USC-SIPI REPORT #175

Spectral Properties of Preconditioned Rational Toeplitz Matrices: The Nonsymmetric Case

by

Ta-Kang Ku and C.-C. Jay Kuo

April 1991

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
Powell Hall of Engineering
University Park/MC-0272
Los Angeles, CA 90089 U.S.A.

Submitted to SIAM Journal on Matrix Analysis and Applications.

SPECTRAL PROPERTIES OF PRECONDITIONED RATIONAL TOEPLITZ MATRICES : THE NONSYMMETRIC CASE *

TA-KANG KU† AND C.-C. JAY KUO†

Abstract. Various preconditioners for symmetric positive-definite (SPD) Toeplitz matrices in circulant matrix form have recently been proposed. The spectral properties of the preconditioned SPD Toeplitz matrices have also been studied. In this research, we apply Strang's preconditioner S_N and our preconditioner K_N to an $N \times N$ nonsymmetric (or nonhermitian) Toeplitz system $T_N\mathbf{x} = \mathbf{b}$. For a large class of Toeplitz matrices, we prove that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except a fixed number independent of N. If T_N is additionally generated by a rational function, we are able to characterize the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ directly. Let the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ be classified into the outliers and the clustered eigenvalues depending on whether they converge to 1 asymptotically. Then, the number of outliers depends on the order of the rational generating function, and the clustering radius is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner. Numerical experiments are provided to illustrate our theoretical study.

Key words. Toeplitz, circulant, nonsymmetric, preconditioners, preconditioned iterative method, CGN, CGS, GMRES.

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction. Research on preconditioning symmetric positive-definite (SPD) Toeplitz matrices with circulant matrices has been active recently [1], [3], [5], [6], [13]. In this research, we generalize Strang's preconditioner S_N [13] and our preconditioner K_N [6] to nonsymmetric (or nonhermitian) Toeplitz matrices. Let T_N be an $N \times N$ nonsymmetric Toeplitz matrix with elements $t_{i,j} = t_{i-j}$. The generalized Strang's preconditioner S_N is obtained by preserving N consecutive diagonals in T_N, i.e. diagonals with elements $t_n, 1 - M \leq n \leq N - M$, and using them to form a circulant matrix. One simple rule to determine M is to choose its value such that $|t_{N-M}| \approx |t_{1-M}|$. Note that half of the elements in T_N are not used in constructing S_N. The generalized preconditioner K_N is obtained from a $2N \times 2N$ circulant matrix in such a way that all elements in T_N are used, and is a circulant matrix itself (See §2). Since S_N and K_N are circulant, the matrix-vector products $S_N^{-1}\mathbf{v}$ and $K_N^{-1}\mathbf{v}$ can be conveniently computed via Fast Fourier Transform (FFT) with $O(N \log N)$ operations. The system of equations associated with the preconditioned Toeplitz matrix is then solved by iterative methods such as CGN (the Conjugate Gradient iteration applied to the Normal equations) [4], GMRES (the Generalized Minimal Residual) [11], and CGS (the Conjugate Gradient Squared) [12].

The convergence rate of preconditioned iterative methods depends on the singular value or eigenvalue distribution of the preconditioned matrices [10]. The spectral properties of preconditioned SPD Toeplitz matrices have been widely studied. Chan and Strang [1] [2] proved that, for a symmetric Toeplitz with a positive generating function in the Wiener class, the preconditioned matrix has eigenvalues clustered around unity except a fixed number independent of N. If the Toeplitz is additionally generated by a rational function, even stronger results were proved by Trefethen [15] and the authors [8]. In contrast, relatively few results for preconditioned nonsymmetric Toeplitz have been obtained so far [9], [17].

In this research, we examine the spectral properties of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ for nonsymmetric T_N in general, and nonsymmetric rational T_N in particular. The main results of our study are stated as

* This version is dated April 12, 1991. This work was supported by the USC Faculty Research and Innovation Fund and by a National Science Foundation Research Initiation Award.

† Signal and Image Processing Institute and Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-0272. E-mail:tkku@sipi.usc.edu and cckuo@sipi.usc.edu

follows. For a large class of general Toeplitz matrices, we prove that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, or equivalently, the eigenvalues of $(S_N^{-1}T_N)^H(S_N^{-1}T_N)$ and $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$, are clustered around unity except a fixed number independent of N. If T_N is additionally generated by a rational function of order $(\alpha, \beta, \gamma, \delta)$, we are able to characterize the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ directly. We classify the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ into two classes, i.e. the outliers and the clustered eigenvalues, depending on whether they converge to 1 asymptotically. Then, (1) the number of outliers is at most $\eta = 2\min(r, s)$ where $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$; and (2) the clustered eigenvalues are confined in a disk centered at 1 with radius ϵ, where the clustering radius ϵ is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner.

With these spectral regularities, we can find appropriate preconditioned iterative methods to solve a nonsymmetric Toeplitz system efficiently. In particular, an $N \times N$ rational Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ can be solved with $O(N \log N)$ operations since the number of iterations required for convergence is independent of the problem size N. To compare the performance of S_N and K_N, the $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have the same number of outliers so that they converge in the same number of iterations asymptotically. However, the performances of S_N and K_N for finite N are determined by the clustering radii of the clustered eigenvalues as well. The magnitudes of the last elements used to construct S_N and K_N are $O(|t_{N-M}| + |t_{1-M}|)$ and $O(|t_N| + |t_{-N}|)$, respectively. Since $O(|t_N| + |t_{-N}|) \leq O(|t_{N-M}| + |t_{1-M}|)$ for large N, iterative methods with preconditioner K_N converges faster than with preconditioner S_N for solving rational Toeplitz systems. This is confirmed by numerical experiments. By the parallelism provided by FFT, the iterative methods with preconditioners in circulant matrix form is highly parallelizable, and the time complexity of the method can be reduced to $O(\log N)$ if $O(N)$ processors are used.

When T_N is a symmetric rational Toeplitz, we have $r = s$ and $t_N = t_{-N}$. Consequently, the number of outliers of $K_N^{-1}T_N$ is $\eta = 2r = 2\max(\alpha, \beta)$ and the clustering radius is $O(|t_N|)$. They reduce to the case given in [8]. Although the results derived in this paper can be viewed as a generalization of the results in [8], we want to point out that the approach adopted in this research is very different from that in [8] and the proof techniques are much more involved. For example, in characterizing the clustering radius of clustered eigenvalues of $K_N^{-1}T_N$ (or $S_N^{-1}T_N$) for symmetric T_N, the intertwinning theorem of eigenvalues was exploited in [8]. However, such a theorem does not exist for nonsymmetric matrices so that we use perturbation theory for eigenvalues instead.

It is worthwhile to mention that there exists a preconditioner based on the minimum-phase LU factorization (MPLU) technique [9] which has a faster or comparable convergence rate than preconditioners S_N and K_N. However, Toeplitz preconditioners in circulant matrix form have two advantages over the MPLU preconditioner. First, the circulant preconditioning technique can be easily generalized to multidimensional Toeplitz systems. See [7] for the two-dimensional case (block Toeplitz matrices). Second, the resulting preconditioned iterative method with preconditioners in circulant form is highly parallelizable while the MPLU preconditioner has to be implemented sequentially.

This paper is organized as follows. The construction of preconditioners S_N and K_N for nonsymmetric Toeplitz T_N is discussed in §2. We describe the singular value distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for general Toeplitz in §3, and characterize the eigenvalue distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for rational Toeplitz in §4 and §5, respectively. Numerical experiments are given in §6 to illustrate the theoretical study.

2. Constructions of Toeplitz preconditioners.

Let T_m be a sequence of $m \times m$ nonsymmetric Toeplitz matrices with generating sequence t_n. Then,

$$T_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{-(N-2)} & t_{-(N-1)} \\ t_1 & t_0 & t_{-1} & \cdot & t_{-(N-2)} \\ \cdot & t_1 & t_0 & \cdot & \cdot \\ t_{N-2} & \cdot & \cdot & \cdot & t_{-1} \\ t_{N-1} & t_{N-2} & \cdot & t_1 & t_0 \end{bmatrix}.$$

Following the idea proposed by Strang [13], we construct the preconditioner S_N by preserving N consecutive diagonals in T_N and bringing them around to form a circulant matrix,

$$S_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 & t_1 \\ t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} \\ t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} \\ t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} \\ t_{-1} & t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 \end{bmatrix}.$$

A simple rule of thumb to decide the value of M is to require $|t_{N-M}| \approx |t_{1-M}|$.

Generalizing the idea in [6], the preconditioner K_N is constructed based on a $2N \times 2N$ circulant matrix R_{2N},

$$R_{2N} = \begin{bmatrix} T_N & \Delta T_N \\ \Delta T_N & T_N \end{bmatrix},$$

where ΔT_N is determined by the elements of T_N to make R_{2N} circulant, i.e.,

$$\Delta T_N = \begin{bmatrix} 0 & t_{N-1} & \cdot & t_2 & t_1 \\ t_{-(N-1)} & 0 & t_{N-1} & \cdot & t_2 \\ \cdot & t_{-(N-1)} & 0 & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & 0 \end{bmatrix}.$$

This construction is motivated by the observation that the augmented circulant system,

$$\begin{bmatrix} T_N & \Delta T_N \\ \Delta T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{b} \end{bmatrix},$$

is equivalent to $(T_N + \Delta T_N)\mathbf{x} = \mathbf{b}$ so that $(T_N + \Delta T_N)^{-1}\mathbf{b}$ can be computed efficiently via FFT and

$$K_N = T_N + \Delta T_N \tag{2.1}$$

can be used as a preconditioner for T_N. Note, however, that K_N itself is also circulant and can be inverted directly via N-point FFT rather than $2N$-point FFT.

3. Spectral properties of preconditioned Toeplitz. We assume that the generating sequence t_n satisfies the following two conditions:

$$\sum_{-\infty}^{\infty} |t_n| \le B_T < \infty, \tag{3.1}$$

$$|T(e^{i\theta})| = \left| \sum_{-\infty}^{\infty} t_n e^{-in\theta} \right| \ge \mu_T > 0, \qquad \forall\theta. \tag{3.2}$$

Since $T(e^{i\theta})$ describes the asymptotic eigenvalue distribution of T_N, the above conditions imply that $||T_N||$ and $||T_N^{-1}||$ are bounded for large N and, consequently, T_N is well conditioned.

With the above conditions, the preconditioners K_N and S_N are also well conditioned for sufficiently large N due to the following theorem.

THEOREM 1. *Let T_N be an $N \times N$ Toeplitz matrix with the corresponding generating sequence satisfying (3.1) and (3.2). The $||(K_N K_N^H)^{-1}||_2$ and $||(S_N S_N^H)^{-1}||_2$ are bounded for sufficiently large N.*

Proof. Since K_N is circulant, we have

$$K_N = F_N^H D_N F_N \quad \text{and} \quad K_N^H = F_N^H D_N^H F_N,$$

where F_N is the $N \times N$ unitary Fourier matrix with $N^{-1/2} e^{-i2\pi(m-1)(n-1)/N}$ as the (m,n) element and D_N a diagonal matrix formed by the eigenvalues of K_N. Thus, K_N, K_N^H and $K_N K_N^H$ share the same eigenvectors, and the eigenvalues of $K_N K_N^H$ are

$$\lambda(K_N K_N^H) = \lambda(K_N)\lambda^*(K_N) = |\lambda(K_N)|^2.$$

Any eigenvalue of K_N belongs to the set of eigenvalues of R_{2N}, which are

$$\rho_n = \lambda_n(R_{2N}) = \sum_{k=-(N-1)}^{N-1} t_k e^{i2\pi kn/2N}, \qquad 1 \le n \le 2N.$$

It is clear that ρ_n is a partial sum of the infinite series $\sum_{-\infty}^{\infty} t_k e^{-ik\theta}$ with $\theta = -n\pi/N$. With (3.2), $|\rho_n| \ge \mu_T - \mu$, where μ can be made arbitrarily small by choosing sufficiently large N so that

$$\|(K_N K_N^H)^{-1}\|_2 \le \frac{1}{(\mu_T - \mu)^2} < \infty.$$

Similar arguments can be used to prove the boundness of $\|(S_N S_N^H)^{-1}\|_2$, and the proof is completed. □

The next theorem describes the clustering property of the singular values of $K_N^{-1}T_N$ and $S_N^{-1}T_N$.

THEOREM 2. *Let T_N be an $N \times N$ Toeplitz matrix with the generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the singular values of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are clustered around unity except a fixed number independent of N*

Proof. Note that the singular value of $K_N^{-1}T_N$ is equal to the square root of the corresponding eigenvalue of $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$. Since $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$ and $(K_N K_N^H)^{-1}(T_N T_N^H)$ are similar, the eigenvalues of $(K_N K_N^H)^{-1}(T_N T_N^H)$ are examined to understand the singular values of $K_N^{-1}T_N$. With the relation $K_N = T_N + \Delta T_N$, we have

$$\lambda[(K_N K_N^H)^{-1}(T_N T_N^H)] = 1 - \lambda[(K_N K_N^H)^{-1}(K_N \Delta T_N^H + \Delta T_N K_N^H - \Delta T_N \Delta T_N^H)].$$

Let us define

$$W_N = K_N \Delta T_N^H + \Delta T_N K_N^H - \Delta T_N \Delta T_N^H,$$

and denote the corresponding $(N-2q) \times (N-2q)$ central diagonal block of $(K_N K_N^H)^{-1}$ and W_N by $\mathcal{K}_{N-2q}^{-1}$ and $\mathcal{W}_{N-2q}$, respectively. By the separation theorem (or intertwining theorem) of eigenvalues [14], [16], there are at least $N-4q$ eigenvalues of $(K_N K_N^H)^{-1}W_N$ bounded by the minimum and the maximum eigenvalues of $\mathcal{K}_{N-2q}^{-1}\mathcal{W}_{N-2q}$.

Since $\mathcal{K}_{N-2q}^{-1}$ is a submatrix of the symmetric circulant matrix $(K_N K_N^H)^{-1}$,

$$\|\mathcal{K}_{N-2q}^{-1}\|_2 \le \|(K_N K_N^H)^{-1}\|_2.$$

According to the definition of $\mathcal{W}_{N-2q}$,

$$\mathcal{W}_{N-2q} = \mathcal{K}\Delta\mathcal{T}^H + \Delta\mathcal{T}\mathcal{K}^H - \Delta\mathcal{T}\Delta\mathcal{T}^H,$$

where $\mathcal{K}$ and $\Delta\mathcal{T}$ are $(N-2q) \times N$ matrices formed by the central $(N-2q)$ rows of K_N and ΔT_N, respectively. It is easy to verify that, for $p = 1, \infty$,

$$\|\mathcal{K}\|_p \le 2 \sum_{n=-(N-1)}^{N-1} |t_n| \le 2 \sum_{n=-\infty}^{\infty} |t_n| \le 2B_T < \infty,$$

and

$$||\Delta\mathcal{T}||_p \le \sum_{n=q+1}^{N-1} (|t_n| + |t_{-n}|) \le \sum_{n=q+1}^{\infty} (|t_n| + |t_{-n}|) = \sigma(q).$$

Since $||A||_2 \le (||A||_1||A||_\infty)^{1/2}$ for an arbitrary matrix A, the above bounds also hold for $p = 2$. Similarly, we can argue that $||\mathcal{K}^H||_2 \le 2B_T < \infty$ and $||\Delta\mathcal{T}^H||_2 \le \sigma(q)$. Thus,

$$\begin{aligned} ||\mathcal{W}_{N-2q}||_2 &\le ||\mathcal{K}||_2||\Delta\mathcal{T}^H||_2 + ||\Delta\mathcal{T}||_2||\mathcal{K}^H||_2 + ||\Delta\mathcal{T}||_2||\Delta\mathcal{T}^H||_2 \\ &\le 4B_T\sigma(q) + \sigma^2(q). \end{aligned}$$

By using Theorem 1 and the fact that $\sigma(q)$ is smaller as q becomes larger due to (3.1), we conclude that for given ϵ there exist q and $\tilde{N}$ such that for all $N \ge \tilde{N}$,

$$||\mathcal{K}_{N-2q}^{-1}||_2||\mathcal{W}_{N-2q}||_2 \le ||(K_N K_N^H)^{-1}||_2||\mathcal{W}_{N-2q}||_2 \le \epsilon.$$

Hence, the eigenvalues of $(K_N K_N^H)^{-1}(T_N T_N^H)$ are confined in the interval $(1-\epsilon, 1+\epsilon)$ except at most $4q$ outlying eigenvalues. Similar arguments can be used to prove the spectral clustering property of the singular values of $S_N^{-1}T_N$. □

With the above spectral clustering property, a Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ can be solved effectively by applying the CGN method to the preconditioned system $K_N^{-1}T_N\mathbf{x} = K_N^{-1}\mathbf{b}$ or $S_N^{-1}T_N\mathbf{x} = S_N^{-1}\mathbf{b}$. When the generating function is additionally rational, we characterize the eigenvalues of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ directly. It will be detailed in the following sections.

4. Spectral properties of preconditioned rational Toeplitz $K_N^{-1}T_N$. The generating function of a sequence of Toeplitz matrices T_m is defined as

$$T(z) = \sum_{n=-\infty}^{\infty} t_n z^{-n}.$$

Let the generating function of T_N be of the form

$$T(z) = \frac{A(z^{-1})}{B(z^{-1})} + \frac{C(z)}{D(z)}, \tag{4.1}$$

where

$$\frac{A(z^{-1})}{B(z^{-1})} = \frac{a_0 + a_1 z^{-1} + \cdots + a_\alpha z^{-\alpha}}{1 + b_1 z^{-1} + \cdots + b_\beta z^{-\beta}}, \qquad \frac{C(z)}{D(z)} = \frac{c_0 + c_1 z + \cdots + c_\gamma z^\gamma}{1 + d_1 z + \cdots + d_\delta z^\delta}.$$

Note that $a_\alpha b_\beta c_\gamma d_\delta \ne 0$ and polynomials $A(z^{-1})$ and $B(z^{-1})$ (or $C(z)$ and $D(z)$) have no common factor. We call $T(z)$ a rational function of order $(\alpha, \beta, \gamma, \delta)$ and T_N a rational Toeplitz matrix. To simplify the notation, we define $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$.

