Band-Toeplitz Preconditioners for Toeplitz Systems with Rational Generating Functions

Hung-Cheong Chan[1], Raymond H. Chan[2] and Michael K. Ng[3]

[1] *Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong*
[2] *Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong*
[3] *Computer Sciences Laboratory, Research School of Information Sciences and Engineering, The Australian National University, Canberra, ACT 0200, Australia.*

In this paper, we employ the preconditioned conjugate gradient (PCG) method to solve Toeplitz systems $A_n x = b$. Here A_n are Toeplitz matrices $\mathcal{T}_n[p/q]$ generated by functions $p(z)/q(z)$, where $p(z) = \sum_{k=-\mu_1}^{\mu_2} p_{-k} z^k$ and $q(z) = \sum_{k=-\nu_1}^{\nu_2} q_{-k} z^k$. Our approach is to precondition A_n by the matrix $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}$. We prove that the rank of the matrix $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}A_n - I_n$ is at most $(\nu_1 + \nu_2)$, where I_n is the n-by-n identity matrix. The PCG method, when applied to solving the preconditioned system, will converge sufficiently fast. In particular, we show that the solutions of $\mathcal{T}_n[p/q]x = b$ can be obtained in $O(n)$ operations.

KEY WORDS: Toeplitz matrix, banded matrix, generating function, preconditioned conjugate gradient method

1 INTRODUCTION

Toeplitz systems arise in a variety of practical applications. For example, finding unknown parameters of stationary autoregressive models in time-series analysis often involves solutions of Toeplitz systems; see King et al. [14]. In signal processing, solving Toeplitz systems is also required to obtain the filter coefficients in the design of recursive digital filters; see Chui and A. Chan [7] and Haykin [12]. There are a number of specialized fast direct methods for solving Toeplitz systems, see for instance Trench [18]. For an n-by-n Toeplitz system $A_n\mathbf{x} = \mathbf{b}$, these methods require $O(n^2)$ operations. Faster methods, which require $O(n \log^2 n)$ operations, have also been developed, see for instance Brent, Gustavson and Yun [2], and Bitmead and Anderson [1]. Recently, preconditioned conjugate gradient (PCG) methods as an iterative method for solving Toeplitz systems has received much attention. The most important result of this method is that the complexity of solving a large class of Toeplitz system can be reduced to $O(n \log n)$; see for instance Chan and Ng [5].

In 1986, G. Strang [17] proposed using preconditioned conjugate gradient method with circulant matrices as preconditioners for solving Toeplitz systems. Since then, several successful approaches of constructing circulant preconditioners have been introduced and analyzed; see for instance T. Chan [6] and R. Chan [3]. In these papers, the Toeplitz matrix A_n is assumed to be generated by a *generating function* f, i.e., the diagonals of A_n are given by the Fourier coefficients of f, see Grenander and Szegö [11]. We emphasize that this class of Toeplitz matrices arises in some practical applications. Typical examples are the kernel of the Wiener-Hopf equation, see Gohberg and Fel'dman [9], the function which gives amplitude characteristic of the recursive digital filters, see Chui and A. Chan [7], the spectral density functions in stationary stochastic process, see Grenander and Szegö [11] and the point-spread functions in image deblurring, see Jain [13]. It has been shown that if f is a positive function in the Wiener class (i.e., the Fourier coefficients of f are absolutely summable), then these circulant preconditioned systems converge superlinearly, see R. Chan [3].

In this paper, we let $\mathcal{T}_n[f]$ be the n-by-n Toeplitz matrix generated by a function f. The (j,k)th entry of $\mathcal{T}_n[f]$ is given by the (j-k)th Fourier coefficient of f, i.e.

$$[\mathcal{T}_n[f]]_{j,k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)e^{-i(j-k)\theta}d\theta, \quad 0 \le j,k \le n-1.$$

We propose a new method to construct preconditioners for Hermitian Toeplitz systems generated by rational functions. More precisely, we consider the generating functions in the form

$$f(z) = \frac{p(z)}{q(z)}, \quad \text{with} \quad z = e^{i\theta}. \tag{1}$$

Here the functions p and q are

$$p(z) = \sum_{k=-\mu}^{\mu} p_{-k}z^k \quad \text{and} \quad q(z) = \sum_{k=-\nu}^{\nu} q_{-k}z^k$$

with $p_{-i} = p_i$ for $i = 1,2,...,\mu$ and $q_{-j} = q_j$ for $j = 1,2,...,\nu$. Our approach is to use the matrices $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}$ to approximate the inverse of $\mathcal{T}_n[f]$ so that the preconditioned matrices are $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[f]$.

For the convergence rate of the preconditioned conjugate gradient (PCG) method, we prove that if the function $f(e^{i\theta})$ is positive for all θ in $[0,2\pi]$, then the

$$\text{rank}(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[f] - I_n) \le 2\nu$$

where I_n is the identity matrix of order n. Since $\mathcal{T}_n[p]$ and $\mathcal{T}_n[q]$ are banded Toeplitz matrices, the matrix-vector multiplications $\mathcal{T}_n[q]\mathbf{v}$ and $\mathcal{T}_n[p]^{-1}\mathbf{v}$ can be computed in $O(\nu n)$ and $O(\mu^2 n)$ operations by banded solvers for any vector $\mathbf{v}$; see for instance Golub and Van Loan [10]. Since $\mathcal{T}_n[f]$ is generally a dense matrix, the matrix-vector products can be computed in $O(n\log n)$ operations by using Fast Fourier Transforms (FFTs), see Chan and Ng [5].

