding the Toeplitz matrix into a circulant block matrix and by applying the Discrete Fourier Transform, we are able to transform the linear least squares problem into a discrete least squares approximation problem for polynomial vectors. We have implemented our algorithm in Matlab. Numerical experiments indicate that our approach is numerically stable even for ill-conditioned problems.

Keywords: Toeplitz linear least squares problems, orthogonal polynomial vectors

1. TOEPLITZ LINEAR LEAST SQUARES PROBLEMS

Let $m \geq n \geq 1$, $t_{-n+1}, \ldots, t_{m-1} \in \mathbf{C}$ 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 $b \in \mathbf{C}^m$. We consider the corresponding Toeplitz linear least squares problem (LS-problem): given T and b, determine the (unique) vector $x \in \mathbf{C}^n$ such that

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

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

Standard algorithms for solving dense linear least squares problems require $\mathcal{O}(mn^2)$ floating point operations (flops). In the case of (1), the amount of work can be reduced by taking into account the special Toeplitz structure of T. Algorithms that require only $\mathcal{O}(mn)$ flops are called *fast*. Several such algorithms have been developed. Sweet[1] was one of the first to introduce a fast algorithm. His method is not numerically stable, though. Other approaches include those by Bojanczyk, Brent and de Hoog,[2] Chun, Kailath and Lev-Ari,[3] Qiao,[4] Cybenko,[5,6] Sweet[7] and many others.

Recently, Ming Gu[8] has developed fast algorithms for solving Toeplitz and also Toeplitz-plus-Hankel linear least squares problems. He first transforms the matrix into a Cauchy-like matrix by using the Fast Fourier Transform or trigonometric transformations and then he solves the corresponding Cauchy-like linear least squares problem. 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 present 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 start by embedding the original LS-problem into an extended LS-problem, whose coefficient matrix is a circulant block matrix. By applying the Discrete Fourier Transform, we obtain a discrete least squares approximation problem for polynomial vectors, which can be solved accurately.

Other author information: (Send correspondence to M.V.B.)
M.V.B.: Email: Marc.VanBarel@cs.kuleuven.ac.be; Telephone: +32-16-327700; Fax: +32-16-327996
G.H.: Email: georg@mcs.sci.kuniv.edu.kw
P.K.: Email: Peter.Kravanja@na-net.ornl.gov

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.

2. TRANSFORMING A TOEPLITZ LINEAR LEAST SQUARES PROBLEM INTO A DISCRETE LEAST SQUARES APPROXIMATION PROBLEM

Let M be an integer that is larger than or equal to $m+n-1$. For the moment, one can choose M arbitrarily but, as will soon become clear, an appropriate choice for M is the smallest power of 2 that is $\geq m+n-1$, which assures that the Discrete Fourier Transform of size M can be computed efficiently. Let A and B be complex matrices of sizes $(M-m)\times n$ and $(M-m)\times(M-m)$, respectively. The matrix B is assumed to be nonsingular. Finally, let $a \in \mathbf{C}^{M-m}$. Instead of looking for the vector $x \in \mathbf{C}^n$ that satisfies (1), let us consider the following linear LS-problem, which we will call the *extended* LS-problem: determine the (unique) vectors $\tilde{x} \in \mathbf{C}^n$ and $y \in \mathbf{C}^{M-m}$ such that the Euclidean norm of the vector

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

is minimal. Then $\tilde{x}=x$. In other words, the first 'component' $\tilde{x}$ of the solution to the extended LS-problem coincides with the solution to the original LS-problem. This can be proved very easily. By considering the square of the Euclidean norm, we obtain that the vectors $\tilde{x}$ and y minimize

$$||A\tilde{x}+By-a||^2 + ||T\tilde{x}-b||^2. \tag{2}$$

The problem of minimizing $||r||$ has only one solution $\begin{bmatrix} \tilde{x} \\ y \end{bmatrix}$. If $\tilde{x}=x$, then the second term in the sum (2) is minimal. Since the square matrix B is nonsingular, we can reduce the first term in (2) to zero by setting $y := B^{-1}(a-A\tilde{x})$ for any given $\tilde{x}$. It follows that indeed $\tilde{x}=x$.

How to choose the matrices A and B? We propose to choose them such that the two block columns

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

are circulant matrices. For example, if we set $M := m+n-1$, then we can choose B as the identity matrix of order $n-1$ and A as the $(n-1)\times n$ Toeplitz matrix

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

with $t_{-n-k} = t_{m-k-1}$ for $k=0,1,\ldots,n-1$. We take a to be the zero vector. As we have already mentioned at the beginning of this section, a more appropriate choice for M is one that allows the Discrete Fourier Transform (of size M) of the columns of the circulant matrices C_1 and C_2 to be computed efficiently.

Define C_3 as the vector

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

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

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

is minimal. Note that the circulant block matrix $\begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix}$ is of size $M \times (n+M-m+1)$. The matrices C_1, C_2 and C_3 are circulant matrices. 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: $\|r\| = \|\mathcal{F}_M r\|$. We will therefore minimize $\|\mathcal{F}_M r\|$ instead of $\|r\|$. The following holds:

$$\begin{aligned} \mathcal{F}_M r &= \mathcal{F}_M \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} && (3) \\ &= \begin{bmatrix} \Lambda_1 \mathcal{F}_{M,n} & \Lambda_2 \mathcal{F}_{M,s} & \Lambda_3 \mathcal{F}_{M,1} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} && (4) \end{aligned}$$

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 the polynomials $x(z)$ 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 (4) now implies that the extended LS-problem can be formulated as the following discrete polynomial least squares approximation problem at roots of unity: determine the polynomials $x(z)$ and $y(z)$, where $\deg x(z) \leq n-1$ and $\deg y(z) \leq s-1$, such that

$$\sum_{k=0}^{M-1} |\lambda_{1,k} x(z_k) + \lambda_{2,k} y(z_k) + \lambda_{3,k} 1|^2 \tag{5}$$

is minimal.