The spectral properties of $K_N^{-1}T_N$ can be determined from that of $T_N^{-1}\Delta T_N$ via

$$[\lambda(K_N^{-1}T_N)]^{-1} = \lambda(T_N^{-1}(T_N + \Delta T_N)) = 1 + \lambda(T_N^{-1}\Delta T_N). \tag{4.2}$$

The eigenvalues of $K_N^{-1}T_N$ clustered around 1 correspond to those of $T_N^{-1}\Delta T_N$ clustered around 0. We summarize the procedures in examing the spectral properties of $T_N^{-1}\Delta T_N$ as follows:

Step 1: Show that the ΔT_N is asymptotically equivalent to a low rank Toeplitz matrix ΔF_N (Lemma 2).

Step 2: Study the rank of ΔF_N by transforming it to a matrix Q_F which has at most $d = r + s$ nonzero columns (Lemma 3).

Step 3: Show that the Q_F is asymptotically equivalent to a matrix $\overline{Q}_F$ which has at most $2\min(r, s)$ nonzero eigenvalues (Lemma 4).

Step 4: Use perturbation theory to determine the radius of the clustered eigenvalues of $T_N^{-1}\Delta T_N$ and $K_N^{-1}T_N$ (Lemmas 5,6 and Theorem 3).

The number of outliers of $K_N^{-1}T_N$, i.e. $2\min(r, s)$, is determined from Steps 1-3, and the clustering radius is determined from Step 4.

4.1. The number of outliers of $K_N^{-1}T_N$. Note that the sequence t_n can be recursively calculated for large $|n|$. This is stated as follows.

LEMMA 1. *The sequence t_n generated by (4.1) follows the recursions,*

$$t_{n+1} = -(b_1 t_n + b_2 t_{n-1} + \cdots + b_\beta t_{n-\beta+1}), \qquad n \geq r = \max(\alpha, \beta),$$
$$t_{n-1} = -(d_1 t_n + d_2 t_{n+1} + \cdots + d_\delta t_{n+\delta-1}), \qquad n \leq -s = -\max(\gamma, \delta). \tag{4.3}$$

Proof. Similar to the proof of Lemma 1 in [8]. □

Since elements t_n satisfy the recursion given in Lemma 1, we construct a low rank Toeplitz matrices ΔF_N as

$$\Delta F_N = F_{1,N} + F_{2,N}, \tag{4.4}$$

where

$$F_{1,N} = \begin{bmatrix} t_N & t_{N-1} & \cdot & t_2 & t_1 \\ t_{N+1} & t_N & t_{N-1} & \cdot & t_2 \\ \cdot & t_{N+1} & t_N & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N \end{bmatrix},$$

and

$$F_{2,N} = \begin{bmatrix} t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{-(N-1)} & t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{-(N-1)} & t_{-N} & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & t_{-N} \end{bmatrix},$$

and where $t_n, n \geq r$ or $n \leq -s$, are recursively defined by (4.3). Due to the recursion given by (4.3), the ranks of $F_{1,N}$ and $F_{2,N}$ are bounded by r and s, respectively. Thus, the rank of ΔF is bounded by $d = r + s$. The following lemma shows that ΔT_N and ΔF_N are in fact asymptotically equivalent.

LEMMA 2. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The ΔT_N and ΔF_N are asymptotically equivalent.*

Proof. Let us denote the difference between ΔF_N and ΔT_N by

$$\Delta E_N = \Delta F_N - \Delta T_N = \begin{bmatrix} t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{N+1} & t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{N+1} & t_N + t_{-N} & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N + t_{-N} \end{bmatrix}. \tag{4.5}$$

It can be easily verified that the l_1 and l_∞ norms of ΔE_N are both bounded by

$$\tau_E = \sum_{n=N}^{2N-1} |t_n| + \sum_{n=-N}^{-(2N-1)} |t_n|. \tag{4.6}$$

Consequently, we have

$$||\Delta E_N||_2 \leq (||\Delta E_N||_1 ||\Delta E_N||_\infty)^{1/2} \leq \tau_E.$$

Since τ_E goes to zero as N goes to infinity due to (3.1), the proof is completed. □

Since ΔT_N is asymptotically equivalent to ΔF_N and the rank of ΔF_N is bounded by d, the number of outliers of $T_N^{-1}\Delta T_N$ (or $K_N^{-1}T_N$) is bounded by d, which is however not tight. We are able to determine a tighter bound by introducing another asymptotically equivalent matrix of ΔT_N (or ΔF_N), which has only $2\min(r, s)$ nonzero eigenvalues in the following. This turns out to be the exact number of outliers actually observed in all our numerical experiments. To exploit the low rank structure of ΔF_N, we transform ΔF_N to

$$Q_F = \Delta F_N U_D L_B, \tag{4.7}$$

where U_D is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $D(z)$ as the first row, and L_B is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $B(z^{-1})$ as the first column. Note that since U_D and L_B are full rank matrices, the Q_F and ΔF_N have the same rank. The structure of Q_F is described in the following lemma.

LEMMA 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The elements of Q_F are zeros except the first s and the last r columns.*

Proof. Note that $F_{1,N}$ and $F_{2,N}$ are Toeplitz matrices with elements

$$(F_{1,N})_{i,j} = t_{N+i-j} \quad \text{and} \quad (F_{2,N})_{i,j} = t_{-N+i-j}.$$

The (i,j) elements of $F_{1,N}U_DL_B$ and $F_{2,N}U_DL_B$ are

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m}d_{n-m}b_{n-j} \quad \text{and} \quad \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m}d_{n-m}b_{n-j},$$

where $b_0 = 1$ $(d_0 = 1)$ and $b_i = 0$ $(d_i = 0)$ if the subscript i is not in the range $0 \le i \le \beta$ $(0 \le i \le \delta)$. If $s < j \le N - r$, we can simplify the above summations as

$$\sum_{m'=0}^{\delta}\left(\sum_{n'=0}^{\beta} t_{N+i+m'-n'-j}b_{n'}\right)d_{m'} = 0 \quad \text{and} \quad \sum_{n'=0}^{\beta}\left(\sum_{m'=0}^{\delta} t_{-N+i+m'-n'-j}d_{m'}\right)b_{n'} = 0,$$

where $m' = n - m$, $n' = n - j$, and the equalities are due to the recursion defined in (4.3). Thus, the elements of

$$Q_F = \Delta F_N U_D L_B = (F_{1,N} + F_{2,N})U_D L_B$$

are zeros except the first s and the last r columns. □

Consequently, we decompose the complex N-tuple space C^N into two orthogonal complement subspaces,

$$\begin{aligned} \mathcal{R}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ s < i \le N - r\}, \\ \mathcal{N}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ 1 \le i \le s \ \text{or}\ N - r < i \le N\}, \end{aligned} \tag{4.8}$$

with dimensions

$$\dim \mathcal{R}(Q_F) = d \quad \text{and} \quad \dim \mathcal{N}(Q_F) = N - d.$$

The subspace $\mathcal{N}(Q_F)$ is contained in the null space of Q_F. Let Q_{NW} denote the northwest $s \times s$ block in Q_F, and Q_{NE}, Q_{SW} and Q_{SE} the corresponding corner blocks in Q_F with sizes $s \times r$, $r \times s$ and $r \times r$, respectively. By using the subspace decomposition (4.8), it is easy to see that the nonzero eigenvalues of Q_F only depend on the corresponding four corner blocks of Q_F, and are also the eigenvalues of the $d \times d$ matrix,

$$P_F = \begin{bmatrix} Q_{NW} & Q_{NE} \\ Q_{SW} & Q_{SE} \end{bmatrix}.$$

In other words, the rank of Q_F is the same as that of P_F.

The bounds for the elements of Q_{NW}, Q_{NE}, Q_{SW} and Q_{SE} are summarized as follows:

$$\begin{aligned} &|(Q_{NW})_{i,j}| \le \tau_{NW}, && \tau_{NW} = O(|t_N| + |t_{-N}|), \\ &|(Q_{SE})_{i,j}| \le \tau_{SE}, && \tau_{SE} = O(|t_N| + |t_{-N}|), \\ &|(Q_{NE})_{i,j}| \le (F_{1,N}U_DL_B)_{i,N-r+j} + \tau_{NE}, && \tau_{NE} = O(|t_{-2N}|), \\ &|(Q_{SW})_{i,j}| \le (F_{2,N}U_DL_B)_{N-s+i,j} + \tau_{SW}, && \tau_{SW} = O(|t_{2N}|). \end{aligned} \tag{4.9}$$

To derive (4.9), recall that the (i,j) element of Q_F is

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m} d_{n-m} b_{n-j} + \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m} d_{n-m} b_{n-j},$$

which is bounded by

$$\sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{N+i+m'-n'-j}||d_{m'}||b_{n'}| + \sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{-N+i+m'-n'-j}||d_{m'}||b_{n'}|.$$

Since the elements of Q_{NW} are the same as those of Q_F with subscript (i,j), $i,j \le s$, they are bounded by

$$\tau_{NW} = \sum_{i=0}^{\beta} |b_i| \sum_{j=0}^{\delta} |d_j| \big(\max_{-(s+\beta)<n<s+\delta} |t_{N+n}| + \max_{-(s+\beta)<n<s+\delta} |t_{-N+n}| \big).$$

To determine the bound for $\sum_{i=0}^{\beta} |b_i|$, we factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1-r_1 z^{-1})(1-r_2 z^{-1})\cdots(1-r_\beta z^{-1}).$$

A direct consequence of (3.1) is that all poles of $A(z^{-1})/B(z^{-1})$ should lie inside the unit circle, i.e. $|r_i| < 1$, $1 \le i \le \beta$, so that

$$|b_k| \le \binom{\beta}{k} (\max |r_i|)^k \le \binom{\beta}{k}, \quad \text{where} \quad \binom{\beta}{k} \equiv \frac{\beta!}{(\beta-k)!k!}.$$

Therefore, we obtain

$$\sum_{k=0}^{\beta} |b_k| \le \sum_{k=0}^{\beta} \binom{\beta}{k} \le 2^\beta.$$

Similarly, $\sum_{k=0}^{\delta} |d_k| \le 2^\delta$ and thus, the elements of Q_{NW} are bounded by

$$\tau_{NW} = 2^{(\beta+\delta)}(|t_{N-s-\beta}| + |t_{-N+s+\delta}|) = O(|t_N| + |t_{-N}|),$$

where the last equality is due to the fact that, for large n, t_n can be approximated by

$$t_n \approx c r_j^n, \quad \text{where} \quad |r_j| = \max_i |r_i|, \qquad (4.10)$$

and where c is a constant. Similarly, we can prove that the elements of Q_{SE} are bounded by

$$\tau_{SE} = 2^{(\beta+\delta)}(|t_{N-r-\beta}| + |t_{-N+r+\delta}|) = O(|t_N| + |t_{-N}|).$$

The (i,j), $1 \le i \le s, 1 \le j \le r$, element of Q_{NE} is the sum of the $(i, N-r+j)$ elements of $F_{1,N}U_D L_B$ and $F_{2,N}U_D L_B$. It is straightforward to verify that the $(i, N-r+j)$ element of $F_{1,N}U_D L_B$ remains unchanged while that of $F_{2,N}U_D L_B$ is bounded by $\tau_{NE} = 2^{(\beta+\delta)}|t_{-2N+d+\delta}| = O(|t_{-2N}|)$ for sufficiently large N. Similarly, we can derive the bound for the elements in Q_{SW} as given by (4.9).

Thus, when N becomes asymptotically large, the P_F converges to

$$\overline{P}_F = \begin{bmatrix} 0 & \overline{Q}_{NE} \\ \overline{Q}_{SW} & 0 \end{bmatrix},$$

where $\overline{Q}_{NE}$ is the converged northeast $s \times r$ block in $F_{1,N}U_D L_B$ and $\overline{Q}_{SW}$ is the converged southwest $r \times s$ block in $F_{2,N}U_D L_B$. Since the ranks of $\overline{Q}_{NE}$ and $\overline{Q}_{SW}$ are both bounded by $\min(r,s)$, the rank of $\overline{P}_F$ is bounded by $\eta = 2\min(r,s)$.

Let us define a matrix $\overline{Q}_F$ by replacing the four corner blocks in Q_F with the corresponding blocks in $\overline{P}_F$. Then, we have

$$\begin{aligned}\tau_Q &= \|Q_F - \overline{Q}_F\|_p = \|P_F - \overline{P}_F\|_p \\ &\leq s\tau_{NW} + r\tau_{SE} + \max(r,s)(\tau_{NE} + \tau_{SW}) \\ &= O(|t_N| + |t_{-N}|),\end{aligned}$$

for $p = 1$ and ∞. The above bounds also hold for $p = 2$ because $\|A\|_2 \leq (\|A\|_1\|A\|_\infty)^{1/2}$ for an arbitrary matrix A. Since τ_Q goes to zero as N goes to infinity due to (3.1), the asymptotic equivalence between Q_F and $\overline{Q}_F$ is established. This result is summarized in the following lemma.

LEMMA 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The Q_F and $\overline{Q}_F$ are asymptotically equivalent.*
Based on Lemmas 2-4, (4.2) and (4.7), $T_N^{-1}\Delta T_N$ is asymptotically equivalent to $T_N^{-1}\overline{Q}_F L_B^{-1} U_D^{-1}$ whose rank is bounded by $\eta = 2\min(r,s)$ and $K_N^{-1}T_N$ has at most η asymptotic eigenvalues not converging to one (outliers).

4.2. The clustering radius of $K_N^{-1}T_N$. We use perturbation theory to estimate the clustering radius of the $N - \eta$ clustered eigenvalues. Instead of examining the eigenvalues of $T_N^{-1}\Delta T_N$ directly, we study those of the similar matrix

$$G_N = L_B^{-1}U_D^{-1}T_N^{-1}\Delta T_N U_D L_B = L_B^{-1}U_D^{-1}T_N^{-1}Q_T,$$

where $Q_T = \Delta T_N U_D L_B$. Let us define

$$H_N = L_B^{-1}U_D^{-1}T_N^{-1}\overline{Q}_F.$$

It is clear that H_N has only d nonzero columns as $\overline{Q}_F$ (or Q_F). The G_N can be viewed as a matrix obtained from H_N by adding the perturbation matrix

$$\Delta G_N = G_N - H_N = L_B^{-1}U_D^{-1}T_N^{-1}(Q_T - \overline{Q}_F). \tag{4.11}$$

A bound of $\|\Delta G_N\|_2$ is given below so that we can estimate the clustering radius of the clustered eigenvalues by using perturbation theory for eigenvalues.

LEMMA 5. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). Then, for sufficiently large N, the $\|\Delta G_N\|_2$ is bounded by $\epsilon = O(|t_N| + |t_{-N}|)$.*

Proof. We first study the 2-norm of $Q_T - \overline{Q}_F$, which is bounded by

$$\|Q_T - \overline{Q}_F\|_2 \leq \|Q_T - Q_F\|_2 + \|Q_F - \overline{Q}_F\|_2.$$

As shown in the proof of Lemma 4, the second term $\|Q_F - \overline{Q}_F\|_2$ is bounded by $\tau_Q = O(|t_N| + |t_{-N}|)$ while the first term $\|Q_T - Q_F\|_2$ is bounded by

$$\|Q_T - Q_F\|_2 \leq \|\Delta T_N - \Delta F_N\|_2\|U_D\|_2\|L_B\|_2 = \|\Delta E_N\|_2\|U_D\|_2\|L_B\|_2.$$

Recall from (4.6) that $\|\Delta E_N\|_2 \leq \sum_{n=N}^{2N-1}(|t_n| + |t_{-n}|)$. By using (4.10), we have

$$\sum_{n=N}^{2N-1}|t_n| \leq \sum_{n=N}^{\infty}|q_j r_j^n| = \frac{|t_N|}{1-|r_j|} = M_B|t_N|, \quad \text{where} \quad M_B = \frac{1}{1-|r_j|}.$$

Similarly, $\sum_{n=N}^{2N-1}|t_{-n}| \leq M_D|t_{-N}|$. Besides, $\|L_B\|_2 \leq \sum_{k=0}^{\beta}|b_k| \leq 2^\beta$ and $\|U_D\|_2 \leq \sum_{k=0}^{\delta}|d_k| \leq 2^\delta$. Thus, we obtain a bound for the first term, i.e.

$$\|Q_T - Q_F\|_2 \leq 2^{(\beta+\delta)}(M_B|t_N| + M_D|t_{-N}|) = O(|t_N| + |t_{-N}|),$$

and conclude that

$$\|Q_T - \overline{Q}_F\|_2 \leq O(|t_N| + |t_{-N}|).$$

With (4.11), we have

$$||\Delta G_N||_2 \le ||L_B^{-1}||_2||U_D^{-1}||_2||T_N^{-1}||_2||(Q_T - \overline{Q}_F)||_2.$$

Due to (3.2), $||T_N^{-1}||_2$ is bounded by a constant c_T independent of N. To show that $||L_B^{-1}||_2$ and $||U_D^{-1}||_2$ are also bounded, we factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1 - r_1 z^{-1})(1 - r_2 z^{-1}) \cdots (1 - r_\beta z^{-1}),$$

where we assume that all roots r_i are distinct for simplicity. By applying the isomorphism between the ring of the power series and the ring of semi-infinite lower (or upper) triangular Toeplitz matrices, the L_B and L_B^{-1} can be decomposed into the products,

$$L_B = L_{r_1} L_{r_2} \cdots L_{r_\beta}, \qquad L_B^{-1} = L_{r_\beta}^{-1} \cdots L_{r_2}^{-1} L_{r_1}^{-1},$$

where $L_{r_i}, 1 \le i \le \beta$ is an $N \times N$ lower triangular Toeplitz matrix with $[1, -r_i, 0, \cdots, 0]^T$ as the first column. It can be easily verified that $L_{r_i}^{-1}$ is a lower triangular Toeplitz matrix with $[1, r_i, r_i^2, \cdots, r_i^{N-1}]^T$ as the first column. Therefore,

$$||L_{r_i}^{-1}||_p \le \sum_{k=0}^{N-1} |r_i^k| \le \sum_{k=0}^{\infty} |r_i^k| = \frac{1}{1 - |r_i|}, \qquad p = 1, 2, \infty,$$

and

$$||L_B^{-1}||_2 \le \prod_{i=1}^{\beta} ||L_{r_i}^{-1}||_2 \le \prod_{i=1}^{\beta} \frac{1}{1 - |r_i|} = c_B.$$