However, Toeplitz matrices generated by $f(z)$ can be expressed as

$$\mathcal{T}_n[f] = L_1 L_2 + U_1 U_2, \tag{2}$$

where L_1 and L_2 are lower triangular band-Toeplitz matrices of lower bandwidths ℓ_1 and ℓ_2 respectively, U_1 and U_2 are upper triangular band-Toeplitz matrices of upper bandwidths u_1 and u_2 respectively; see Dickinson [8]. Here the values of ℓ_1, ℓ_2, u_1 and u_2 depends on μ and ν in $p(z)$ and $q(z)$ respectively. Therefore for any vector $\mathbf{v}$, the matrix-vector product $\mathcal{T}_n[f]\mathbf{v}$ can also be computed in $O(\tau n)$ operations, where $\tau = \ell_1 + \ell_2 + u_1 + u_2$. Hence the total operations per each iteration of the PCG method is $O((\tau + \nu + \mu^2)n)$. It follows that the complexity of using our method to solve Toeplitz systems generated by f in the form of (1) is $O(\nu(\tau + \nu + \mu^2)n)$.

We remark that our preconditioners are different from that recently proposed by Ku and Kuo [15] for these Toeplitz matrices generated by $p(z)/q(z)$. Their method basically used the decomposition (2) to construct the minimum-phase LU factorization (MPLU) preconditioner K_n for $\mathcal{T}_n[p/q]$. They proved that $\text{rank}(K_n^{-1}\mathcal{T}_n[p/q] - I_n)$ is fixed independent of n. In general, the rank depends on both the degree of p and the degree of q. On the other hand, we show that the bound for the rank of our preconditioned matrices depends on the degree of q only. Thus our method is particularly favourable for generating functions with a small degree in the denominator. In addition, the construction of such K_n requires factorization of polynomials, see Ku and Kuo [15], whereas the construction of our preconditioners is not required to do so.

The outline of the chapter is as follows. In §2, we prove the clustering property of the resulting preconditioned matrices. In §3, we mention a possible application of our method to Wiener filtering in digital signal processing. Finally, numerical examples will be given in §4.

2 CONVERGENCE ANALYSIS

Let us consider a family of Hermitian Toeplitz matrices $\mathcal{T}_n[f]$ whose generating function f is in the form

$$f(z) = \frac{a(z^{-1})}{b(z^{-1})} + \frac{c(z)}{d(z)}, \tag{3}$$

where $a(z)$, $b(z)$, $c(z)$ and $d(z)$ are polynomials in z. We remark that these Toeplitz matrices appear practically in digital signal processing and the applications will be discussed in the next section.

For convenience, we express f in (3) as the form p/q in (1) in the following discussion. In the following, we assume that $f(e^{i\theta}) > 0$ for all $\theta \in [0, 2\pi]$. We also assume that both p and q have no roots on the unit circle $|z| = 1$. We emphasize that these assumptions are always satisfied in digital signal processing, which will be discussed in §3.

Our approach is to use $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}$ to approximate the inverse of $\mathcal{T}_n[p/q]$. Now, let us discuss the convergence rate of our algorithm. Once again, we want to prove that $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[f] = I_n + L_n,$ where I_n is the identity matrix of order n and L_n is a low-rank matrix. Then the number of distinct eigenvalues of the preconditioned matrices is uniformly bounded by a constant.

In the following, after proving a prerequisite lemma, we shall show that $(\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q])$ is a low-rank matrix. The fact that $(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q] - I_n)$ is a low-rank matrix will follow. We shall also find a tight bound for the rank. We remark that in the following lemmas and theorems, the polynomials $p(z)$ and $q(z)$ are assumed to have no roots on the unit circle $|z| = 1$.

Lemma 2.1 *Let* $q(z) = \sum_{k=-\nu_1}^{\nu_2} q_k z^{-k}$ *and* $s_1, s_2, ..., s_b, t_1, t_2, ..., t_d$ *be the roots of* q *(*$b + d = \nu_1 + \nu_2 + 1$*) such that* $|s_i| < 1$ *and* $|t_i| > 1$*. Suppose*

$$\frac{1}{q(z)} = \sum_{k=-\infty}^{\infty} \alpha_k z^{-k} = \frac{q_A(z^{-1})}{q_B(z^{-1})} + \frac{q_C(z)}{q_D(z)}, \quad \max_i |s_i| < |z| < \min_i |t_i|,$$

where

$$q_B(z^{-1}) = \sum_{k=0}^{b} b_k z^{-k} \quad \text{and} \quad q_D(z) = \sum_{k=0}^{d} d_{-k} z^k.$$

Then for any positive integers i *and* n*,*

$$\sum_{j=0}^{b} b_j \mathbf{u}_{b+i-j} = \mathbf{0} \quad \text{and} \quad \sum_{j=0}^{d} d_{-j} \mathbf{v}_{d+i-j} = \mathbf{0},$$

where

$$\mathbf{u}_i = (\alpha_{n+i-1}, \alpha_{n+i-2}, ..., \alpha_i) \quad \text{and} \quad \mathbf{v}_i = (\alpha_{-i}, \alpha_{-(i+1)}, ..., \alpha_{-(n+i-1)}).$$

Proof. Let

$$q_A(z^{-1}) = \sum_{k=0}^{a} a_k z^{-k} \quad \text{and} \quad q_C(z) = \sum_{k=0}^{c} c_{-k} z^k.$$