3. ORTHOGONAL POLYNOMIAL VECTORS

The minimisation problem (5) can be solved within the framework of orthogonal polynomial vectors developed by Van Barel and Bultheel.[9–12] The following notation will be used: to indicate that the degree of the first component of a polynomial vector $P \in \mathbf{C}[z]^{3\times 1}$ 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 3.1 (inner product, norm). *Consider the subspace $\mathcal{P} \subset \mathbf{C}[z]^{3\times 1}$ of polynomial vectors P of degree*

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

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

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

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

$$\langle P, Q\rangle := \sum_{k=1}^{M} P^H(z_k) F_k^H F_k Q(z_k). \tag{6}$$

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 (6) 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 [12], 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 $a_1, \ldots, a_\delta \in \mathbf{C}$. 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:

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

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.

4. 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 [8]. 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.

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	1.5×10^4
	320	300	6.4×10^2	1.1×10^3	1.2×10^5
	480	450	4.7×10^2	7.7×10^2	2.5×10^5
	640	600	7.5×10^2	9.5×10^2	5.6×10^5
2	160	150	2.1×10^{16}	3.9×10^1	2.0×10^2
	320	300	1.5×10^{16}	2.8×10^0	6.2×10^2
	480	450	1.3×10^{16}	2.4×10^0	3.3×10^2
	640	600	1.3×10^{16}	2.0×10^0	2.7×10^3

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.[13,14]

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.[15] See also [16].

Theorem 4.1. *Let $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $0 \neq x \in \mathbf{R}^n$, and $r := b - Ax$. Let $\theta \in \mathbf{R}$. 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.

5. 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.

Table 2. Normwise backward error (large residuals)

Matrix type	Order		$\kappa(T)$	$\eta_F(x)/u$	
	m	n		QR	NEW
1	160	150	5.4×10^2	4.1×10^1	3.4×10^3
	320	300	4.5×10^2	6.1×10^1	3.9×10^4
	480	450	4.4×10^2	1.2×10^2	8.0×10^4
	640	600	9.1×10^2	1.0×10^2	1.5×10^5
2	160	150	2.1×10^{16}	1.3×10^2	3.9×10^0
	320	300	1.5×10^{16}	3.2×10^0	7.4×10^0
	480	450	1.3×10^{16}	2.4×10^0	7.2×10^0
	640	600	1.3×10^{16}	4.6×10^0	1.7×10^1

REFERENCES

1. D. R. Sweet, *Numerical Methods for Toeplitz matrices.* PhD thesis, University of Adelaide, Adelaide, Australia, 1982.
2. A. W. Bojanczyk, R. P. Brent, and F. R. de Hoog, "*QR* factorization of Toeplitz matrices," *Numer. Math.* **49**, pp. 81–94, July 1986.
3. J. Chun, T. Kailath, and H. Lev-Ari, "Fast parallel algorithms for *QR* and triangular factorization," *SIAM J. Sci. Stat. Comput.* **8**, pp. 899–913, Nov. 1987.
4. S. Qiao, "Hybrid algorithm for fast Toeplitz orthogonalization," *Numer. Math.* **53**, pp. 351–366, 1988.
5. G. Cybenko, "A general orthogonalization technique with applications to time series analysis and signal processing," *Math. Comp.* **40**, pp. 323–336, 1983.
6. G. Cybenko, "Fast Toeplitz orthogonalization using inner products," *SIAM J. Sci. Stat. Comput.* **8**, pp. 734–740, 1987.
7. D. R. Sweet, "Fast Toeplitz orthogonalization," *Numer. Math.* **43**, pp. 1–21, Jan. 1984.
8. M. Gu, "Stable and efficient algorithms for structured systems of linear equations," *SIAM J. Matrix. Anal. Appl.* **19**(2), pp. 279–306, 1998.
9. A. Bultheel and M. Van Barel, "Vector orthogonal polynomials and least squares approximation," *SIAM J. Matrix Anal. Appl.* **16**(3), pp. 863–885, 1995.
10. M. Van Barel and A. Bultheel, "A parallel algorithm for discrete least squares rational approximation," *Numer. Math.* **63**, pp. 99–121, 1992.
11. M. Van Barel and A. Bultheel, "Discrete linearized least squares approximation on the unit circle," *J. Comput. Appl. Math.* **50**, pp. 545–563, 1994.
12. M. V. Barel and A. Bultheel, "Orthonormal polynomial vectors and least squares approximation for a discrete inner product," *Electronic Transactions on Numerical Analysis* **3**, pp. 1–23, Mar. 1995.
13. I. Gohberg, T. Kailath, and V. Olshevsky, "Fast Gaussian elimination with partial pivoting for matrices with displacement structure," *Math. Comput.* **64**(212), pp. 1557–1576, 1995.
14. J. M. Varah, "The Prolate matrix," *Linear Algebr. Appl.* **187**, pp. 269–278, July 1993.
15. B. Waldén, R. Karlson, and J. guang Sun, "Optimal backward perturbation bounds for the linear least squares problem," *Numerical Linear Algebra with Applications* **2**(3), pp. 271–286, 1995.
16. N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, 1996.