Similar arguments can be used to prove that $||U_D^{-1}||_2 \le c_D$. Finally, we have

$$||\Delta G_N||_2 \le \epsilon \equiv c_B c_D c_T ||(Q_T - \overline{Q}_F)||_2 = O(|t_N| + |t_{-N}|). \tag{4.12}$$

The proof is completed. □

Let us denote the rank of $H_N = L_B^{-1} U_D^{-1} T_N^{-1} \overline{Q}_F$ by $\tilde{\eta}$. Clearly, $\tilde{\eta} \le \eta = 2\min(r, s)$. We arrange the eigenvalues of H_N in a descending order so that $|\lambda_n| \ge |\lambda_{n+1}|$ ($\lambda_n = 0$ for $\tilde{\eta} < n \le N$), and denote the corresponding normalized right-hand and left-hand eigenvectors by $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ and $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$, respectively. Besides, vectors $\mathbf{x}_n$ with $\tilde{\eta} < n \le N$ are chosen to be othorgonal. The complex N-tuple space is decomposed into the row and the null spaces of H_N,

$$\text{Row}(H_N) = \text{span}\{\mathbf{x}_n, n \le \tilde{\eta}\}, \qquad \text{Null}(H_N) = \text{span}\{\mathbf{x}_n, \tilde{\eta} < n \le N\}.$$

Since $G_N = H_N + \Delta G_N$ and $||\Delta G_N||_2 \le \epsilon$, the eigenvalues and the right-hand eigenvectors of G_N are denoted by $\lambda_n(\epsilon)$ and $\mathbf{x}_n(\epsilon)$, respectively. By using results from perturbation theory for repeated eigenvalues [16], the eigenvectors $\mathbf{x}_n(\epsilon)$ with $\tilde{\eta} < n \le N$ must take the form

$$\mathbf{x}_n(\epsilon) = \sum_{m=1}^{\tilde{\eta}} \frac{\xi_{mn}}{(\lambda_n - \lambda_m)s_m} \mathbf{x}_m + \sum_{m=\tilde{\eta}+1}^{N} g_{mn} \mathbf{x}_m + O(\epsilon^2), \tag{4.13}$$

where $\xi_{mn} = \mathbf{y}_m^H \Delta G_N \mathbf{x}_n$, $\lambda_n = 0$, $s_m = \mathbf{y}_m^H \mathbf{x}_m$ and $g_{nn} = 1$. Due to the construction, we know that

$$||\mathbf{x}_n(\epsilon)||_2 \ge ||\mathbf{x}_n||_2 = 1. \tag{4.14}$$

The factor $|\xi_{mn}|$ is bounded by

$$|\xi_{mn}| = |\mathbf{y}_m^H \Delta G_N \mathbf{x}_n| \le ||\mathbf{y}_m||_2 ||\Delta G_N||_2 ||\mathbf{x}_n||_2 \le \epsilon.$$

The $|s_m^{-1}|, 1 \le m \le \tilde{\eta}$, is also bounded as given in the following lemma.

LEMMA 6. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). Then, the $|s_m^{-1}|, 1 \le m \le \bar{\eta}$, of H_N is bounded by a constant independent of N.*

Proof. The eigenvalues λ and the right-hand eigenvectors $\mathbf{x}$ of H_N satisfy

$$L_B\overline{Q}_F\mathbf{x} = \lambda L_B T_N U_D L_B \mathbf{x}.$$

Since the elements of $\overline{Q}_F$ are zeros except the first s and the last r columns, so are the elements of $L_B\overline{Q}_F$. Thus, the nonzero eigenvalues of H_N only depend on the northwest $s \times s$, northeast $s \times r$, southwest $r \times s$ and southeast $r \times r$ blocks of $L_B\overline{Q}_F$ and $L_B T_N U_D L_B$. The boundness of $|s_m^{-1}|, 1 \le m \le \bar{\eta}$, is guaranteed if the elements of the four corner blocks of $L_B\overline{Q}_F$ and $L_B T_N U_D L_B$ remain unchanged for sufficiently large N.

By using the band structure of L_B and the special structure of $\overline{Q}_F$, it is straightforward to verify that the four blocks of $L_B\overline{Q}_F$ remain unchanged for large N. Next, we examine the matrix $L_B T_N U_D L_B$. By using (4.1) and the isomorphism between the ring of the power series and the ring of the semi-infinite lower (or upper) triangular Toeplitz matrices, we can express T_N as

$$T_N = L_A L_B^{-1} + U_C U_D^{-1},$$

where L_A is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $A(z^{-1})$ as the first column, and U_C is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $C(z)$ as the first row. Then, we have

$$L_B T_N U_D L_B = L_A U_D L_B + L_B U_C L_B,$$

whose four corner blocks remain unchanged for large N. Thus, λ_m and $s_m = \mathbf{y}_m^H \mathbf{x}_m$ with $1 \le m \le \bar{\eta}$, do not change with N, when N becomes sufficiently large. □

Let $\mathbf{v}_n(\epsilon)$ be the normalized vector of $\mathbf{x}_n(\epsilon)$,

$$\mathbf{v}_n(\epsilon) = \frac{\mathbf{x}_n(\epsilon)}{||\mathbf{x}_n(\epsilon)||_2},$$

which can be decomposed as

$$\mathbf{v}_n(\epsilon) = \mathbf{v}_{\mathrm{N}}(\epsilon) + \mathbf{v}_{\mathrm{R}}(\epsilon),$$

where $\mathbf{v}_{\mathrm{N}}(\epsilon) \in \mathrm{Null}(H_N)$ and $\mathbf{v}_{\mathrm{R}}(\epsilon) \in \mathrm{Row}(H_N)$. The magnitude of $\lambda_n(\epsilon)$, $\bar{\eta} < n \le N$, of G_N is approximated by

$$|\lambda_n(\epsilon)| = ||G_N \mathbf{v}_n(\epsilon)||_2 = ||H_N \mathbf{v}_{\mathrm{R}}(\epsilon) + \Delta G_N \mathbf{v}_n(\epsilon)||_2.$$

By using (4.12)-(4.14), we obtain that

$$\begin{aligned}
\max_{\bar{\eta}<n\le N} |\lambda_n(\epsilon)| &\le \max_{\bar{\eta}<n\le N} ||H_N \mathbf{v}_{\mathrm{R}}(\epsilon)||_2 + \max_{\bar{\eta}<n\le N} ||\Delta G_N \mathbf{v}_n(\epsilon)||_2 \\
&\le \sum_{m=1}^{\bar{\eta}} \frac{||\xi_{mn} H_N \mathbf{x}_m||_2}{|\lambda_m s_m|\ ||\mathbf{x}_n(\epsilon)||_2} + ||\Delta G_N||_2 \\
&\le \sum_{m=1}^{\bar{\eta}} \frac{\epsilon}{|s_m|} + \epsilon = \epsilon_K \\
&= O(|t_N| + |t_{-N}|),
\end{aligned}$$

for sufficiently large N. The above analysis is concluded in the following theorem.

THEOREM 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the preconditioned Toeplitz matrix $K_N^{-1}T_N$ has the following two properties:*

P1: The number of outliers is at most $\eta = 2\min(r, s)$.

P2: There are at least $N - \eta$ eigenvalues confined in the disk centered at 1 with radius ϵ_K, where

$$\epsilon_K = O(|t_N| + |t_{-N}|).$$

5. Spectral properties of preconditioned rational Toeplitz $S_N^{-1}T_N$. The preconditioned Toeplitz matrix $S_N^{-1}T_N$ has similar spectral properties as $K_N^{-1}T_N$. The number of outliers of $S_N^{-1}T_N$ can be obtained by proving that $\Delta S_N = S_N - T_N$ and ΔF_N given by (4.4) are asymptotically equivalent.

LEMMA 7. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). $S_N^{-1}T_N$ has asymptotically at most $\eta = 2\min(r,s)$ eigenvalues not converging to 1.*

Proof. Let us define $\Delta S_N = S_N - T_N$, and express the difference between ΔF_N in (4.4) and ΔS_N as

$$\Delta F_N - \Delta S_N = E_{1,N} + E_{2,N},$$

where $E_{1,N}$ and $E_{2,N}$ are $N \times N$ Toeplitz matrices with elements

$$(E_{1,N})_{i,j} = \begin{cases} t_{N+i-j}, & -(M-1) \le i-j \le N-1, \\ t_{i-j}, & -(N-1) \le i-j \le -M, \end{cases}$$

and

$$(E_{2,N})_{i,j} = \begin{cases} t_{i-j}, & N-(M-1) \le i-j \le N-1, \\ t_{i-j-N}, & -(N-1) \le i-j \le N-M, \end{cases}$$

respectively. By using similar arguments in deriving Lemma 2, we can prove that ΔS_N and ΔF_N are asymptotically equivalent. Since ΔF_N is asymptotically equivalent to the matrix $\overline{Q}_F L_B^{-1} U_D^{-1}$ with rank $\tilde{\eta} \le \eta = 2\min(r,s)$ as described in Lemma 4, the proof is completed. □

Similar arguments used in §4.2 can be applied to derive the following theorem.

THEOREM 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the preconditioned Toeplitz matrix $S_N^{-1}T_N$ has the following two properties:*
P1: The number of outliers is at most $\eta = 2\min(r,s)$.
P2: There are at least $N - \eta$ eigenvalues confined in the disk centered at 1 with radius ϵ_S, where

$$\epsilon_S = O(|t_{N-M}| + |t_{1-M}|).$$

6. Numerical results. Four test problems, including both rational and nonrational T_N, are used to illustrate the above analysis. For the nonsymmetric Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ to be solved, we choose $\mathbf{b} = (1,\cdots,1)^T$ and zero initial guess in all experiments. Without further specification, M is chosen such that $|t_{N-M}| \approx |t_{1-M}|$ to construct preconditioner S_N. We use the first test problem, which is generated by a nonrational function, to examine the clustering effect of singular values. Test problems 2-4 are generated by rational functions so that the number of outliers and the clustering radius can be observed, which confirm the theoretical results developed in §4 and §5.

Test Problem 1. Nonrational T_N.

Let T_N be a Toeplitz matrix with generating sequence

$$t_n = \begin{cases} 1/\log(2-n), & n \le -1, \\ 1/\log(2-n) + 1/(1+n), & n = 0, \\ 1/(1+n), & n \ge 1. \end{cases}$$

The singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are plotted in Fig. 1(a) for $N = 32$, 64 and 128. Both $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have clustered singular values. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 1(b). It is clear that the eigenvalues possess a certain clustering property. We apply both the CGN and CGS methods to solve the preconditioned Toeplitz system $P_N^{-1}T_N\mathbf{x} = P_N^{-1}\mathbf{b}$. The numbers of iterations required for the CGN and CGS methods to achieve $\|\mathbf{b} - T_N\mathbf{x}\|_2 < 10^{-12}$ are summarized in Tables 1 and 2, respectively. The case without preconditioning is also included for comparison. The use of preconditioners does accelerate the convergence rate of iterative methods. The numbers of

N	T_N	S_N	K_N
32	24	12	9
64	33	15	11
128	49	17	13

TABLE 1
The numbers of iterations required for the CGN method.

N	T_N	S_N	K_N
32	15	7	9
64	21	8	10
128	26	9	10

TABLE 2
The numbers of iterations required for the CGS method.

iterations required for S_N and K_N increase slightly as N becomes large. The K_N performs better than S_N in the CGN method. However, their performances are comparable for the CGS method. Since the CGN method in general requires more iterations than the CGS method and the convergence rate of the CGS method is related to the eigenvalue distribution of the iteration matrix, we will only present the results of the CGS method for the remaining three test problems.

Test Problem 2. Rational T_N with $(r, s) = (1, 1)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{1 + 0.7z^{-1}}{1 - 0.9z^{-1}} + \frac{1 - 0.8z}{1 + 0.7z}.$$

To show that the simple rule for choosing M, i.e. $|t_{N-M}| \approx |t_{1-M}|$, does provide a better spectral clustering property and a better convergence rate for $S_N^{-1}T_N$, two preconditioners S_N and $\tilde{S}_N$ are constructed. The S_N is constructed with M such that $|t_{N-M}| \approx |t_{1-M}|$ while the $\tilde{S}_N$ is constructed with $M = \lceil N/2 \rceil$. The eigenvalues of T_N, $\tilde{S}_N^{-1}T_N$, $S_N^{-1}T_N$ and $K_N^{-1}T_N$ with $N = 32$ are plotted in Figs. 2(a)-(d). All preconditioned Toeplitz matrices have eigenvalues clustered around 1 except $2 = 2\min(r, s)$ outliers. The $K_N^{-1}T_N$ has the best clustering effect, and the eigenvalues of $S_N^{-1}T_N$ are more closely clustered than those of $\tilde{S}_N^{-1}T_N$. The sums of magnitudes of the last elements in constructing S_N and K_N and the corresponding clustering radii are listed in Table 3. They are approximately of the same order, as stated in Theorems 3 and 4.

The convergence history of the CGS method with various preconditioners is plotted in Fig. 3 with $N = 32$. The convergence rate of the CGS method without preconditioning (the curve denoted by T_N) is very slow. This phenomenon is not surprising by examining the eigenvalue distribution given in Fig. 2(a). Preconditioning improves the convergence behavior dramatically. It is clear that K_N performs the best while S_N performs better than $\tilde{S}_N$.

Test Problem 3. Rational T_N with $(r, s) = (3, 1)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{(1 + 0.5z^{-1})(1 + 0.7z^{-1})}{(1 - 0.4z^{-1})(1 - 0.6z^{-1})(1 - 0.8z^{-1})} + \frac{1 + 0.8z}{1 + 0.9z}.$$

The eigenvalues of T_N, $S_N^{-1}T_N$ and $K_N^{-1}T_N$ with $N = 64$ are plotted in Figs. 4(a)-(c). It is clear that $K_N^{-1}T_N$ has $2 = 2\min(r, s)$ outliers. The outliers of $S_N^{-1}T_N$ are not easy to identify for this case.

N	ϵ_S	$\lvert t_{N-M}\rvert + \lvert t_{1-M}\rvert$	ϵ_K	$\lvert t_{N-1}\rvert + \lvert t_{1-N}\rvert$
32	8.2×10^{-2}	2.8×10^{-1}	3.5×10^{-2}	6.8×10^{-2}
64	4.6×10^{-3}	2.1×10^{-2}	1.2×10^{-3}	2.3×10^{-3}
128	3.3×10^{-5}	1.1×10^{-4}	1.4×10^{-6}	2.7×10^{-6}

TABLE 3
The clustering radii ϵ of preconditioners S_N and K_N for Test Problem 2.

N	ϵ_S	$\lvert t_{N-M}\rvert + \lvert t_{1-M}\rvert$	ϵ_K	$\lvert t_{N-1}\rvert + \lvert t_{1-N}\rvert$
32	1.7×10^{-1}	1.4×10^{-1}	6.1×10^{-2}	2.8×10^{-2}
64	2.7×10^{-2}	1.3×10^{-2}	5.1×10^{-4}	1.6×10^{-4}
128	1.7×10^{-3}	1.6×10^{-4}	5.8×10^{-7}	1.7×10^{-7}

TABLE 4
The clustering radii ϵ of preconditioners S_N and K_N for Test Problem 3.

However, two outliers can be observed more easily for larger N. Besides, the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$. We list in Table 4 the sums of magnitudes of the last elements in constructing S_N and K_N and the corresponding clustering radii. The convergence history of the CGS method with $N = 64$ is plotted in Fig. 5. We observe that the CGS method without preconditioning does not converge and that the CGS method with preconditioners K_N and S_N converges in 4 and 6 iterations, respectively. This seems to suggest that the use of preconditioners does not only accelerate the convergence rate by providing better spectral properties but also improves the convergence of nonsymmetric iterative algorithms by making the preconditioned matrix more close to normal.

Test Problem 4. Rational triangular T_N with $(r, s) = (1, 0)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{1 - 0.7z^{-1}}{1 + 0.5z^{-1}}.$$

Since there are only N nonzero elements in T_N, we can make S_N the same as K_N. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 6(a). We see that all eigenvalues are clustered around 1 with radius $\epsilon_K = O(|t_N|) = 10^{-9}$. This is consistent with Theorem 3, which predicts that $K_N^{-1}T_N$ has $0 = 2\min(r, s)$ outliers. The convergence history of the CGS method with $N = 32$ is plotted in Fig. 6(b). The CGS method with preconditioner K_N converges in two iterations while the CGS method without preconditioning does not converge.

7. Conclusion. In this paper, we generalized the circulant preconditioning technique from symmetric to nonsymmetric Toeplitz matrices. The resulting preconditioned Toeplitz systems are then solved by various iterative methods such as CGN and CGS. For a large class of Toeplitz matrices, we proved that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except a fixed number independent of N. When the generating function is rational, the eigenvalues of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are classified into clustered eigenvalues and outliers. The number of outliers depends on the order of the rational generating function. The clustered eigenvalues are confined in the disk centered at 1 with the radii $\epsilon_K = O(|t_N| + |t_{-N}|)$ and $\epsilon_S = O(|t_{N-M}| + |t_{1-M}|)$ for $K_N^{-1}T_N$ and $S_N^{-1}T_N$, respectively. Since the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$, preconditioner K_N performs better than S_N for solving rational Toeplitz systems.

REFERENCES

[1] R. H. CHAN, *Circulant preconditioners for Hermitian Toeplitz system*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.

[2] R. H. CHAN AND G. STRANG, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 104–119.