Then

$$\frac{q_A(z^{-1})}{q_B(z^{-1})} = \alpha_0^{\ell} + \sum_{k=1}^{\infty} \alpha_k z^{-k} \quad \text{and} \quad \frac{q_C(z)}{q_D(z)} = \alpha_0^{u} + \sum_{k=-\infty}^{-1} \alpha_k z^{-k},$$

where $\alpha_0^\ell + \alpha_0^u = \alpha_0$; see Dickinson [8]. Hence

$$\begin{aligned} q_A(z^{-1}) &= \left(\alpha_0^\ell + \sum_{k=1}^{\infty} \alpha_k z^{-k}\right) q_B(z^{-1}) \\ \sum_{k=0}^{a} a_k z^{-k} &= \left(\alpha_0^\ell + \sum_{k=1}^{\infty} \alpha_k z^{-k}\right)\left(\sum_{k=0}^{b} b_k z^{-k}\right) \\ &= \alpha_0^\ell \sum_{k=0}^{b} b_k z^{-k} + \sum_{k=0}^{\infty} \gamma_k z^{-k}, \end{aligned}$$

where

$$\gamma_k = \sum_{j=0}^{b} b_j \alpha_{k-j}, \quad k \in \mathbf{Z}^+.$$

Without loss of generality, assume $b \geq a$. Then $\gamma_k = 0$ for any $k > b$. Thus for any positive integer i, we have

$$\left\{ \begin{array}{lcl} \gamma_{b+i} & = & 0 \\ \gamma_{b+i+1} & = & 0 \\ & \vdots & \\ & \vdots & \\ \gamma_{b+i+n-1} & = & 0 \end{array} \right.$$

which implies

$$\left\{ \begin{array}{lcl} \displaystyle\sum_{j=0}^{b} b_j \alpha_{(b+i)-j} & = & 0 \\ \displaystyle\sum_{j=0}^{b} b_j \alpha_{(b+i+1)-j} & = & 0 \\ & \vdots & \\ & \vdots & \\ \displaystyle\sum_{j=0}^{b} b_j \alpha_{(b+i+n-1)-j} & = & 0 \end{array} \right.$$

and then

$$\sum_{j=0}^{b} b_j \begin{pmatrix} \alpha_{(b+i)-j} \\ \alpha_{(b+i+1)-j} \\ \vdots \\ \vdots \\ \alpha_{(b+i+n-1)-j} \end{pmatrix} = \mathbf{0}.$$

Hence $\displaystyle\sum_{j=0}^{b} b_j \mathbf{u}_{b+i-j} = \mathbf{0}$. Similarly,

$$\begin{aligned}
q_C(z) &= \left(\alpha_0^u + \sum_{k=-\infty}^{-1} \alpha_k z^{-k}\right) q_D(z) \\
\sum_{k=0}^{c} c_{-k} z^k &= \left(\alpha_0^u + \sum_{k=-\infty}^{-1} \alpha_k z^{-k}\right) \left(\sum_{k=0}^{d} d_{-k} z^k\right) \\
&= \left(\alpha_0^u + \sum_{k=1}^{\infty} \alpha_{-k} z^{k}\right) \left(\sum_{k=0}^{d} d_{-k} z^k\right) \\
&= \alpha_0^u \sum_{k=0}^{d} d_{-k} z^k + \sum_{k=0}^{\infty} \delta_{-k} z^k,
\end{aligned}$$

where

$$\delta_{-k} = \sum_{j=0}^{d} d_{-j} \alpha_{-(k-j)}, \quad k \in \mathbf{Z}^+.$$

Assume $d \geq c$. Then $\delta_{-k} = 0$ for $k > d$. Thus for any positive integer i, we have

$$\left\{\begin{array}{lcl}
\delta_{-(d+i)} & = & 0 \\
\delta_{-(d+i+1)} & = & 0 \\
 & \vdots & \\
 & \vdots & \\
\delta_{-(d+i+n-1)} & = & 0
\end{array}\right.$$

which implies

$$\left\{\begin{array}{lcl}
\displaystyle\sum_{j=0}^{d} d_{-j} \alpha_{-(d+i-j)} & = & 0 \\
\displaystyle\sum_{j=0}^{d} d_{-j} \alpha_{-(d+i+1-j)} & = & 0 \\
 & \vdots & \\
 & \vdots & \\
\displaystyle\sum_{j=0}^{d} d_{-j} \alpha_{-(d+i+n-1-j)} & = & 0
\end{array}\right.$$

and then

$$\sum_{j=0}^{d} d_{-j} \begin{pmatrix} \alpha_{-(d+i-j)} \\ \alpha_{-(d+i+1-j)} \\ \vdots \\ \vdots \\ \alpha_{-(d+i+n-1-j)} \end{pmatrix} = \mathbf{0}.$$

Hence $\sum_{j=0}^{d} d_{-j}\mathbf{v}_{d+i-j} = \mathbf{0}$. □

It follows from Lemma 2.1 that the vectors $\mathbf{u}_i, \mathbf{u}_{i+1}, ..., \mathbf{u}_{i+b}$ are linearly dependent and the vectors $\mathbf{v}_i, \mathbf{v}_{i+1}, ..., \mathbf{v}_{i+d}$ are also linearly dependent for any positive integer i. Using Lemma 2.1, we now show that $(\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q])$ is a low-rank matrix.

Lemma 2.2 *Let* $p(z) = \sum_{k=-\mu_1}^{\mu_2} p_k z^{-k}$ *and* $q(z) = \sum_{k=-\nu_1}^{\nu_2} q_k z^{-k}$. *Suppose* $q(z)$ *has respectively* b *and* d *roots inside and outside the unit circle* $|z| = 1$ *(*$b + d = \nu_1 + \nu_2 + 1$*). Then for* $n > \max\{\mu_1 + \mu_2, \nu_1 + \nu_2\}$,

$$\mathrm{rank}\,(\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q]) \le \min\{\mu_1, b\} + \min\{\mu_2, d\}.$$

Proof. Let

$$\frac{1}{q(z)} = \sum_{k=-\infty}^{\infty} \alpha_k z^{-k}.$$

Then

$$\frac{p(z)}{q(z)} = \sum_{k=-\infty}^{\infty} \beta_k z^{-k},$$

where

$$\beta_k = \sum_{t=-\mu_1}^{\mu_2} p_t \alpha_{k-t}.$$

Let $\mathcal{T}_\infty[p] = [p_{ij}]$, $\mathcal{T}_\infty[1/q] = [\alpha_{ij}]$ and $X = [x_{ij}] = \mathcal{T}_\infty[p]\mathcal{T}_\infty[1/q]$. Then

$$\begin{aligned} x_{ij} &= \sum_{s=-\infty}^{\infty} p_{is}\alpha_{sj} \\ &= \sum_{s=-\infty}^{\infty} p_{i-s}\alpha_{s-j} \\ &= \sum_{t=-\infty}^{\infty} p_t \alpha_{(i-j)-t}. \end{aligned}$$

Therefore X is a Toeplitz matrix generated by $p(z)/q(z)$, i.e., $\mathcal{T}_\infty[p]\mathcal{T}_\infty[1/q] = \mathcal{T}_\infty[p/q]$. To clarify the finite case, let us introduce some notations. Define

$$M_p = \begin{pmatrix} p_{-\mu_1} & 0 & 0 & \cdots & \cdots & 0 \\ p_{-(\mu_1-1)} & p_{-\mu_1} & 0 & \cdots & \cdots & 0 \\ p_{-(\mu_1-2)} & p_{-(\mu_1-1)} & p_{-\mu_1} & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ p_{-1} & p_{-2} & \cdots & \cdots & p_{-(\mu_1-1)} & p_{-\mu_1} \end{pmatrix}$$

and

$$N_p = \begin{pmatrix} p_1 & p_2 & \cdots & \cdots & p_{\mu_2-1} & p_{\mu_2} \\ p_2 & p_3 & \cdots & \cdots & p_{\mu_2} & 0 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & 0 & 0 \\ p_{\mu_2-1} & p_{\mu_2} & 0 & \cdots & \cdots & 0 \\ p_{\mu_2} & 0 & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

For any positive integer k, let

$$E_k = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \vdots \\ \mathbf{u}_k \end{pmatrix} \quad \text{and} \quad F_k = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \vdots \\ \mathbf{v}_k \end{pmatrix},$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$, $i = 1, 2, ..., k$, are defined in Lemma 2.1. Also, for any positive integers m and n, let $\mathbf{0}_{m,n}$ denote the m-by-n zero matrix. Then it is straightforward to show that

$$\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q] = \begin{pmatrix} N_p F_{\mu_2} \\ \mathbf{0}_{(n-\mu_1-\mu_2),n} \\ M_p E_{\mu_1} \end{pmatrix}. \tag{4}$$

Hence the entries of $\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q]$ are zeros except possibly entries in its first μ_2 rows and last μ_1 rows. It follows that

$$\operatorname{rank}\left(\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q]\right) \le \mu_1 + \mu_2. \tag{5}$$

If $\mu_1 > b$, then by the result of Lemma 2.1, $\mathbf{u}_{b+1}$, $\mathbf{u}_{b+2}$, ..., $\mathbf{u}_{\mu_1}$ can be expressed as linear combinations of $\mathbf{u}_1, \mathbf{u}_2$, ..., $\mathbf{u}_b$. For $k = 1, 2, ..., (\mu_1 - b)$, let

$$\mathbf{u}_{b+k} = \sum_{j=1}^{b} f_{k,j}\mathbf{u}_j.$$

Also, for $i = 1, 2, ..., \mu_1$ and $j = 1, 2, ..., b$, define $S_{\mu_1,b} = [s_{ij}]$, where

$$s_{ij} = \begin{cases} 1 & \text{if} \quad i = j,\, i \le b \\ 0 & \text{if} \quad i \ne j,\, i \le b \\ f_{(i-b),j} & \text{if} \quad i > b \end{cases}.$$

Then $E_{\mu_1} = S_{\mu_1,b}E_b$ and hence

$$\operatorname{rank}(E_{\mu_1}) \le \operatorname{rank}(E_b) \le b \quad \text{for } \mu_1 > b. \tag{6}$$

If $\mu_2 > d$, by the result of Lemma 2.1 again, $\mathbf{v}_{d+1}$, $\mathbf{v}_{d+2}$, ..., $\mathbf{v}_{\mu_2}$ can be expressed as linear combinations of $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_d$. For $k = 1, 2, ..., (\mu_2 - d)$, let

$$\mathbf{v}_{d+k} = \sum_{j=1}^{d} g_{k,j} \mathbf{v}_j .$$

Also, for $i = 1, 2, ..., \mu_2$ and $j = 1, 2, ..., d$, define $W_{\mu_2,d} = [w_{ij}]$, where

$$w_{ij} = \begin{cases} 1 & \text{if} \quad i = j,\, i \le d \\ 0 & \text{if} \quad i \ne j,\, i \le d \\ g_{(i-d),j} & \text{if} \quad i > d \end{cases} .$$

Then $F_{\mu_2} = W_{\mu_2,d} F_d$ and hence

$$\operatorname{rank}(F_{\mu_2}) \le \operatorname{rank}(F_d) \le d \quad \text{for } \mu_2 > d. \tag{7}$$

Our final result follows from (4), (5), (6) and (7):

$$\begin{aligned} \operatorname{rank}(\mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q]) &\le \operatorname{rank}(E_{\mu_1}) + \operatorname{rank}(F_{\mu_2}) \\ &\le \min\{\mu_1, b\} + \min\{\mu_2, d\}. \end{aligned}$$

□

Now we show that $(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q] - I_n)$ is a low-rank matrix by using Lemma 2.2.

Theorem 2.3 *Let* $p(z) = \displaystyle\sum_{k=-\mu_1}^{\mu_2} p_k z^{-k}$ *and* $q(z) = \displaystyle\sum_{k=-\nu_1}^{\nu_2} q_k z^{-k}$. *Suppose* $\mathcal{T}_n[p]$ *is invertible. Then for* $n > \max\{\mu_1 + \mu_2, \nu_1 + \nu_2\}$,

$$\operatorname{rank}[(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1})\mathcal{T}_n[p/q] - I_n] \le \nu_1 + \nu_2,$$

where I_n *is the identity matrix of order* n.