[3] T. F. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 766–771.

[4] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.

[5] T. HUCKLE, *Circulant and skew-circulant matrices for solving Toeplitz matrices problems*, in Cooper Mountain Conference on Iterative Methods, Cooper Mountain, Colorado, 1990.

[6] T. K. KU AND C. J. KUO, *Design and analysis of Toeplitz preconditioners*, Tech. Rep. 155, USC, Signal and Image Processing Institute, May 1990. To appear in IEEE Trans. on Signal Processing, Jan. 1992.

[7] ——, *On the spectrum of a family of preconditioned block Toeplitz matrices*, Tech. Rep. 164, USC, Signal and Image Processing Institute, Nov. 1990.

[8] ——, *Spectral properties of preconditioned rational Toeplitz matrices*, Tech. Rep. 163, USC, Signal and Image Processing Institute, Sept. 1990. to appear in SIAM J. Matrix Anal. Appl.

[9] ——, *A minimum-phase LU factorization preconditioner for Toeplitz matrices*, Tech. Rep. 171, USC, Signal and Image Processing Institute, Feb. 1991.

[10] N. M. NACHTIGAL, S. C. REDDY, AND L. N. TREFETHEN, *How fast are nonsymmetric matrix iterations*, in Cooper Mountain Conference on Iterative Methods, Cooper Mountain, Colorado, 1990.

[11] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[12] P. SONNEVELD, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.

[13] G. STRANG, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[14] ——, *Linear Algebra and Its Applications*, Harcourt Brace Jonanovich, Inc., Orlando, Florida, third ed., 1988.

[15] L. N. TREFETHEN, *Approximantion theory and numerical linear algebra*, in Algorithms for Approximation II, M. Cox and J. C. Mason, eds., 1988.

[16] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

[17] A. YAMASAKI, *New preconditioners based on low-rank elimination*, Tech. Rep. Numerical Analysis 89-10, MIT, Dept. of Math., Dec. 1989.

Figure Captions

Figure 1: (a) The singular value distribution of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, and (b) the eigenvalue distribution of $K_N^{-1}T_N$ for Test Problem 1.

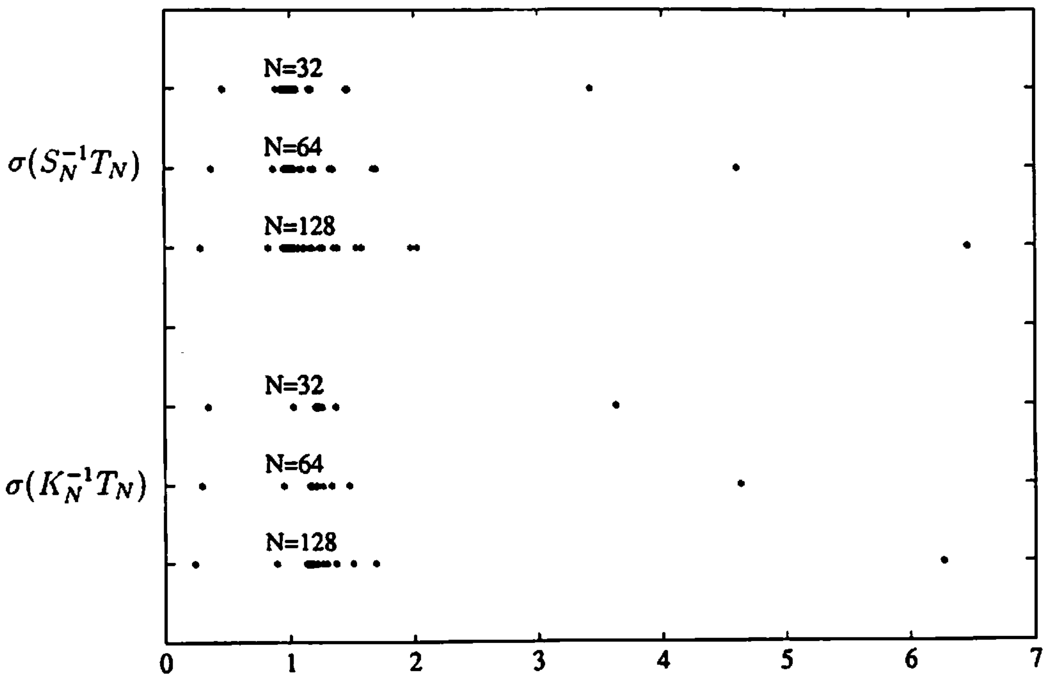
Figure 2: The eigenvalue distribution of (a) T_N, (b) $\tilde{S}_N^{-1}T_N$, (c) $S_N^{-1}T_N$ and (d) $K_N^{-1}T_N$ for Test Problem 2.

Figure 3: The convergence history of the CGS method for Test Problem 2.

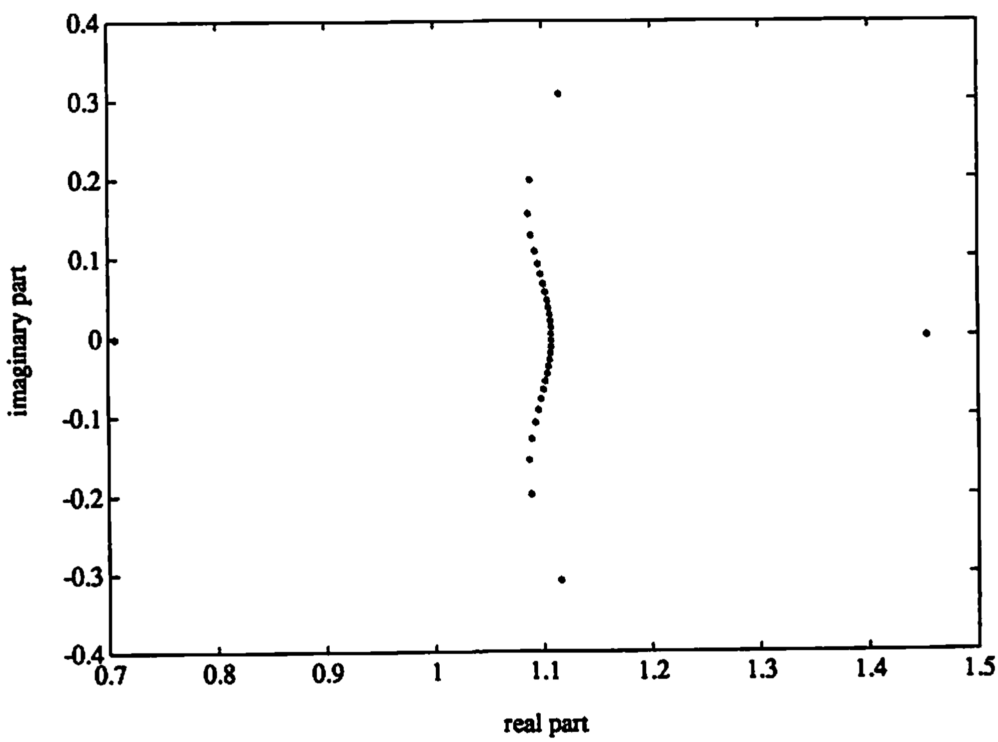
Figure 4: The eigenvalue distribution of (a) T_N, (b) $S_N^{-1}T_N$ and (c) $K_N^{-1}T_N$ for Test Problem 3.

Figure 5: The convergence history of the CGS method for Test Problem 3.

Figure 6: (a) The eigenvalue distribution of $K_N^{-1}T_N$, and (b) the convergence history of the CGS method for Test Problem 4.

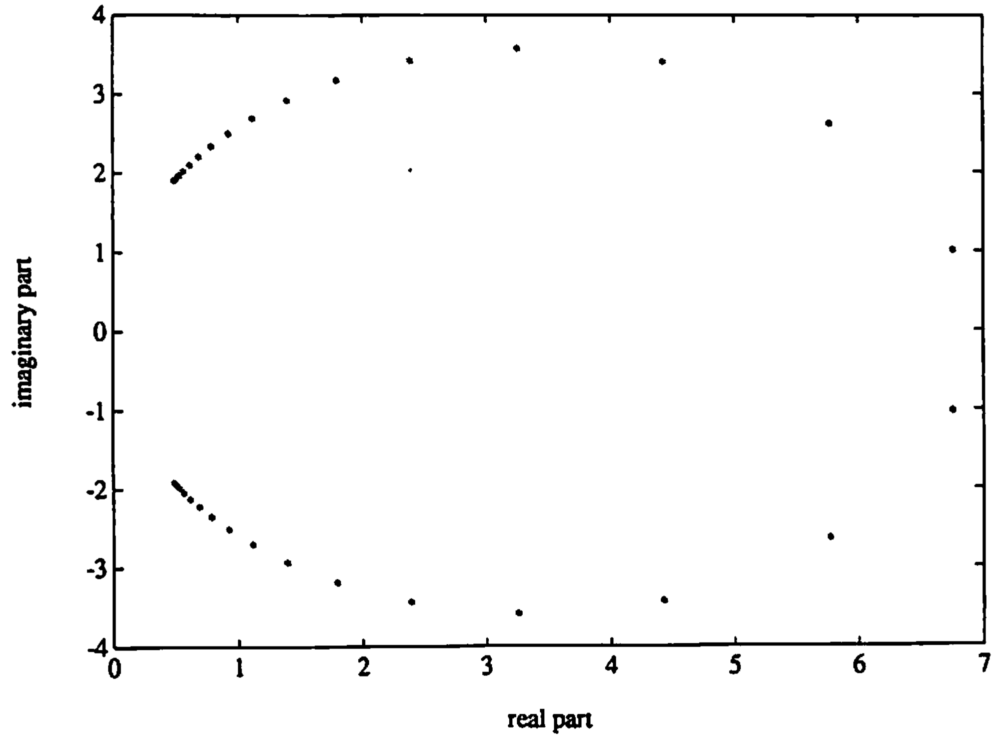


(a)

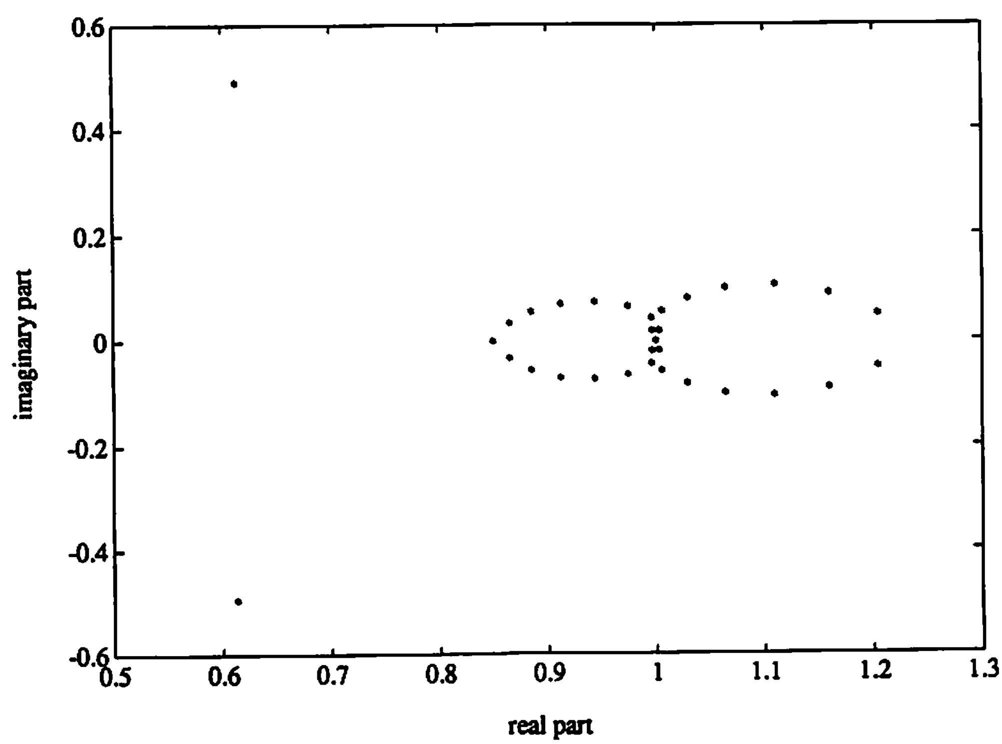


(b)

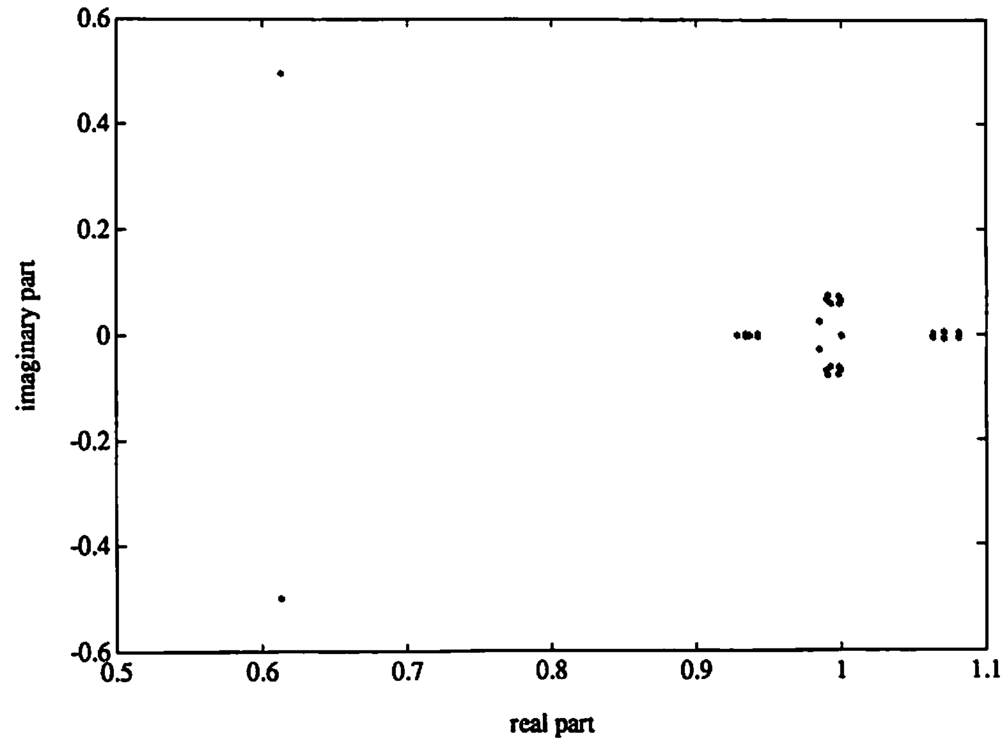
FIG. 1. *(a) The singular value distribution of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, and (b) the eigenvalue distribution of $K_N^{-1}T_N$ for Test Problem 1.*


imaginary part
real part
4
3
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5
6
7

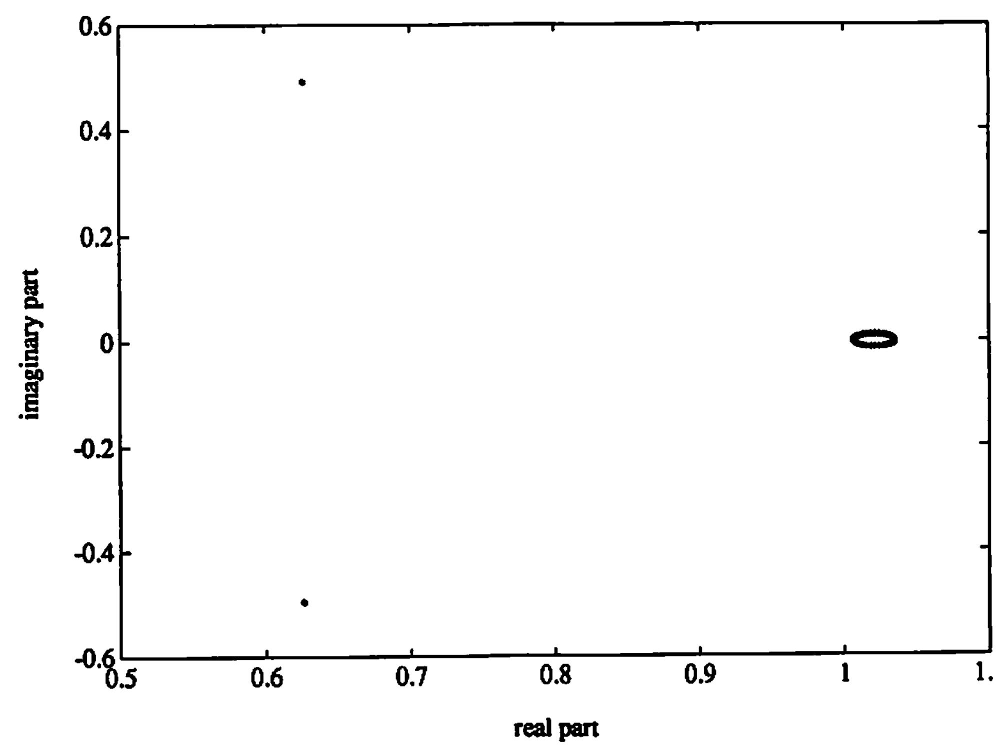
(a)


imaginary part
real part
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3

(b)



(c)



(d)

FIG. 2. *The eigenvalue distribution of (a) T_N, (b) $\tilde{S}_N^{-1}T_N$, (c) $S_N^{-1}T_N$ and (d) $K_N^{-1}T_N$ for Test Problem 2.*