Proof. Suppose $q(z)$ has respectively b and d roots inside and outside the unit circle $|z| = 1$. Let

$$R_n(p,q) = \mathcal{T}_n[p/q] - \mathcal{T}_n[p]\mathcal{T}_n[1/q].$$

With the notations defined in Lemma 2.2, we have

$$\begin{aligned} & (\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1})\mathcal{T}_n[p/q] \\ = \; & \mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}(\mathcal{T}_n[p]\mathcal{T}_n[1/q] + R_n(p,q)) \\ = \; & \mathcal{T}_n[q]\mathcal{T}_n[1/q] + \mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}R_n(p,q) \\ = \; & I_n - R_n(q,q) + \mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}R_n(p,q) \\ = \; & I_n - \begin{pmatrix} N_q W_{\nu_2,d} F_d \\ \mathbf{0}_{(n-\nu_1-\nu_2),n} \\ M_q S_{\nu_1,b} E_b \end{pmatrix} + \mathcal{T}_n[q]\mathcal{T}_n[p]^{-1} \begin{pmatrix} N_p W_{\mu_2,d} F_d \\ \mathbf{0}_{(n-\mu_1-\mu_2),n} \\ M_p S_{\mu_1,b} E_b \end{pmatrix} \end{aligned}$$

$$= \quad I_n + \left[- \begin{pmatrix} N_q W_{\nu_2,d} & \mathbf{0}_{\nu_2,(n-b-d)} & \mathbf{0}_{\nu_2,b} \\ \mathbf{0}_{(n-\nu_1-\nu_2),d} & \mathbf{0}_{(n-\nu_1-\nu_2),(n-b-d)} & \mathbf{0}_{(n-\nu_1-\nu_2),b} \\ \mathbf{0}_{\nu_1,d} & \mathbf{0}_{\nu_1,(n-b-d)} & M_q S_{\nu_1,b} \end{pmatrix} + \right.$$
$$\left. \mathcal{T}_n[q]\mathcal{T}_n[p]^{-1} \begin{pmatrix} M_p W_{\mu_2,d} & \mathbf{0}_{\mu_2,(n-b-d)} & \mathbf{0}_{\mu_2,b} \\ \mathbf{0}_{(n-\mu_1-\mu_2),d} & \mathbf{0}_{(n-\mu_1-\mu_2),(n-b-d)} & \mathbf{0}_{(n-\mu_1-\mu_2),b} \\ \mathbf{0}_{\mu_1,d} & \mathbf{0}_{\mu_1,(n-b-d)} & M_p S_{\mu_1,b} \end{pmatrix} \right]$$
$$\begin{pmatrix} F_d \\ \mathbf{0}_{(n-b-d),n} \\ E_b \end{pmatrix}.$$

Hence

$$\operatorname{rank}[(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1})\mathcal{T}_n[p/q] - I_n] \le \operatorname{rank}\left[\begin{pmatrix} F_d \\ \mathbf{0}_{(n-b-d),n} \\ E_b \end{pmatrix}\right] \le b + d = \nu_1 + \nu_2.$$

□

Now let us discuss the cost of solving $\mathcal{T}_n[f]\mathbf{x} = \mathbf{b}$ by preconditioned conjugate gradient method with $\mathcal{B}_n[p/q]$ as preconditioners, where f is in the form described in (1). The cost per iteration in the PCG method is about $10n$ operations plus the cost of computing $\mathcal{T}_n[p/q]\mathbf{v}$, $\mathcal{T}_n[q]\mathbf{v}$ and $\mathcal{T}_n[p]^{-1}\mathbf{v}$ for some vectors $\mathbf{v}$. Since the matrix $\mathcal{T}_n[p/q]$ can be expressed in terms of the sum of two products of band-Toeplitz matrices as described in (2), the matrix-vector multiplication $\mathcal{T}_n[p/q]\mathbf{v}$ can be computed in $O(n)$ operations. Since $\mathcal{T}_n[p]$ and $\mathcal{T}_n[q]$ are both band-Toeplitz matrices, the product $\mathcal{T}_n[q]\mathbf{v}$ and $\mathcal{T}_n[p]^{-1}\mathbf{v}$ can also be found in $O((\nu_1+\nu_2)n)$ and $O((\mu_1\mu_2)n)$ operations. Therefore we conclude that the cost per PCG iteration is $O((\nu_1+\nu_2+\mu_1\mu_2)n)$ operations.

Moreover, $\nu_1 = \nu_2 = \nu$, say, for Hermitian cases. As $\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}$ is not Hermitian in general, we use

$$\mathcal{B}_n[p/q] = \frac{1}{2}\left(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1} + \mathcal{T}_n[p]^{-1}\mathcal{T}_n[q]\right)$$

as our preconditioner for $\mathcal{T}_n[p/q]$ so that PCG method can be applied. Then it follows from Theorem 2.3 that

$$\operatorname{rank}(\mathcal{B}_n[p/q]\mathcal{T}_n[p/q] - I_n) \le 4\nu.$$

Thus the preconditioned matrix $\mathcal{B}_n[p/q]\mathcal{T}_n[p/q]$ can be written as the sum of the identity matrix and a matrix of rank not greater than 4ν and hence it has at most $(4\nu+1)$ distinct eigenvalues. Consequently, if we apply conjugate gradient method to solve the preconditioned system

$$\mathcal{B}_n[p/q]\mathcal{T}_n[p/q]\mathbf{x} = \mathcal{B}_n[p/q]\mathbf{b},$$

the method will converge in at most $(4\nu+1)$ iterations in exact arithmetic.

We remark that if $\mathcal{B}_n[p/q]$ is not positive definite, we can consider applying the conjugate gradient method to the following normalized preconditioning system

$$\begin{aligned} & \left(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q]\right)^* \left(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q]\right)\mathbf{x} \\ = \; & \left(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q]\right)^* \left(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\right)\mathbf{b}. \end{aligned}$$

We see that it is also possible to generalize this preconditioning technique to non-Hermitian Toeplitz matrices generated by p/q, where

$$p(z) = \sum_{k=-\mu_1}^{\mu_2} p_k z^{-k} \quad \text{and} \quad q(z) = \sum_{k=-\nu_1}^{\nu_2} q_k z^{-k}$$

with $\mu_1 \neq \mu_2$, $\nu_1 \neq \nu_2$, $p_{-k} \neq p_k$ and $q_{-k} \neq q_k$ in general. The corresponding clustering property can be determined by the following corollary of Theorem 2.3.

Corollary 2.4 *Let* $p(z) = \sum_{k=-\mu_1}^{\mu_2} p_k z^{-k}$ *and* $q(z) = \sum_{k=-\nu_1}^{\nu_2} q_k z^{-k}$. *Suppose* $\mathcal{T}_n[p]$ *is invertible. Then for* $n > 2\max\{\mu_1+\mu_2, \nu_1+\nu_2\}$,

$$\operatorname{rank}[(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q])^*(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q]) - I_n] \leq 2(\nu_1+\nu_2),$$

where I_n *is the identity matrix of order* n.

Proof. By Theorem 2.3, there exists a matrix L_q with $\operatorname{rank}(L_q) \leq (\nu_1+\nu_2)$ such that

$$\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q] = I_n + L_q.$$

Then

$$\begin{aligned} (\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q])^*(\mathcal{T}_n[q]\mathcal{T}_n[p]^{-1}\mathcal{T}_n[p/q]) &= (I_n+L_q)^*(I_n+L_q) \\ &= (I_n+L_q^*)(I_n+L_q) \\ &= I_n + L_q + L_q^*(I_n+L_q). \end{aligned}$$

Hence the result follows. □

3 APPLICATIONS TO WIENER FILTERING

In this section we consider possible applications of our methods to Wiener filtering in digital signal processing. To present the problem properly, let us introduce some terminologies used in signal processing. Let $\{x_i\}$ be a discrete-time stationary zero-mean complex-valued input process. A transversal filter of order n is of the form

$$\hat{x}_i = \sum_{k=1}^{n} w_k x_{i-k},$$

where $\hat{x}_i$ is the predicted value of x_i based on the data $\{x_k\}_{k=i-1}^{i-n}$ and $\{w_k\}_{k=1}^{n}$ are the impulse responses of the filter. It is studied in the book by Haykin [12] that given the correlation (second order statistics) of the input process $\{x_i\}$ and the cross-correlation between the input process and the desired response $\{d_i\}$, the optimal impulse responses of the filter are given by the solution of the linear system of equations

$$R_n \mathbf{w} = \mathbf{b}, \tag{8}$$

where R_n is an n-by-n Hermitian Toeplitz matrix given by

$$R_n = \begin{pmatrix} r_0 & \bar{r}_1 & \cdots & \bar{r}_{n-2} & \bar{r}_{n-1} \\ r_1 & r_0 & \bar{r}_1 & \ddots & \bar{r}_{n-2} \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{n-2} & \ddots & \ddots & \ddots & \bar{r}_1 \\ r_{n-1} & r_{n-2} & \cdots & r_1 & r_0 \end{pmatrix}$$

and $\mathbf{b}$ is a vector of the form $(b_1, b_2, \cdots, b_n)^T$. The entries $r_j = \mathcal{E}(x_i \bar{x}_{i-j})$ are the correlations of the input process and $b_j = \mathcal{E}(x_i \bar{d}_{i+j-1})$ are the cross-correlations of the input process and the desired responses, where $\mathcal{E}$ is the expectation operator. The Toeplitz system (8) is commonly called the discrete Wiener-Hopf equations, see Haykin [12].

We note that for a discrete-time stationary process, if the correlations of the process are absolutely summable, i.e. $\sum_{k=-\infty}^{\infty} |r_k| < \infty$, then r_k can be expressed in the form

$$r_k = \int_{-\pi}^{\pi} f(e^{i\theta}) e^{ik\theta} \mathrm{d}\theta.$$

Here f, called the *spectral density function* of the stationary process, is given by

$$f(z) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} r_k z^{-k},$$

where $|z| = 1$; see Haykin [12]. We note that the Toeplitz matrix R_n is generated by $f(z)$. In many practical applications in signal processing, the underlying spectral density functions of the input stochastic processes are always rational functions; see Haykin [12] and Proakis et al. [16]. These stochastic processes include, for instance, moving average (MA) processes, autoregressive (AR) processes and autoregressive-moving average (ARMA) processes. We remark that the ARMA process is driven by adding white noise to a filter with transfer function $H(z)$,

$$H(z) = \frac{\sum_{k=0}^{\mu} a_{-k} z^k}{\sum_{k=0}^{\nu} b_{-k} z^k},$$

and its corresponding spectral density function $f(z)$ is given by

$$f(z) = \sigma^2 H(z) H(z^{-1}),$$

where σ^2 is the variance of white noise. If we apply the preconditioned conjugate gradient method with our preconditioners proposed and analyzed in the previous section, the impulse responses of the discrete Wiener-Hopf equations can be efficiently determined.

4 NUMERICAL EXAMPLES

In this section, we compare the convergence rate of our preconditioner with T. Chan's circulant preconditioner[6], Ku and Kuo's MPLU factorization preconditioner[15] and the Toeplitz preconditioner $\mathcal{T}_n[1/f]$ discussed in Chan and Ng [4]. We first test their performance on the following five generating functions:

(i) $\dfrac{1-0.1z^{-1}}{1-0.8z^{-1}} + \dfrac{1-0.1z}{1-0.8z};$

(ii) $\dfrac{z^2}{(z-0.999)(1-0.999z)(z-0.5)(1-0.5z)};$

(iii) $\dfrac{\sum_{k=1}^{8} kz^{-k} + 10 + \sum_{k=1}^{8} kz^{k}}{z^{-1}+3+z};$

(iv) $\dfrac{z^{-1}+3+z}{\sum_{k=1}^{8} kz^{-k} + 10 + \sum_{k=1}^{8} kz^{k}};$

(v) $\dfrac{(1-0.2z^{-1})(1+0.3z^{-1})(1-0.5z^{-1})}{(1-0.7z^{-1})(1+0.5z^{-1})} + \dfrac{1+2z}{(1.5-z)(2+z)(2-z)}.$