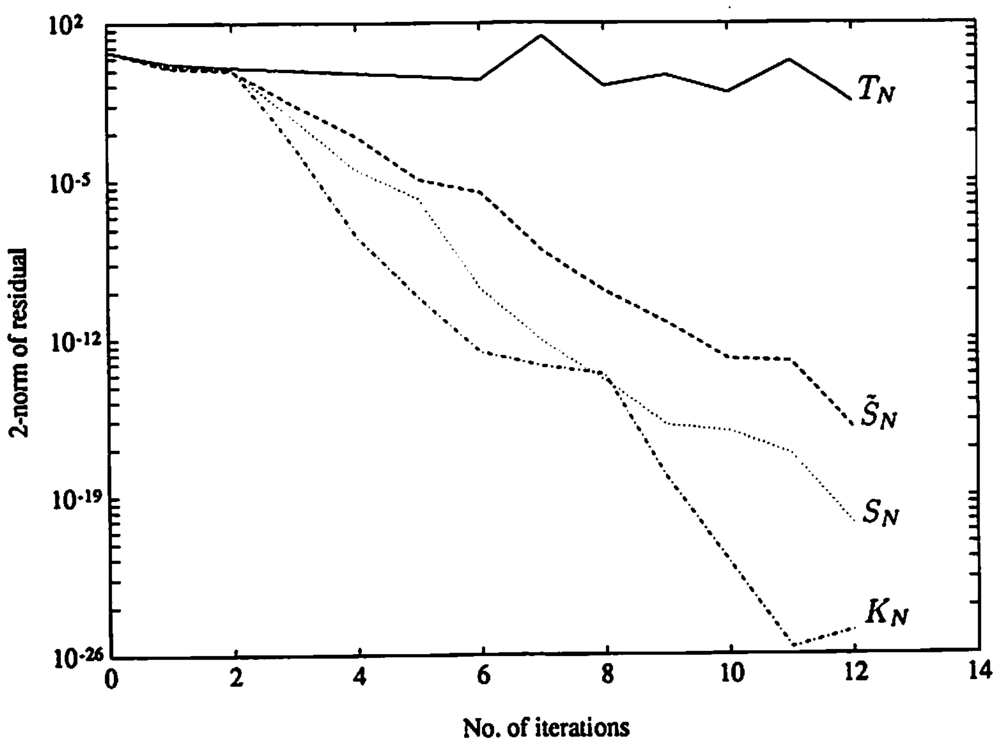
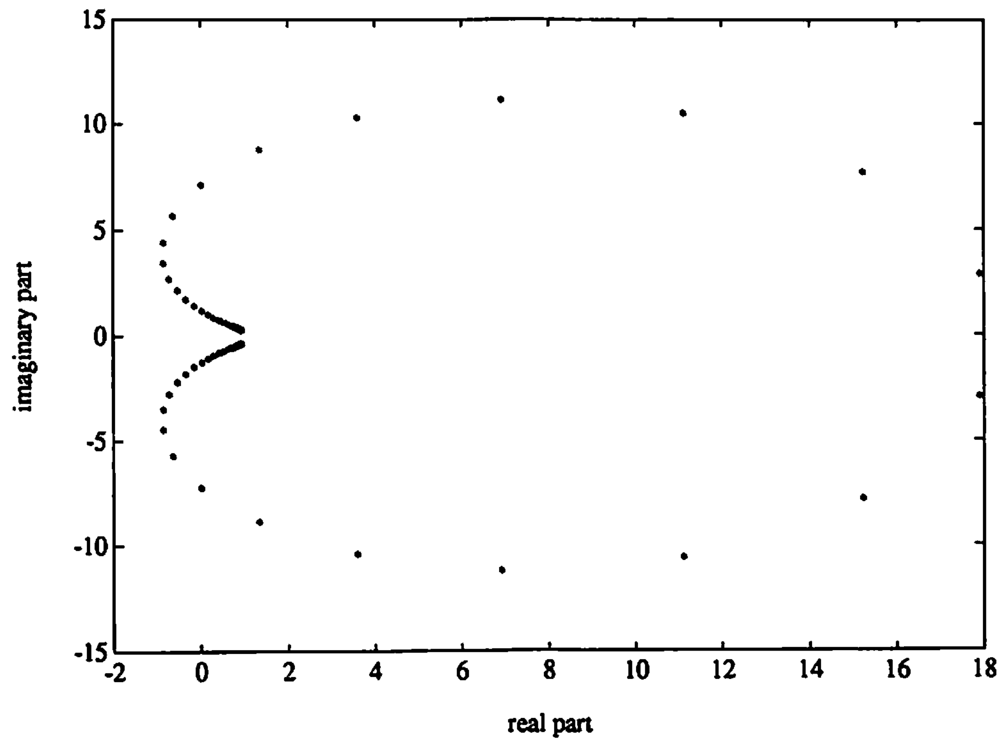
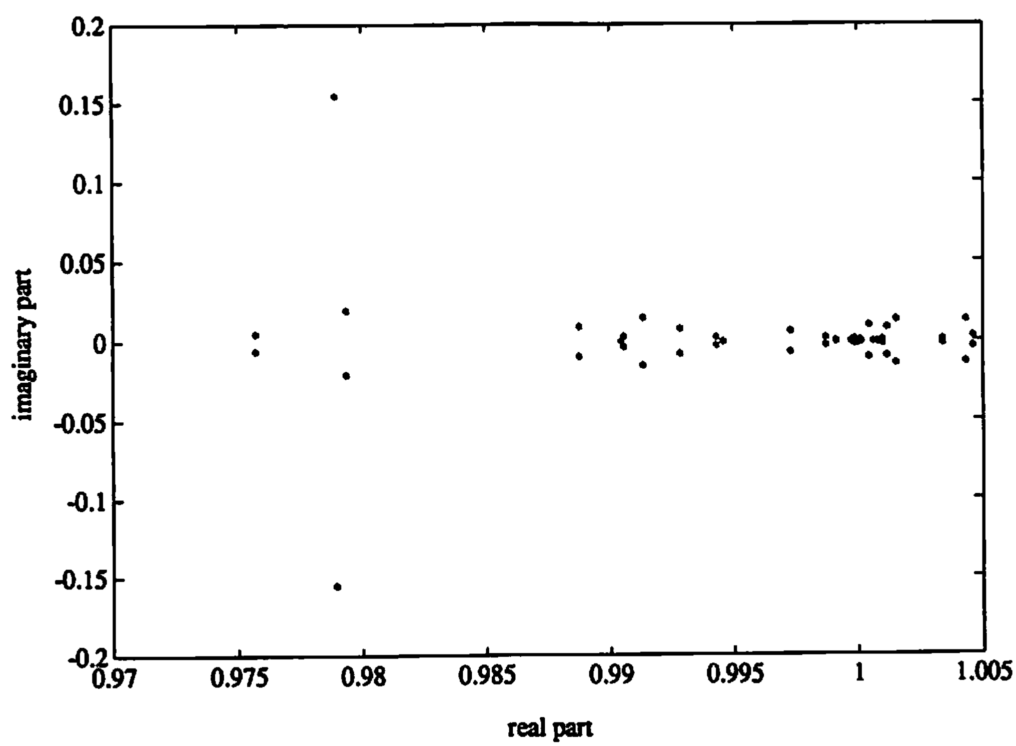


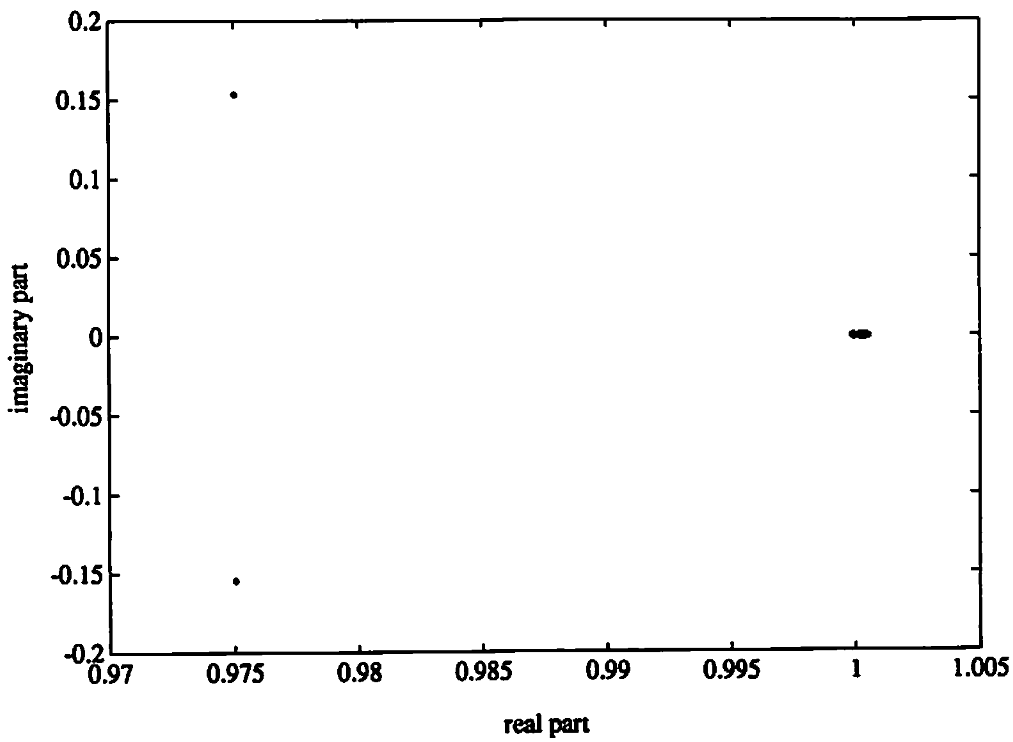
FIG. 3. *The convergence history of the CGS method for Test Problem 2.*


15
10
5
0
-5
-10
-15
-2
0
2
4
6
8
10
12
14
16
18
imaginary part
real part

(a)


0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0.97
0.975
0.98
0.985
0.99
0.995
1
1.005
imaginary part
real part

(b)



(c)

FIG. 4. *The eigenvalue distribution of (a) T_N, (b) $S_N^{-1}T_N$ and (c) $K_N^{-1}T_N$ for Test Problem 3.*

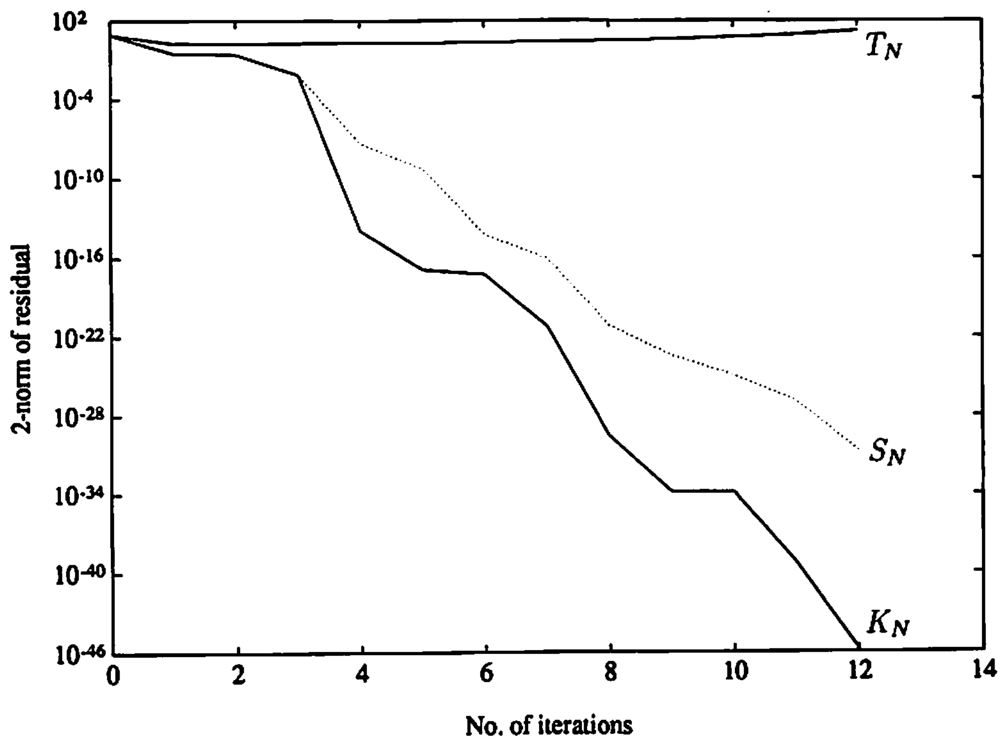


FIG. 5. *The convergence history of the CGS method for Test Problem 3.*

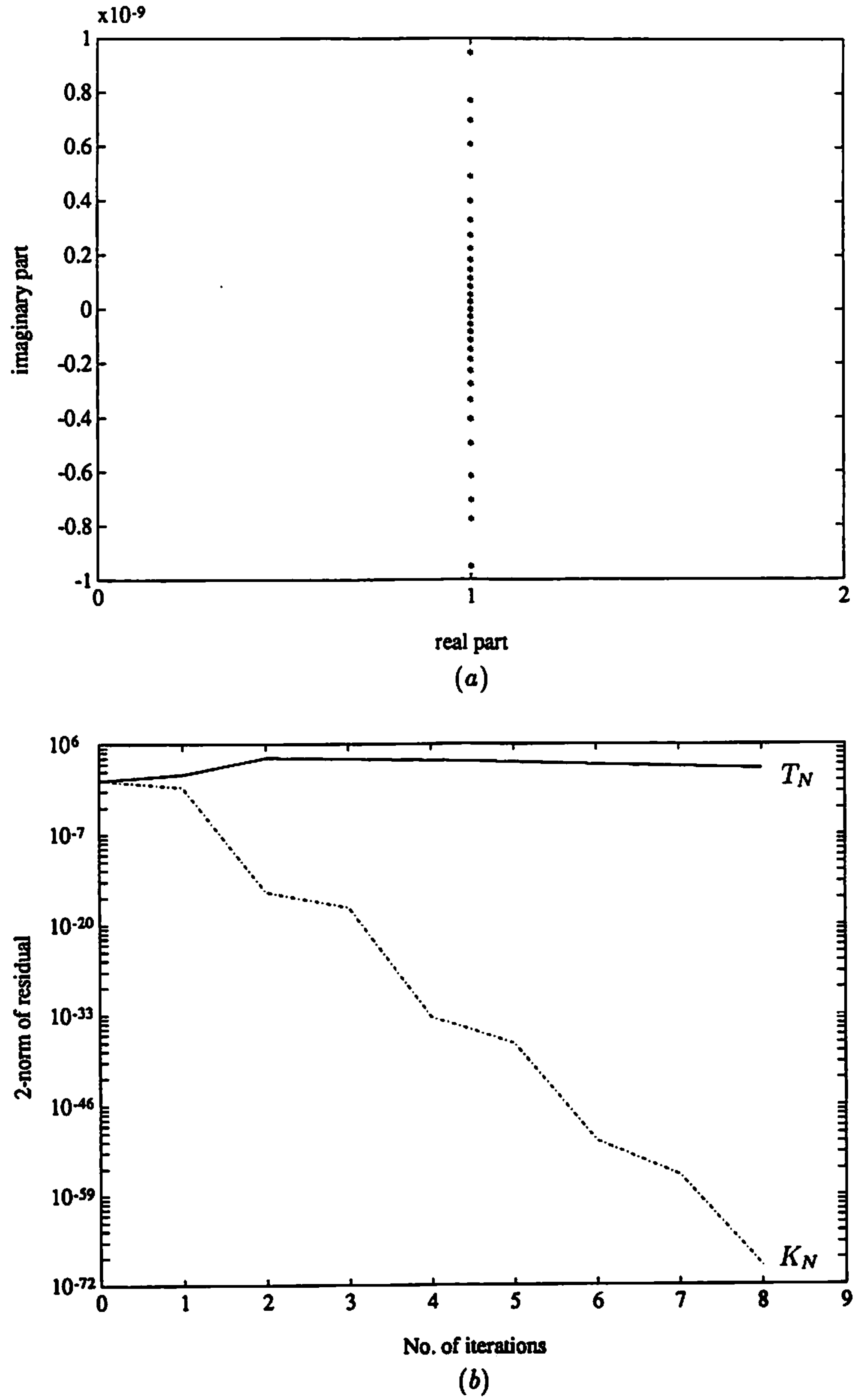


FIG. 6. *(a) The eigenvalue distribution of $K_N^{-1}T_N$, and (b) the convergence history of the CGS method for Test Problem 4.*

USC-SIPI REPORT #175

Spectral Properties of Preconditioned Rational Toeplitz Matrices: The Nonsymmetric Case

by

Ta-Kang Ku and C.-C. Jay Kuo

April 1991

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
Powell Hall of Engineering
University Park/MC-0272
Los Angeles, CA 90089 U.S.A.

Submitted to SIAM Journal on Matrix Analysis and Applications.

SPECTRAL PROPERTIES OF PRECONDITIONED RATIONAL TOEPLITZ MATRICES : THE NONSYMMETRIC CASE *

TA-KANG KU† AND C.-C. JAY KUO†

Abstract. Various preconditioners for symmetric positive-definite (SPD) Toeplitz matrices in circulant matrix form have recently been proposed. The spectral properties of the preconditioned SPD Toeplitz matrices have also been studied. In this research, we apply Strang's preconditioner S_N and our preconditioner K_N to an $N \times N$ nonsymmetric (or nonhermitian) Toeplitz system $T_N\mathbf{x} = \mathbf{b}$. For a large class of Toeplitz matrices, we prove that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except a fixed number independent of N. If T_N is additionally generated by a rational function, we are able to characterize the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ directly. Let the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ be classified into the outliers and the clustered eigenvalues depending on whether they converge to 1 asymptotically. Then, the number of outliers depends on the order of the rational generating function, and the clustering radius is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner. Numerical experiments are provided to illustrate our theoretical study.

Key words. Toeplitz, circulant, nonsymmetric, preconditioners, preconditioned iterative method, CGN, CGS, GMRES.