We remark that the Toeplitz matrices generated by the functions (i)–(iv) are symmetric while the Toeplitz matrices generated by the function (v) are non-symmetric. In the test, we used the vector of all ones as the right-hand-side vector and the zero vector as the initial guess. The stopping criterion is $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 \le 10^{-7}$, where $\mathbf{r}_k$ is the residual vector after k iterations. All computations are done by Matlab (V4.2) on an HP 715/50 workstation. Tables 1–4 show the numbers of CG iterations required for convergence with different choices of preconditioners, while Table 5 shows the number of CG iterations for the normalized preconditioned systems. In the tables, I denotes that no preconditioners are used, C is T.Chan's circulant preconditioner[6], K is Ku and Kuo's MPLU factorization preconditioner[15], T is the Toeplitz preconditioner $\mathcal{T}_n[1/f]$ discussed in the paper Chan and Ng [4] and B is our proposed band-Toeplitz product preconditioner. We remark that the cost of the method using circulant preconditioner is $O(n \log n)$ operations per CG iteration, whereas the others are just $O(n)$ operations per CG iteration.

As for the comparison of floating point operations in conjugate gradient iterations for different preconditioners, Table 6 shows the numbers of matrix-vector products and solvers required in the PCG iteration for the n-by-n band-Toeplitz,

TABLE 1: Number of CG iterations for generating function.

n	I	C	K	T	B
16	6	5	2	2	2
32	9	5	2	2	2
64	11	5	2	2	2
128	15	5	2	2	2
256	18	4	2	2	2

TABLE 2: Number of CG iterations for generating function (ii).

n	I	C	K	T	B
16	8	6	3	3	3
32	16	9	3	3	3
64	42	10	3	3	3
128	112	11	3	4	3
256	276	12	3	4	3

banded triangular Toeplitz preconditioners. We remark that the number of floating point operations required of band-Toeplitz, banded triangular Toeplitz matrix-vector products and solvers can be found in Golub and van Loan [10]. For simplicity, we list their number of floating point operations in Table 7. Tables 8–10 show the bandwidths of band-Toeplitz, banded triangular Toeplitz matrices involved and their corresponding number of floating point operations for different examples. We note that the cost per iteration of using various preconditioners are almost the same except for generating function (vi).

From the numerical results, we see that the convergence performances of our preconditioners and MPLU factorization preconditioner are both better than those of T. Chan's circulant preconditioner and without using preconditioner. We note that the inverse of a Toepltiz matrix generated by rational function is non-Toeplitz, but it is closely related to Toepltiz matrices. Since our preconditioners and MPLU factorization preconditioner are the approximations of the inverse of the Toepltiz matrix, we expect that their performance are better than that of circulant preconditioner. For the generating function (iii), number of iterations required for convergence using our band-Toeplitz product preconditioners is less than that using MPLU factorization preconditioner. Therefore, we see that our method is useful for generating functions with small ν.

Next we apply our method to a finite impulse response (FIR) system identification problem. Figure 1 is a block diagram of an FIR system identification model. The input signal x_k drives the unknown system to produce the output sequence y_k. We model the unknown system as an FIR filter. If the unknown system is actually an FIR system, then the model is exact. We formulate the discrete Wiener-Hopf

TABLE 3: Number of CG iterations for generating function (iii).

n	I	C	K	T	B
16	8	8	6	6	3
32	16	9	6	6	3
64	24	10	6	6	3
128	31	9	6	6	3
256	38	9	6	6	3

TABLE 4: Number of CG iterations for generating function (iv).

n	I	C	K	T	B
16	8	8	6	6	6
32	15	11	6	6	6
64	26	11	6	6	6
128	47	10	6	6	6
256	56	10	6	6	6

TABLE 5: Number of CG iterations for generating function (v).

n	I	C	K	T	B
16	11	8	6	6	6
32	13	8	5	6	6
64	17	7	5	6	6
128	23	7	5	6	6
256	24	7	5	6	6

TABLE 6: The number of matrix-vector (MV) products and solvers (SO) required in each PCG iteration for different preconditioners.

Matrices	K	T	B
Band-Toeplitz B_1: MV	0	0	1
Band-Toeplitz B_2: SO	0	0	1
Banded Lower Triangular Toeplitz L_1: MV	1	1	0
Banded Lower Triangular Toeplitz L_2: SO	1	1	0
Banded Upper Triangular Toeplitz U_1: MV	1	1	0
Banded Upper Triangular Toeplitz U_2: SO	1	1	0

TABLE 7: Flop counts of matrix-vector (MV) products and solvers (SO) for band Toeplitz type preconditioners with lower bandwidths p and upper bandwidths q.

Matrices	lower	upper	flops
Band-Toeplitz B_1: MV	p	q	$2n(p+q)$
Band-Toeplitz B_2: SO	p	q	$2n(pq+p+q)$
Banded Lower Triangular Toeplitz L_1: MV	p	0	$2np$
Banded Lower Triangular Toeplitz L_2: SO	p	0	$2np$
Banded Upper Triangular Toeplitz U_1: MV	0	q	$2nq$
Banded Upper Triangular Toeplitz U_2: SO	0	q	$2nq$

TABLE 8: The lower bandwidths of L_1, L_2 and the upper bandwidths of U_1, U_2 for Examples (i)–(v) and the corresponding flops for MPLU factorization preconditioner K.

	L_1	L_2	U_1	U_2	flops
(i)	1	1	1	1	$8n$
(ii)	2	2	2	2	$16n$
(iii)	1	8	1	8	$36n$
(iv)	1	8	1	8	$36n$
(v)	3	3	3	3	$24n$

TABLE 9: The lower bandwidths of L_1, L_2 and the upper bandwidths of U_1, U_2 for Examples (i)–(v) and the corresponding flops for Toeplitz preconditioner T.