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction. Research on preconditioning symmetric positive-definite (SPD) Toeplitz matrices with circulant matrices has been active recently [1], [3], [5], [6], [13]. In this research, we generalize Strang's preconditioner S_N [13] and our preconditioner K_N [6] to nonsymmetric (or nonhermitian) Toeplitz matrices. Let T_N be an $N \times N$ nonsymmetric Toeplitz matrix with elements $t_{i,j} = t_{i-j}$. The generalized Strang's preconditioner S_N is obtained by preserving N consecutive diagonals in T_N, i.e. diagonals with elements $t_n, 1-M \leq n \leq N-M$, and using them to form a circulant matrix. One simple rule to determine M is to choose its value such that $|t_{N-M}| \approx |t_{1-M}|$. Note that half of the elements in T_N are not used in constructing S_N. The generalized preconditioner K_N is obtained from a $2N \times 2N$ circulant matrix in such a way that all elements in T_N are used, and is a circulant matrix itself (See §2). Since S_N and K_N are circulant, the matrix-vector products $S_N^{-1}\mathbf{v}$ and $K_N^{-1}\mathbf{v}$ can be conveniently computed via Fast Fourier Transform (FFT) with $O(N \log N)$ operations. The system of equations associated with the preconditioned Toeplitz matrix is then solved by iterative methods such as CGN (the Conjugate Gradient iteration applied to the Normal equations) [4], GMRES (the Generalized Minimal Residual) [11], and CGS (the Conjugate Gradient Squared) [12].

The convergence rate of preconditioned iterative methods depends on the singular value or eigenvalue distribution of the preconditioned matrices [10]. The spectral properties of preconditioned SPD Toeplitz matrices have been widely studied. Chan and Strang [1] [2] proved that, for a symmetric Toeplitz with a positive generating function in the Wiener class, the preconditioned matrix has eigenvalues clustered around unity except a fixed number independent of N. If the Toeplitz is additionally generated by a rational function, even stronger results were proved by Trefethen [15] and the authors [8]. In contrast, relatively few results for preconditioned nonsymmetric Toeplitz have been obtained so far [9], [17].

In this research, we examine the spectral properties of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ for nonsymmetric T_N in general, and nonsymmetric rational T_N in particular. The main results of our study are stated as

* This version is dated April 12, 1991. This work was supported by the USC Faculty Research and Innovation Fund and by a National Science Foundation Research Initiation Award.

† Signal and Image Processing Institute and Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-0272. E-mail:tkku@sipi.usc.edu and cckuo@sipi.usc.edu

follows. For a large class of general Toeplitz matrices, we prove that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, or equivalently, the eigenvalues of $(S_N^{-1}T_N)^H(S_N^{-1}T_N)$ and $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$, are clustered around unity except a fixed number independent of N. If T_N is additionally generated by a rational function of order $(\alpha, \beta, \gamma, \delta)$, we are able to characterize the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ directly. We classify the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ into two classes, i.e. the outliers and the clustered eigenvalues, depending on whether they converge to 1 asymptotically. Then, (1) the number of outliers is at most $\eta = 2\min(r, s)$ where $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$; and (2) the clustered eigenvalues are confined in a disk centered at 1 with radius ϵ, where the clustering radius ϵ is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner.

With these spectral regularities, we can find appropriate preconditioned iterative methods to solve a nonsymmetric Toeplitz system efficiently. In particular, an $N \times N$ rational Toeplitz system $T_N\mathrm{x} = \mathrm{b}$ can be solved with $O(N \log N)$ operations since the number of iterations required for convergence is independent of the problem size N. To compare the performance of S_N and K_N, the $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have the same number of outliers so that they converge in the same number of iterations asymptotically. However, the performances of S_N and K_N for finite N are determined by the clustering radii of the clustered eigenvalues as well. The magnitudes of the last elements used to construct S_N and K_N are $O(|t_{N-M}| + |t_{1-M}|)$ and $O(|t_N| + |t_{-N}|)$, respectively. Since $O(|t_N| + |t_{-N}|) \leq O(|t_{N-M}| + |t_{1-M}|)$ for large N, iterative methods with preconditioner K_N converges faster than with preconditioner S_N for solving rational Toeplitz systems. This is confirmed by numerical experiments. By the parallelism provided by FFT, the iterative methods with preconditioners in circulant matrix form is highly parallelizable, and the time complexity of the method can be reduced to $O(\log N)$ if $O(N)$ processors are used.

When T_N is a symmetric rational Toeplitz, we have $r = s$ and $t_N = t_{-N}$. Consequently, the number of outliers of $K_N^{-1}T_N$ is $\eta = 2r = 2\max(\alpha, \beta)$ and the clustering radius is $O(|t_N|)$. They reduce to the case given in [8]. Although the results derived in this paper can be viewed as a generalization of the results in [8], we want to point out that the approach adopted in this research is very different from that in [8] and the proof techniques are much more involved. For example, in characterizing the clustering radius of clustered eigenvalues of $K_N^{-1}T_N$ (or $S_N^{-1}T_N$) for symmetric T_N, the intertwinning theorem of eigenvalues was exploited in [8]. However, such a theorem does not exist for nonsymmetric matrices so that we use perturbation theory for eigenvalues instead.

It is worthwhile to mention that there exists a preconditioner based on the minimum-phase LU factorization (MPLU) technique [9] which has a faster or comparable convergence rate than preconditioners S_N and K_N. However, Toeplitz preconditioners in circulant matrix form have two advantages over the MPLU preconditioner. First, the circulant preconditioning technique can be easily generalized to multidimensional Toeplitz systems. See [7] for the two-dimensional case (block Toeplitz matrices). Second, the resulting preconditioned iterative method with preconditioners in circulant form is highly parallelizable while the MPLU preconditioner has to be implemented sequentially.

This paper is organized as follows. The construction of preconditioners S_N and K_N for nonsymmetric Toeplitz T_N is discussed in §2. We describe the singular value distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for general Toeplitz in §3, and characterize the eigenvalue distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for rational Toeplitz in §4 and §5, respectively. Numerical experiments are given in §6 to illustrate the theoretical study.

2. Constructions of Toeplitz preconditioners. Let T_m be a sequence of $m \times m$ nonsymmetric Toeplitz matrices with generating sequence t_n. Then,

$$T_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{-(N-2)} & t_{-(N-1)} \\ t_1 & t_0 & t_{-1} & \cdot & t_{-(N-2)} \\ \cdot & t_1 & t_0 & \cdot & \cdot \\ t_{N-2} & \cdot & \cdot & \cdot & t_{-1} \\ t_{N-1} & t_{N-2} & \cdot & t_1 & t_0 \end{bmatrix}.$$

Following the idea proposed by Strang [13], we construct the preconditioner S_N by preserving N consecutive diagonals in T_N and bringing them around to form a circulant matrix,

$$S_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 & t_1 \\ t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} \\ t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} \\ t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} \\ t_{-1} & t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 \end{bmatrix}.$$

A simple rule of thumb to decide the value of M is to require $|t_{N-M}| \approx |t_{1-M}|$.

Generalizing the idea in [6], the preconditioner K_N is constructed based on a $2N \times 2N$ circulant matrix R_{2N},

$$R_{2N} = \begin{bmatrix} T_N & \Delta T_N \\ \Delta T_N & T_N \end{bmatrix},$$

where ΔT_N is determined by the elements of T_N to make R_{2N} circulant, i.e.,

$$\Delta T_N = \begin{bmatrix} 0 & t_{N-1} & \cdot & t_2 & t_1 \\ t_{-(N-1)} & 0 & t_{N-1} & \cdot & t_2 \\ \cdot & t_{-(N-1)} & 0 & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & 0 \end{bmatrix}.$$

This construction is motivated by the observation that the augmented circulant system,

$$\begin{bmatrix} T_N & \Delta T_N \\ \Delta T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{b} \end{bmatrix},$$

is equivalent to $(T_N + \Delta T_N)\mathbf{x} = \mathbf{b}$ so that $(T_N + \Delta T_N)^{-1}\mathbf{b}$ can be computed efficiently via FFT and

$$K_N = T_N + \Delta T_N \tag{2.1}$$

can be used as a preconditioner for T_N. Note, however, that K_N itself is also circulant and can be inverted directly via N-point FFT rather than $2N$-point FFT.

3. Spectral properties of preconditioned Toeplitz. We assume that the generating sequence t_n satisfies the following two conditions:

$$\sum_{-\infty}^{\infty} |t_n| \le B_T < \infty, \tag{3.1}$$

$$|T(e^{i\theta})| = \left| \sum_{-\infty}^{\infty} t_n e^{-in\theta} \right| \ge \mu_T > 0, \qquad \forall\theta. \tag{3.2}$$

Since $T(e^{i\theta})$ describes the asymptotic eigenvalue distribution of T_N, the above conditions imply that $||T_N||$ and $||T_N^{-1}||$ are bounded for large N and, consequently, T_N is well conditioned.

With the above conditions, the preconditioners K_N and S_N are also well conditioned for sufficiently large N due to the following theorem.

THEOREM 1. *Let T_N be an $N \times N$ Toeplitz matrix with the corresponding generating sequence satisfying (3.1) and (3.2). The $||(K_N K_N^H)^{-1}||_2$ and $||(S_N S_N^H)^{-1}||_2$ are bounded for sufficiently large N.*

Proof. Since K_N is circulant, we have

$$K_N = F_N^H D_N F_N \quad \text{and} \quad K_N^H = F_N^H D_N^H F_N,$$

where F_N is the $N \times N$ unitary Fourier matrix with $N^{-1/2}e^{-i2\pi(m-1)(n-1)/N}$ as the (m,n) element and D_N a diagonal matrix formed by the eigenvalues of K_N. Thus, K_N, K_N^H and $K_N K_N^H$ share the same eigenvectors, and the eigenvalues of $K_N K_N^H$ are

$$\lambda(K_N K_N^H) = \lambda(K_N)\lambda^*(K_N) = |\lambda(K_N)|^2.$$

Any eigenvalue of K_N belongs to the set of eigenvalues of R_{2N}, which are

$$\rho_n = \lambda_n(R_{2N}) = \sum_{k=-(N-1)}^{N-1} t_k e^{i2\pi kn/2N}, \qquad 1 \leq n \leq 2N.$$

It is clear that ρ_n is a partial sum of the infinite series $\sum_{-\infty}^{\infty} t_k e^{-ik\theta}$ with $\theta = -n\pi/N$. With (3.2), $|\rho_n| \geq \mu_T - \mu$, where μ can be made arbitrarily small by choosing sufficiently large N so that

$$||(K_N K_N^H)^{-1}||_2 \leq \frac{1}{(\mu_T - \mu)^2} < \infty.$$

Similar arguments can be used to prove the boundness of $||(S_N S_N^H)^{-1}||_2$, and the proof is completed. □

The next theorem describes the clustering property of the singular values of $K_N^{-1}T_N$ and $S_N^{-1}T_N$.

THEOREM 2. *Let T_N be an $N \times N$ Toeplitz matrix with the generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the singular values of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are clustered around unity except a fixed number independent of N*

Proof. Note that the singular value of $K_N^{-1}T_N$ is equal to the square root of the corresponding eigenvalue of $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$. Since $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$ and $(K_N K_N^H)^{-1}(T_N T_N^H)$ are similar, the eigenvalues of $(K_N K_N^H)^{-1}(T_N T_N^H)$ are examined to understand the singular values of $K_N^{-1}T_N$. With the relation $K_N = T_N + \Delta T_N$, we have

$$\lambda[(K_N K_N^H)^{-1}(T_N T_N^H)] = 1 - \lambda[(K_N K_N^H)^{-1}(K_N \Delta T_N^H + \Delta T_N K_N^H - \Delta T_N \Delta T_N^H)].$$

Let us define

$$W_N = K_N \Delta T_N^H + \Delta T_N K_N^H - \Delta T_N \Delta T_N^H,$$

and denote the corresponding $(N-2q) \times (N-2q)$ central diagonal block of $(K_N K_N^H)^{-1}$ and W_N by $\mathcal{K}_{N-2q}^{-1}$ and $\mathcal{W}_{N-2q}$, respectively. By the separation theorem (or intertwining theorem) of eigenvalues [14], [16], there are at least $N-4q$ eigenvalues of $(K_N K_N^H)^{-1}W_N$ bounded by the minimum and the maximum eigenvalues of $\mathcal{K}_{N-2q}^{-1}\mathcal{W}_{N-2q}$.

Since $\mathcal{K}_{N-2q}^{-1}$ is a submatrix of the symmetric circulant matrix $(K_N K_N^H)^{-1}$,

$$||\mathcal{K}_{N-2q}^{-1}||_2 \leq ||(K_N K_N^H)^{-1}||_2.$$

According to the definition of $\mathcal{W}_{N-2q}$,

$$\mathcal{W}_{N-2q} = \mathcal{K}\Delta\mathcal{T}^H + \Delta\mathcal{T}\mathcal{K}^H - \Delta\mathcal{T}\Delta\mathcal{T}^H,$$

where $\mathcal{K}$ and $\Delta\mathcal{T}$ are $(N-2q) \times N$ matrices formed by the central $(N-2q)$ rows of K_N and ΔT_N, respectively. It is easy to verify that, for $p = 1, \infty$,

$$||\mathcal{K}||_p \leq 2 \sum_{n=-(N-1)}^{N-1} |t_n| \leq 2 \sum_{n=-\infty}^{\infty} |t_n| \leq 2B_T < \infty,$$

and

$$||\Delta\mathcal{T}||_p \leq \sum_{n=q+1}^{N-1} (|t_n| + |t_{-n}|) \leq \sum_{n=q+1}^{\infty} (|t_n| + |t_{-n}|) = \sigma(q).$$

Since $||A||_2 \leq (||A||_1||A||_\infty)^{1/2}$ for an arbitrary matrix A, the above bounds also hold for $p = 2$. Similarly, we can argue that $||\mathcal{K}^H||_2 \leq 2B_T < \infty$ and $||\Delta\mathcal{T}^H||_2 \leq \sigma(q)$. Thus,

$$\begin{aligned} ||\mathcal{W}_{N-2q}||_2 &\leq ||\mathcal{K}||_2||\Delta\mathcal{T}^H||_2 + ||\Delta\mathcal{T}||_2||\mathcal{K}^H||_2 + ||\Delta\mathcal{T}||_2||\Delta\mathcal{T}^H||_2 \\ &\leq 4B_T\sigma(q) + \sigma^2(q). \end{aligned}$$

By using Theorem 1 and the fact that $\sigma(q)$ is smaller as q becomes larger due to (3.1), we conclude that for given ϵ there exist q and $\tilde{N}$ such that for all $N \geq \tilde{N}$,

$$||\mathcal{K}_{N-2q}^{-1}||_2||\mathcal{W}_{N-2q}||_2 \leq ||(K_N K_N^H)^{-1}||_2||\mathcal{W}_{N-2q}||_2 \leq \epsilon.$$

Hence, the eigenvalues of $(K_N K_N^H)^{-1}(T_N T_N^H)$ are confined in the interval $(1-\epsilon, 1+\epsilon)$ except at most $4q$ outlying eigenvalues. Similar arguments can be used to prove the spectral clustering property of the singular values of $S_N^{-1}T_N$. □

With the above spectral clustering property, a Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ can be solved effectively by applying the CGN method to the preconditioned system $K_N^{-1}T_N\mathbf{x} = K_N^{-1}\mathbf{b}$ or $S_N^{-1}T_N\mathbf{x} = S_N^{-1}\mathbf{b}$. When the generating function is additionally rational, we characterize the eigenvalues of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ directly. It will be detailed in the following sections.

4. Spectral properties of preconditioned rational Toeplitz $K_N^{-1}T_N$. The generating function of a sequence of Toeplitz matrices T_m is defined as

$$T(z) = \sum_{n=-\infty}^{\infty} t_n z^{-n}.$$

Let the generating function of T_N be of the form

$$T(z) = \frac{A(z^{-1})}{B(z^{-1})} + \frac{C(z)}{D(z)}, \tag{4.1}$$

where

$$\frac{A(z^{-1})}{B(z^{-1})} = \frac{a_0 + a_1 z^{-1} + \cdots + a_\alpha z^{-\alpha}}{1 + b_1 z^{-1} + \cdots + b_\beta z^{-\beta}}, \qquad \frac{C(z)}{D(z)} = \frac{c_0 + c_1 z + \cdots + c_\gamma z^\gamma}{1 + d_1 z + \cdots + d_\delta z^\delta}.$$

Note that $a_\alpha b_\beta c_\gamma d_\delta \neq 0$ and polynomials $A(z^{-1})$ and $B(z^{-1})$ (or $C(z)$ and $D(z)$) have no common factor. We call $T(z)$ a rational function of order $(\alpha, \beta, \gamma, \delta)$ and T_N a rational Toeplitz matrix. To simplify the notation, we define $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$.

The spectral properties of $K_N^{-1}T_N$ can be determined from that of $T_N^{-1}\Delta T_N$ via

$$[\lambda(K_N^{-1}T_N)]^{-1} = \lambda(T_N^{-1}(T_N + \Delta T_N)) = 1 + \lambda(T_N^{-1}\Delta T_N). \tag{4.2}$$

The eigenvalues of $K_N^{-1}T_N$ clustered around 1 correspond to those of $T_N^{-1}\Delta T_N$ clustered around 0. We summarize the procedures in examing the spectral properties of $T_N^{-1}\Delta T_N$ as follows:

Step 1: Show that the ΔT_N is asymptotically equivalent to a low rank Toeplitz matrix ΔF_N (Lemma 2).

Step 2: Study the rank of ΔF_N by transforming it to a matrix Q_F which has at most $d = r + s$ nonzero columns (Lemma 3).

Step 3: Show that the Q_F is asymptotically equivalent to a matrix $\overline{Q}_F$ which has at most $2\min(r, s)$ nonzero eigenvalues (Lemma 4).

Step 4: Use perturbation theory to determine the radius of the clustered eigenvalues of $T_N^{-1}\Delta T_N$ and $K_N^{-1}T_N$ (Lemmas 5,6 and Theorem 3).

The number of outliers of $K_N^{-1}T_N$, i.e. $2\min(r, s)$, is determined from Steps 1-3, and the clustering radius is determined from Step 4.

4.1. The number of outliers of $K_N^{-1}T_N$. Note that the sequence t_n can be recursively calculated for large $|n|$. This is stated as follows.

LEMMA 1. *The sequence t_n generated by (4.1) follows the recursions,*

$$t_{n+1} = -(b_1 t_n + b_2 t_{n-1} + \cdots + b_\beta t_{n-\beta+1}), \qquad n \geq r = \max(\alpha, \beta),$$
$$t_{n-1} = -(d_1 t_n + d_2 t_{n+1} + \cdots + d_\delta t_{n+\delta-1}), \qquad n \leq -s = -\max(\gamma, \delta). \tag{4.3}$$

Proof. Similar to the proof of Lemma 1 in [8]. □

Since elements t_n satisfy the recursion given in Lemma 1, we construct a low rank Toeplitz matrices ΔF_N as

$$\Delta F_N = F_{1,N} + F_{2,N}, \tag{4.4}$$

where

$$F_{1,N} = \begin{bmatrix} t_N & t_{N-1} & \cdot & t_2 & t_1 \\ t_{N+1} & t_N & t_{N-1} & \cdot & t_2 \\ \cdot & t_{N+1} & t_N & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N \end{bmatrix},$$

and

$$F_{2,N} = \begin{bmatrix} t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{-(N-1)} & t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{-(N-1)} & t_{-N} & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & t_{-N} \end{bmatrix},$$

and where $t_n, n \geq r$ or $n \leq -s$, are recursively defined by (4.3). Due to the recursion given by (4.3), the ranks of $F_{1,N}$ and $F_{2,N}$ are bounded by r and s, respectively. Thus, the rank of ΔF is bounded by $d = r + s$. The following lemma shows that ΔT_N and ΔF_N are in fact asymptotically equivalent.

LEMMA 2. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The ΔT_N and ΔF_N are asymptotically equivalent.*

Proof. Let us denote the difference between ΔF_N and ΔT_N by

$$\Delta E_N = \Delta F_N - \Delta T_N = \begin{bmatrix} t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{N+1} & t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{N+1} & t_N + t_{-N} & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N + t_{-N} \end{bmatrix}. \tag{4.5}$$

It can be easily verified that the l_1 and l_∞ norms of ΔE_N are both bounded by

$$\tau_E = \sum_{n=N}^{2N-1} |t_n| + \sum_{n=-N}^{-(2N-1)} |t_n|. \tag{4.6}$$

Consequently, we have

$$||\Delta E_N||_2 \leq (||\Delta E_N||_1 ||\Delta E_N||_\infty)^{1/2} \leq \tau_E.$$

Since τ_E goes to zero as N goes to infinity due to (3.1), the proof is completed. □

Since ΔT_N is asymptotically equivalent to ΔF_N and the rank of ΔF_N is bounded by d, the number of outliers of $T_N^{-1}\Delta T_N$ (or $K_N^{-1}T_N$) is bounded by d, which is however not tight. We are able to determine a tighter bound by introducing another asymptotically equivalent matrix of ΔT_N (or ΔF_N), which has only $2\min(r, s)$ nonzero eigenvalues in the following. This turns out to be the exact number of outliers actually observed in all our numerical experiments. To exploit the low rank structure of ΔF_N, we transform ΔF_N to

$$Q_F = \Delta F_N U_D L_B, \tag{4.7}$$

where U_D is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $D(z)$ as the first row, and L_B is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $B(z^{-1})$ as the first column. Note that since U_D and L_B are full rank matrices, the Q_F and ΔF_N have the same rank. The structure of Q_F is described in the following lemma.

LEMMA 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The elements of Q_F are zeros except the first s and the last r columns.*

Proof. Note that $F_{1,N}$ and $F_{2,N}$ are Toeplitz matrices with elements

$$(F_{1,N})_{i,j} = t_{N+i-j} \quad \text{and} \quad (F_{2,N})_{i,j} = t_{-N+i-j}.$$

The (i,j) elements of $F_{1,N}U_DL_B$ and $F_{2,N}U_DL_B$ are

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m}d_{n-m}b_{n-j} \quad \text{and} \quad \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m}d_{n-m}b_{n-j},$$

where $b_0 = 1$ ($d_0 = 1$) and $b_i = 0$ ($d_i = 0$) if the subscript i is not in the range $0 \le i \le \beta$ ($0 \le i \le \delta$). If $s < j \le N-r$, we can simplify the above summations as

$$\sum_{m'=0}^{\delta}\left(\sum_{n'=0}^{\beta} t_{N+i+m'-n'-j}b_{n'}\right)d_{m'} = 0 \quad \text{and} \quad \sum_{n'=0}^{\beta}\left(\sum_{m'=0}^{\delta} t_{-N+i+m'-n'-j}d_{m'}\right)b_{n'} = 0,$$

where $m' = n - m$, $n' = n - j$, and the equalities are due to the recursion defined in (4.3). Thus, the elements of

$$Q_F = \Delta F_N U_D L_B = (F_{1,N} + F_{2,N})U_D L_B$$

are zeros except the first s and the last r columns. □

Consequently, we decompose the complex N-tuple space C^N into two orthogonal complement subspaces,

$$\begin{aligned} \mathcal{R}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ s < i \le N-r\}, \\ \mathcal{N}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ 1 \le i \le s \ \text{ or } \ N-r < i \le N\}, \end{aligned} \tag{4.8}$$

with dimensions

$$\dim \mathcal{R}(Q_F) = d \quad \text{and} \quad \dim \mathcal{N}(Q_F) = N - d.$$

The subspace $\mathcal{N}(Q_F)$ is contained in the null space of Q_F. Let Q_{NW} denote the northwest $s \times s$ block in Q_F, and Q_{NE}, Q_{SW} and Q_{SE} the corresponding corner blocks in Q_F with sizes $s \times r$, $r \times s$ and $r \times r$, respectively. By using the subspace decomposition (4.8), it is easy to see that the nonzero eigenvalues of Q_F only depend on the corresponding four corner blocks of Q_F, and are also the eigenvalues of the $d \times d$ matrix,

$$P_F = \begin{bmatrix} Q_{NW} & Q_{NE} \\ Q_{SW} & Q_{SE} \end{bmatrix}.$$

In other words, the rank of Q_F is the same as that of P_F.

The bounds for the elements of Q_{NW}, Q_{NE}, Q_{SW} and Q_{SE} are summarized as follows:

$$\begin{aligned} &|(Q_{NW})_{i,j}| \le \tau_{NW}, && \tau_{NW} = O(|t_N| + |t_{-N}|), \\ &|(Q_{SE})_{i,j}| \le \tau_{SE}, && \tau_{SE} = O(|t_N| + |t_{-N}|), \\ &|(Q_{NE})_{i,j}| \le (F_{1,N}U_DL_B)_{i,N-r+j} + \tau_{NE}, && \tau_{NE} = O(|t_{-2N}|), \\ &|(Q_{SW})_{i,j}| \le (F_{2,N}U_DL_B)_{N-s+i,j} + \tau_{SW}, && \tau_{SW} = O(|t_{2N}|). \end{aligned} \tag{4.9}$$

To derive (4.9), recall that the (i,j) element of Q_F is

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m}d_{n-m}b_{n-j} + \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m}d_{n-m}b_{n-j},$$

which is bounded by

$$\sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{N+i+m'-n'-j}||d_{m'}||b_{n'}| + \sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{-N+i+m'-n'-j}||d_{m'}||b_{n'}|.$$

Since the elements of Q_{NW} are the same as those of Q_F with subscript (i,j), $i,j \leq s$, they are bounded by

$$\tau_{NW} = \sum_{i=0}^{\beta}|b_i|\sum_{j=0}^{\delta}|d_j|(\max_{-(s+\beta)<n<s+\delta}|t_{N+n}| + \max_{-(s+\beta)<n<s+\delta}|t_{-N+n}|).$$

To determine the bound for $\sum_{i=0}^{\beta}|b_i|$, we factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1-r_1z^{-1})(1-r_2z^{-1})\cdots(1-r_\beta z^{-1}).$$

A direct consequence of (3.1) is that all poles of $A(z^{-1})/B(z^{-1})$ should lie inside the unit circle, i.e. $|r_i|<1$, $1\leq i\leq\beta$, so that

$$|b_k| \leq \binom{\beta}{k}(\max|r_i|)^k \leq \binom{\beta}{k}, \quad \text{where} \quad \binom{\beta}{k} \equiv \frac{\beta!}{(\beta-k)!k!}.$$

Therefore, we obtain

$$\sum_{k=0}^{\beta}|b_k| \leq \sum_{k=0}^{\beta}\binom{\beta}{k} \leq 2^\beta.$$

Similarly, $\sum_{k=0}^{\delta}|d_k| \leq 2^\delta$ and thus, the elements of Q_{NW} are bounded by

$$\tau_{NW} = 2^{(\beta+\delta)}(|t_{N-s-\beta}| + |t_{-N+s+\delta}|) = O(|t_N| + |t_{-N}|),$$

where the last equality is due to the fact that, for large n, t_n can be approximated by

$$t_n \approx cr_j^n, \quad \text{where} \quad |r_j| = \max_i |r_i|, \tag{4.10}$$

and where c is a constant. Similarly, we can prove that the elements of Q_{SE} are bounded by

$$\tau_{SE} = 2^{(\beta+\delta)}(|t_{N-r-\beta}| + |t_{-N+r+\delta}|) = O(|t_N| + |t_{-N}|).$$

The (i,j), $1\leq i\leq s, 1\leq j\leq r$, element of Q_{NE} is the sum of the $(i, N-r+j)$ elements of $F_{1,N}U_DL_B$ and $F_{2,N}U_DL_B$. It is straightforward to verify that the $(i,N-r+j)$ element of $F_{1,N}U_DL_B$ remains unchanged while that of $F_{2,N}U_DL_B$ is bounded by $\tau_{NE} = 2^{(\beta+\delta)}|t_{-2N+d+\delta}| = O(|t_{-2N}|)$ for sufficiently large N. Similarly, we can derive the bound for the elements in Q_{SW} as given by (4.9).

Thus, when N becomes asymptotically large, the P_F converges to

$$\overline{P}_F = \begin{bmatrix} 0 & \overline{Q}_{NE} \\ \overline{Q}_{SW} & 0 \end{bmatrix},$$

where $\overline{Q}_{NE}$ is the converged northeast $s\times r$ block in $F_{1,N}U_DL_B$ and $\overline{Q}_{SW}$ is the converged southwest $r\times s$ block in $F_{2,N}U_DL_B$. Since the ranks of $\overline{Q}_{NE}$ and $\overline{Q}_{SW}$ are both bounded by $\min(r,s)$, the rank of $\overline{P}_F$ is bounded by $\eta = 2\min(r,s)$.

Let us define a matrix $\overline{Q}_F$ by replacing the four corner blocks in Q_F with the corresponding blocks in $\overline{P}_F$. Then, we have

$$\begin{aligned}\tau_Q &= \|Q_F - \overline{Q}_F\|_p = \|P_F - \overline{P}_F\|_p \\ &\leq s\tau_{NW} + r\tau_{SE} + \max(r,s)(\tau_{NE} + \tau_{SW}) \\ &= O(|t_N| + |t_{-N}|),\end{aligned}$$

for $p = 1$ and ∞. The above bounds also hold for $p = 2$ because $\|A\|_2 \leq (\|A\|_1\|A\|_\infty)^{1/2}$ for an arbitrary matrix A. Since τ_Q goes to zero as N goes to infinity due to (3.1), the asymptotic equivalence between Q_F and $\overline{Q}_F$ is established. This result is summarized in the following lemma.

LEMMA 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). The Q_F and $\overline{Q}_F$ are asymptotically equivalent.*

Based on Lemmas 2-4, (4.2) and (4.7), $T_N^{-1}\Delta T_N$ is asymptotically equivalent to $T_N^{-1}\overline{Q}_F L_B^{-1}U_D^{-1}$ whose rank is bounded by $\eta = 2\min(r,s)$ and $K_N^{-1}T_N$ has at most η asymptotic eigenvalues not converging to one (outliers).

4.2. The clustering radius of $K_N^{-1}T_N$. We use perturbation theory to estimate the clustering radius of the $N-\eta$ clustered eigenvalues. Instead of examining the eigenvalues of $T_N^{-1}\Delta T_N$ directly, we study those of the similar matrix

$$G_N = L_B^{-1}U_D^{-1}T_N^{-1}\Delta T_N U_D L_B = L_B^{-1}U_D^{-1}T_N^{-1}Q_T,$$

where $Q_T = \Delta T_N U_D L_B$. Let us define

$$H_N = L_B^{-1}U_D^{-1}T_N^{-1}\overline{Q}_F.$$

It is clear that H_N has only d nonzero columns as $\overline{Q}_F$ (or Q_F). The G_N can be viewed as a matrix obtained from H_N by adding the perturbation matrix

$$\Delta G_N = G_N - H_N = L_B^{-1}U_D^{-1}T_N^{-1}(Q_T - \overline{Q}_F). \tag{4.11}$$

A bound of $\|\Delta G_N\|_2$ is given below so that we can estimate the clustering radius of the clustered eigenvalues by using perturbation theory for eigenvalues.

LEMMA 5. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). Then, for sufficiently large N, the $\|\Delta G_N\|_2$ is bounded by $\epsilon = O(|t_N| + |t_{-N}|)$.*

Proof. We first study the 2-norm of $Q_T - \overline{Q}_F$, which is bounded by

$$\|Q_T - \overline{Q}_F\|_2 \leq \|Q_T - Q_F\|_2 + \|Q_F - \overline{Q}_F\|_2.$$

As shown in the proof of Lemma 4, the second term $\|Q_F - \overline{Q}_F\|_2$ is bounded by $\tau_Q = O(|t_N| + |t_{-N}|)$ while the first term $\|Q_T - Q_F\|_2$ is bounded by

$$\|Q_T - Q_F\|_2 \leq \|\Delta T_N - \Delta F_N\|_2\|U_D\|_2\|L_B\|_2 = \|\Delta E_N\|_2\|U_D\|_2\|L_B\|_2.$$

Recall from (4.6) that $\|\Delta E_N\|_2 \leq \sum_{n=N}^{2N-1}(|t_n| + |t_{-n}|)$. By using (4.10), we have

$$\sum_{n=N}^{2N-1}|t_n| \leq \sum_{n=N}^{\infty}|q_j r_j^n| = \frac{|t_N|}{1-|r_j|} = M_B|t_N|, \quad \text{where} \quad M_B = \frac{1}{1-|r_j|}.$$

Similarly, $\sum_{n=N}^{2N-1}|t_{-n}| \leq M_D|t_{-N}|$. Besides, $\|L_B\|_2 \leq \sum_{k=0}^{\beta}|b_k| \leq 2^\beta$ and $\|U_D\|_2 \leq \sum_{k=0}^{\delta}|d_k| \leq 2^\delta$. Thus, we obtain a bound for the first term, i.e.

$$\|Q_T - Q_F\|_2 \leq 2^{(\beta+\delta)}(M_B|t_N| + M_D|t_{-N}|) = O(|t_N| + |t_{-N}|),$$

and conclude that

$$\|Q_T - \overline{Q}_F\|_2 \leq O(|t_N| + |t_{-N}|).$$

With (4.11), we have

$$||\triangle G_N||_2 \le ||L_B^{-1}||_2||U_D^{-1}||_2||T_N^{-1}||_2||(Q_T - \overline{Q}_F)||_2.$$

Due to (3.2), $||T_N^{-1}||_2$ is bounded by a constant c_T independent of N. To show that $||L_B^{-1}||_2$ and $||U_D^{-1}||_2$ are also bounded, we factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1 - r_1 z^{-1})(1 - r_2 z^{-1}) \cdots (1 - r_\beta z^{-1}),$$

where we assume that all roots r_i are distinct for simplicity. By applying the isomorphism between the ring of the power series and the ring of semi-infinite lower (or upper) triangular Toeplitz matrices, the L_B and L_B^{-1} can be decomposed into the products,

$$L_B = L_{r_1} L_{r_2} \cdots L_{r_\beta}, \qquad L_B^{-1} = L_{r_\beta}^{-1} \cdots L_{r_2}^{-1} L_{r_1}^{-1},$$

where $L_{r_i}, 1 \le i \le \beta$ is an $N \times N$ lower triangular Toeplitz matrix with $[1, -r_i, 0, \cdots, 0]^T$ as the first column. It can be easily verified that $L_{r_i}^{-1}$ is a lower triangular Toeplitz matrix with $[1, r_i, r_i^2, \cdots, r_i^{N-1}]^T$ as the first column. Therefore,

$$||L_{r_i}^{-1}||_p \le \sum_{k=0}^{N-1} |r_i^k| \le \sum_{k=0}^{\infty} |r_i^k| = \frac{1}{1 - |r_i|}, \qquad p = 1, 2, \infty,$$

and

$$||L_B^{-1}||_2 \le \prod_{i=1}^{\beta} ||L_{r_i}^{-1}||_2 \le \prod_{i=1}^{\beta} \frac{1}{1 - |r_i|} = c_B.$$

Similar arguments can be used to prove that $||U_D^{-1}||_2 \le c_D$. Finally, we have

$$||\triangle G_N||_2 \le \epsilon \equiv c_B c_D c_T ||(Q_T - \overline{Q}_F)||_2 = O(|t_N| + |t_{-N}|). \tag{4.12}$$

The proof is completed. □

Let us denote the rank of $H_N = L_B^{-1} U_D^{-1} T_N^{-1} \overline{Q}_F$ by $\tilde{\eta}$. Clearly, $\tilde{\eta} \le \eta = 2\min(r, s)$. We arrange the eigenvalues of H_N in a descending order so that $|\lambda_n| \ge |\lambda_{n+1}|$ ($\lambda_n = 0$ for $\tilde{\eta} < n \le N$), and denote the corresponding normalized right-hand and left-hand eigenvectors by $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ and $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$, respectively. Besides, vectors $\mathbf{x}_n$ with $\tilde{\eta} < n \le N$ are chosen to be othorgonal. The complex N-tuple space is decomposed into the row and the null spaces of H_N,

$$\text{Row}(H_N) = \text{span}\{\mathbf{x}_n, n \le \tilde{\eta}\}, \qquad \text{Null}(H_N) = \text{span}\{\mathbf{x}_n, \tilde{\eta} < n \le N\}.$$

Since $G_N = H_N + \triangle G_N$ and $||\triangle G_N||_2 \le \epsilon$, the eigenvalues and the right-hand eigenvectors of G_N are denoted by $\lambda_n(\epsilon)$ and $\mathbf{x}_n(\epsilon)$, respectively. By using results from perturbation theory for repeated eigenvalues [16], the eigenvectors $\mathbf{x}_n(\epsilon)$ with $\tilde{\eta} < n \le N$ must take the form

$$\mathbf{x}_n(\epsilon) = \sum_{m=1}^{\tilde{\eta}} \frac{\xi_{mn}}{(\lambda_n - \lambda_m) s_m} \mathbf{x}_m + \sum_{m=\tilde{\eta}+1}^{N} g_{mn} \mathbf{x}_m + O(\epsilon^2), \tag{4.13}$$

where $\xi_{mn} = \mathbf{y}_m^H \triangle G_N \mathbf{x}_n$, $\lambda_n = 0$, $s_m = \mathbf{y}_m^H \mathbf{x}_m$ and $g_{nn} = 1$. Due to the construction, we know that

$$||\mathbf{x}_n(\epsilon)||_2 \ge ||\mathbf{x}_n||_2 = 1. \tag{4.14}$$

The factor $|\xi_{mn}|$ is bounded by

$$|\xi_{mn}| = |\mathbf{y}_m^H \triangle G_N \mathbf{x}_n| \le ||\mathbf{y}_m||_2 ||\triangle G_N||_2 ||\mathbf{x}_n||_2 \le \epsilon.$$

The $|s_m^{-1}|, 1 \le m \le \tilde{\eta}$, is also bounded as given in the following lemma.