	L_1	L_2	U_1	U_2	flops
(i)	1	1	1	1	$8n$
(ii)	2	3	3	2	$20n$
(iii)	1	8	1	8	$32n$
(iv)	8	1	8	1	$32n$
(v)	2	3	3	2	$20n$

TABLE 10: The lower and upper bandwidths of B_1 and B_2 for Examples (i)–(v); and the corresponding flops for our preconditioner B.

	B_1		B_2		flops
	lower	upper	lower	upper	
(i)	1	1	1	1	$10n$
(ii)	0	0	2	2	$16n$
(iii)	8	8	1	1	$38n$
(iv)	1	1	8	8	$164n$
(v)	3	3	2	3	$34n$

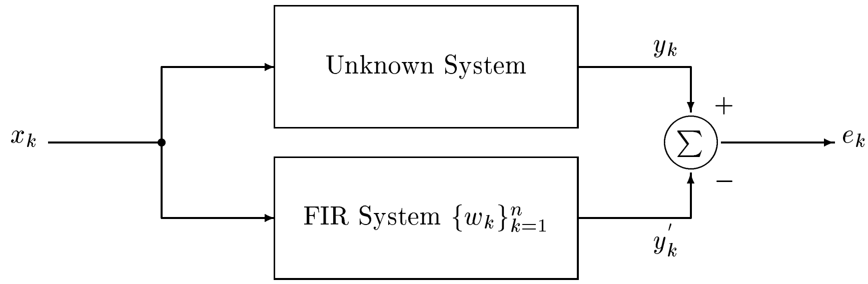

FIGURE 1: FIR System Identification Model

n	I	C	K	T	B
8	4	4	2	2	2
16	8	4	2	2	2
32	13	4	2	2	2
64	17	4	2	2	2
128	19	4	2	2	2

TABLE 11: Number of iterations for ARMA process.

equations (8) by using the known statistics (correlations) of the input process and estimating the cross-correlations between the input process and the desired responses from the data samples to construct the Toeplitz matrix and the right-hand side vector respectively. By solving these discrete Wiener-Hopf equations, the FIR system impulse responses can be found.

We used ARMA process as the input process and considered transfer function $H(z)$ given by

$$H(z) = \frac{10}{\sqrt{2}}\left(\frac{z-0.01}{z-0.5}\right).$$

The corresponding spectral density function (generating function of Toeplitz matrices) is

$$f(z) = \frac{-z+100.01-z^{-1}}{-z+2.5-z^{-1}}.$$

Table 6 shows the number of iterations for different orders of the FIR filter. We see that our methods converge very fast.

ACKNOWLEDGEMENTS

Research by R. Chan was supported by HKRGC grants no. CUHK 316/94E. Research by R. M. Ng was supported the Cooperative Research Centre for Advanced Computational Systems.

REFERENCES

1. R. Bitmead and B. Anderson, *Asymptotically Fast Solution of Toe-plitz and Related Systems of Equations*, Lin. Alg. Appl., V34 (1980), pp. 103-116.
2. R. Brent, F. Gustavson, and D. Yun, *Fast Solution of Toeplitz Systems of Equations and Computations of Pade Approximations*, J. Algorithms, V1 (1980), pp. 259-295.
3. R. Chan, *Circulant Preconditioners for Hermitian Toeplitz Systems*, SIAM J. Matrix Anal. Appl., V10 (1989), pp. 542–506.
4. R. Chan and M. Ng, *Toeplitz Preconditioners for Hermitian Toeplitz Systems*, Lin. Alg. Appl., V190 (1993), pp. 181–208.
5. R. Chan and M. Ng, *Conjugate Gradient Methods for Toeplitz Systems*, SIAM Review, V38, No. 3 (1996), pp. 427–482.
6. T. Chan, *An Optimal Circulant Preconditioner for Toeplitz Systems*, SIAM J. Sci. Statist. Comput., V9 (1988), pp. 766–771.
7. C. Chui and A. Chan, *Application of Approximation Theory Methods to Recursive Digital Filter Design*, IEEE Trans. Acoust. Speech Signal Process, V30 (1982), pp. 18-24.
8. B. Dickinson, *Solution of Linear Equations with Rational Toeplitz Matrices*, Mathematics of Computation, V34 (1980), pp. 227-233.
9. I. Gohberg and I. Feĺman, *Convolution Equations and Projection Methods for Their Solution*, Transl. Math. Monographs, Vol. 41, Amer. Math. Soc., Providence, RI, 1974.
10. G. Golub and C. Van Loan, *Matrix Computations* , 2nd Ed., The Johns Hopkins University Press, Maryland, 1989.
11. U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*, 2nd Ed., Chelsea Pub. Co., New York, 1984.
12. S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, New Jersey, 1991.
13. A. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliiffs, NJ, 1989.
14. R. King, M. Ahmadi, R.Gorgui-Naguib, A. Kwabwe, and M. Azimi-Sadjadi, *Digital Filtering in One and Two Dimensions: Design and Applications*, Plenum, New York, 1989.
15. K. Ku and C. Kuo, *Minimum-Phase LU Factorization Preconditioners for Toeplitz Matrices*, SIAM J. Sci. Stat. Comp., V13 (1992), pp.1470–1487.
16. J. Proakis, C. Rader, F. Ling and C. Nikias, *Advanced Digital Signal Processing*, Maxwell Macmillan, 1992.
17. G. Strang, *A Proposal for Toeplitz Matrix Calculations*, Stud. Appl. Math., V74 (1986), pp. 171–176.
18. W. Trench, *An Algorithm for the Inversion of Finite Toeplitz Matrices* , SIAM J. Appl. Math., V12 (1964), pp. 515–522.