LEMMA 6. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). Then, the $|s_m^{-1}|, 1 \le m \le \tilde{\eta}$, of H_N is bounded by a constant independent of N.*

Proof. The eigenvalues λ and the right-hand eigenvectors $\mathbf{x}$ of H_N satisfy

$$L_B \overline{Q}_F \mathbf{x} = \lambda L_B T_N U_D L_B \mathbf{x}.$$

Since the elements of $\overline{Q}_F$ are zeros except the first s and the last r columns, so are the elements of $L_B\overline{Q}_F$. Thus, the nonzero eigenvalues of H_N only depend on the northwest $s \times s$, northeast $s \times r$, southwest $r \times s$ and southeast $r \times r$ blocks of $L_B\overline{Q}_F$ and $L_B T_N U_D L_B$. The boundness of $|s_m^{-1}|, 1 \le m \le \tilde{\eta}$, is guaranteed if the elements of the four corner blocks of $L_B\overline{Q}_F$ and $L_B T_N U_D L_B$ remain unchanged for sufficiently large N.

By using the band structure of L_B and the special structure of $\overline{Q}_F$, it is straightforward to verify that the four blocks of $L_B\overline{Q}_F$ remain unchanged for large N. Next, we examine the matrix $L_B T_N U_D L_B$. By using (4.1) and the isomorphism between the ring of the power series and the ring of the semi-infinite lower (or upper) triangular Toeplitz matrices, we can express T_N as

$$T_N = L_A L_B^{-1} + U_C U_D^{-1},$$

where L_A is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $A(z^{-1})$ as the first column, and U_C is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $C(z)$ as the first row. Then, we have

$$L_B T_N U_D L_B = L_A U_D L_B + L_B U_C L_B,$$

whose four corner blocks remain unchanged for large N. Thus, λ_m and $s_m = \mathbf{y}_m^H \mathbf{x}_m$ with $1 \le m \le \tilde{\eta}$, do not change with N, when N becomes sufficiently large. □

Let $\mathbf{v}_n(\epsilon)$ be the normalized vector of $\mathbf{x}_n(\epsilon)$,

$$\mathbf{v}_n(\epsilon) = \frac{\mathbf{x}_n(\epsilon)}{||\mathbf{x}_n(\epsilon)||_2},$$

which can be decomposed as

$$\mathbf{v}_n(\epsilon) = \mathbf{v}_{\mathrm{N}}(\epsilon) + \mathbf{v}_{\mathrm{R}}(\epsilon),$$

where $\mathbf{v}_{\mathrm{N}}(\epsilon) \in \mathrm{Null}(H_N)$ and $\mathbf{v}_{\mathrm{R}}(\epsilon) \in \mathrm{Row}(H_N)$. The magnitude of $\lambda_n(\epsilon)$, $\tilde{\eta} < n \le N$, of G_N is approximated by

$$|\lambda_n(\epsilon)| = ||G_N \mathbf{v}_n(\epsilon)||_2 = ||H_N \mathbf{v}_{\mathrm{R}}(\epsilon) + \Delta G_N \mathbf{v}_n(\epsilon)||_2.$$

By using (4.12)-(4.14), we obtain that

$$\begin{aligned}
\max_{\tilde{\eta}<n\le N} |\lambda_n(\epsilon)| &\le \max_{\tilde{\eta}<n\le N} ||H_N \mathbf{v}_{\mathrm{R}}(\epsilon)||_2 + \max_{\tilde{\eta}<n\le N} ||\Delta G_N \mathbf{v}_n(\epsilon)||_2 \\
&\le \sum_{m=1}^{\tilde{\eta}} \frac{||\xi_{mn} H_N \mathbf{x}_m||_2}{|\lambda_m s_m| \, ||\mathbf{x}_n(\epsilon)||_2} + ||\Delta G_N||_2 \\
&\le \sum_{m=1}^{\tilde{\eta}} \frac{\epsilon}{|s_m|} + \epsilon = \epsilon_K \\
&= O(|t_N| + |t_{-N}|),
\end{aligned}$$

for sufficiently large N. The above analysis is concluded in the following theorem.

THEOREM 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the preconditioned Toeplitz matrix $K_N^{-1}T_N$ has the following two properties:*
P1: The number of outliers is at most $\eta = 2\min(r, s)$.
P2: There are at least $N - \eta$ eigenvalues confined in the disk centered at 1 with radius ϵ_K, where

$$\epsilon_K = O(|t_N| + |t_{-N}|).$$

5. Spectral properties of preconditioned rational Toeplitz $S_N^{-1}T_N$. The preconditioned Toeplitz matrix $S_N^{-1}T_N$ has similar spectral properties as $K_N^{-1}T_N$. The number of outliers of $S_N^{-1}T_N$ can be obtained by proving that $\Delta S_N = S_N - T_N$ and ΔF_N given by (4.4) are asymptotically equivalent.

LEMMA 7. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). $S_N^{-1}T_N$ has asymptotically at most $\eta = 2\min(r,s)$ eigenvalues not converging to 1.*

Proof. Let us define $\Delta S_N = S_N - T_N$, and express the difference between ΔF_N in (4.4) and ΔS_N as

$$\Delta F_N - \Delta S_N = E_{1,N} + E_{2,N},$$

where $E_{1,N}$ and $E_{2,N}$ are $N \times N$ Toeplitz matrices with elements

$$(E_{1,N})_{i,j} = \begin{cases} t_{N+i-j}, & -(M-1) \le i-j \le N-1, \\ t_{i-j}, & -(N-1) \le i-j \le -M, \end{cases}$$

and

$$(E_{2,N})_{i,j} = \begin{cases} t_{i-j}, & N-(M-1) \le i-j \le N-1, \\ t_{i-j-N}, & -(N-1) \le i-j \le N-M, \end{cases}$$

respectively. By using similar arguments in deriving Lemma 2, we can prove that ΔS_N and ΔF_N are asymptotically equivalent. Since ΔF_N is asymptotically equivalent to the matrix $\overline{Q}_F L_B^{-1} U_D^{-1}$ with rank $\tilde{\eta} \le \eta = 2\min(r,s)$ as described in Lemma 4, the proof is completed. □

Similar arguments used in §4.2 can be applied to derive the following theorem.

THEOREM 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in (4.1) with the corresponding generating sequence satisfying (3.1) and (3.2). For sufficiently large N, the preconditioned Toeplitz matrix $S_N^{-1}T_N$ has the following two properties:*
P1: The number of outliers is at most $\eta = 2\min(r,s)$.
P2: There are at least $N - \eta$ eigenvalues confined in the disk centered at 1 with radius ϵ_S, where

$$\epsilon_S = O(|t_{N-M}| + |t_{1-M}|).$$

6. Numerical results. Four test problems, including both rational and nonrational T_N, are used to illustrate the above analysis. For the nonsymmetric Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ to be solved, we choose $\mathbf{b} = (1, \cdots, 1)^T$ and zero initial guess in all experiments. Without further specification, M is chosen such that $|t_{N-M}| \approx |t_{1-M}|$ to construct preconditioner S_N. We use the first test problem, which is generated by a nonrational function, to examine the clustering effect of singular values. Test problems 2-4 are generated by rational functions so that the number of outliers and the clustering radius can be observed, which confirm the theoretical results developed in §4 and §5.

Test Problem 1. Nonrational T_N.

Let T_N be a Toeplitz matrix with generating sequence

$$t_n = \begin{cases} 1/\log(2-n), & n \le -1, \\ 1/\log(2-n) + 1/(1+n), & n = 0, \\ 1/(1+n), & n \ge 1. \end{cases}$$

The singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are plotted in Fig. 1(a) for N = 32, 64 and 128. Both $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have clustered singular values. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 1(b). It is clear that the eigenvalues possess a certain clustering property. We apply both the CGN and CGS methods to solve the preconditioned Toeplitz system $P_N^{-1}T_N\mathbf{x} = P_N^{-1}\mathbf{b}$. The numbers of iterations required for the CGN and CGS methods to achieve $\|\mathbf{b} - T_N\mathbf{x}\|_2 < 10^{-12}$ are summarized in Tables 1 and 2, respectively. The case without preconditioning is also included for comparison. The use of preconditioners does accelerate the convergence rate of iterative methods. The numbers of

N	T_N	S_N	K_N
32	24	12	9
64	33	15	11
128	49	17	13

TABLE 1
The numbers of iterations required for the CGN method.

N	T_N	S_N	K_N
32	15	7	9
64	21	8	10
128	26	9	10

TABLE 2
The numbers of iterations required for the CGS method.

iterations required for S_N and K_N increase slightly as N becomes large. The K_N performs better than S_N in the CGN method. However, their performances are comparable for the CGS method. Since the CGN method in general requires more iterations than the CGS method and the convergence rate of the CGS method is related to the eigenvalue distribution of the iteration matrix, we will only present the results of the CGS method for the remaining three test problems.

Test Problem 2. Rational T_N with $(r, s) = (1, 1)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{1+0.7z^{-1}}{1-0.9z^{-1}} + \frac{1-0.8z}{1+0.7z}.$$

To show that the simple rule for choosing M, i.e. $|t_{N-M}| \approx |t_{1-M}|$, does provide a better spectral clustering property and a better convergence rate for $S_N^{-1}T_N$, two preconditioners S_N and $\tilde{S}_N$ are constructed. The S_N is constructed with M such that $|t_{N-M}| \approx |t_{1-M}|$ while the $\tilde{S}_N$ is constructed with $M = \lceil N/2 \rceil$. The eigenvalues of T_N, $\tilde{S}_N^{-1}T_N$, $S_N^{-1}T_N$ and $K_N^{-1}T_N$ with $N = 32$ are plotted in Figs. 2(a)-(d). All preconditioned Toeplitz matrices have eigenvalues clustered around 1 except $2 = 2\min(r, s)$ outliers. The $K_N^{-1}T_N$ has the best clustering effect, and the eigenvalues of $S_N^{-1}T_N$ are more closely clustered than those of $\tilde{S}_N^{-1}T_N$. The sums of magnitudes of the last elements in constructing S_N and K_N and the corresponding clustering radii are listed in Table 3. They are approximately of the same order, as stated in Theorems 3 and 4.

The convergence history of the CGS method with various preconditioners is plotted in Fig. 3 with $N = 32$. The convergence rate of the CGS method without preconditioning (the curve denoted by T_N) is very slow. This phenomenon is not surprising by examining the eigenvalue distribution given in Fig. 2(a). Preconditioning improves the convergence behavior dramatically. It is clear that K_N performs the best while S_N performs better than $\tilde{S}_N$.

Test Problem 3. Rational T_N with $(r, s) = (3, 1)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{(1+0.5z^{-1})(1+0.7z^{-1})}{(1-0.4z^{-1})(1-0.6z^{-1})(1-0.8z^{-1})} + \frac{1+0.8z}{1+0.9z}.$$

The eigenvalues of T_N, $S_N^{-1}T_N$ and $K_N^{-1}T_N$ with $N = 64$ are plotted in Figs. 4(a)-(c). It is clear that $K_N^{-1}T_N$ has $2 = 2\min(r, s)$ outliers. The outliers of $S_N^{-1}T_N$ are not easy to identify for this case.

N	ϵ_S	$\|t_{N-M}\|+\|t_{1-M}\|$	ϵ_K	$\|t_{N-1}\|+\|t_{1-N}\|$
32	8.2×10^{-2}	2.8×10^{-1}	3.5×10^{-2}	6.8×10^{-2}
64	4.6×10^{-2}	2.1×10^{-2}	1.2×10^{-3}	2.3×10^{-3}
128	3.3×10^{-5}	1.1×10^{-4}	1.4×10^{-6}	2.7×10^{-6}

TABLE 3
The clustering radii ϵ of preconditioners S_N and K_N for Test Problem 2.

N	ϵ_S	$\|t_{N-M}\|+\|t_{1-M}\|$	ϵ_K	$\|t_{N-1}\|+\|t_{1-N}\|$
32	1.7×10^{-1}	1.4×10^{-1}	6.1×10^{-2}	2.8×10^{-2}
64	2.7×10^{-2}	1.3×10^{-2}	5.1×10^{-4}	1.6×10^{-4}
128	1.7×10^{-3}	1.6×10^{-4}	5.8×10^{-7}	1.7×10^{-7}

TABLE 4
The clustering radii ϵ of preconditioners S_N and K_N for Test Problem 3.

However, two outliers can be observed more easily for larger N. Besides, the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$. We list in Table 4 the sums of magnitudes of the last elements in constructing S_N and K_N and the corresponding clustering radii. The convergence history of the CGS method with $N = 64$ is plotted in Fig. 5. We observe that the CGS method without preconditioning does not converge and that the CGS method with preconditioners K_N and S_N converges in 4 and 6 iterations, respectively. This seems to suggest that the use of preconditioners does not only accelerate the convergence rate by providing better spectral properties but also improves the convergence of nonsymmetric iterative algorithms by making the preconditioned matrix more close to normal.

Test Problem 4. Rational triangular T_N with $(r,s) = (1,0)$.
The generating function of T_N is chosen to be

$$T(z) = \frac{1-0.7z^{-1}}{1+0.5z^{-1}}.$$

Since there are only N nonzero elements in T_N, we can make S_N the same as K_N. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 6(a). We see that all eigenvalues are clustered around 1 with radius $\epsilon_K = O(|t_N|) = 10^{-9}$. This is consistent with Theorem 3, which predicts that $K_N^{-1}T_N$ has $0 = 2\min(r,s)$ outliers. The convergence history of the CGS method with $N = 32$ is plotted in Fig. 6(b). The CGS method with preconditioner K_N converges in two iterations while the CGS method without preconditioning does not converge.

7. Conclusion. In this paper, we generalized the circulant preconditioning technique from symmetric to nonsymmetric Toeplitz matrices. The resulting preconditioned Toeplitz systems are then solved by various iterative methods such as CGN and CGS. For a large class of Toeplitz matrices, we proved that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except a fixed number independent of N. When the generating function is rational, the eigenvalues of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are classified into clustered eigenvalues and outliers. The number of outliers depends on the order of the rational generating function. The clustered eigenvalues are confined in the disk centered at 1 with the radii $\epsilon_K = O(|t_N| + |t_{-N}|)$ and $\epsilon_S = O(|t_{N-M}| + |t_{1-M}|)$ for $K_N^{-1}T_N$ and $S_N^{-1}T_N$, respectively. Since the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$, preconditioner K_N performs better than S_N for solving rational Toeplitz systems.

REFERENCES

[1] R. H. CHAN, *Circulant preconditioners for Hermitian Toeplitz system*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.

[2] R. H. CHAN AND G. STRANG, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 104–119.

[3] T. F. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 766–771.

[4] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.

[5] T. HUCKLE, *Circulant and skew-circulant matrices for solving Toeplitz matrices problems*, in Cooper Mountain Conference on Iterative Methods, Cooper Mountain, Colorado, 1990.

[6] T. K. KU AND C. J. KUO, *Design and analysis of Toeplitz preconditioners*, Tech. Rep. 155, USC, Signal and Image Processing Institute, May 1990. To appear in IEEE Trans. on Signal Processing, Jan. 1992.

[7] ——, *On the spectrum of a family of preconditioned block Toeplitz matrices*, Tech. Rep. 164, USC, Signal and Image Processing Institute, Nov. 1990.

[8] ——, *Spectral properties of preconditioned rational Toeplitz matrices*, Tech. Rep. 163, USC, Signal and Image Processing Institute, Sept. 1990. to appear in SIAM J. Matrix Anal. Appl.

[9] ——, *A minimum-phase LU factorization preconditioner for Toeplitz matrices*, Tech. Rep. 171, USC, Signal and Image Processing Institute, Feb. 1991.

[10] N. M. NACHTIGAL, S. C. REDDY, AND L. N. TREFETHEN, *How fast are nonsymmetric matrix iterations*, in Cooper Mountain Conference on Iterative Methods, Cooper Mountain, Colorado, 1990.

[11] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[12] P. SONNEVELD, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.

[13] G. STRANG, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[14] ——, *Linear Algebra and Its Applications*, Harcourt Brace Jonanovich, Inc., Orlando, Florida, third ed., 1988.

[15] L. N. TREFETHEN, *Approximantion theory and numerical linear algebra*, in Algorithms for Approximation II, M. Cox and J. C. Mason, eds., 1988.

[16] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

[17] A. YAMASAKI, *New preconditioners based on low-rank elimination*, Tech. Rep. Numerical Analysis 89-10, MIT, Dept. of Math., Dec. 1989.

Figure Captions

Figure 1: (a) The singular value distribution of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, and (b) the eigenvalue distribution of $K_N^{-1}T_N$ for Test Problem 1.

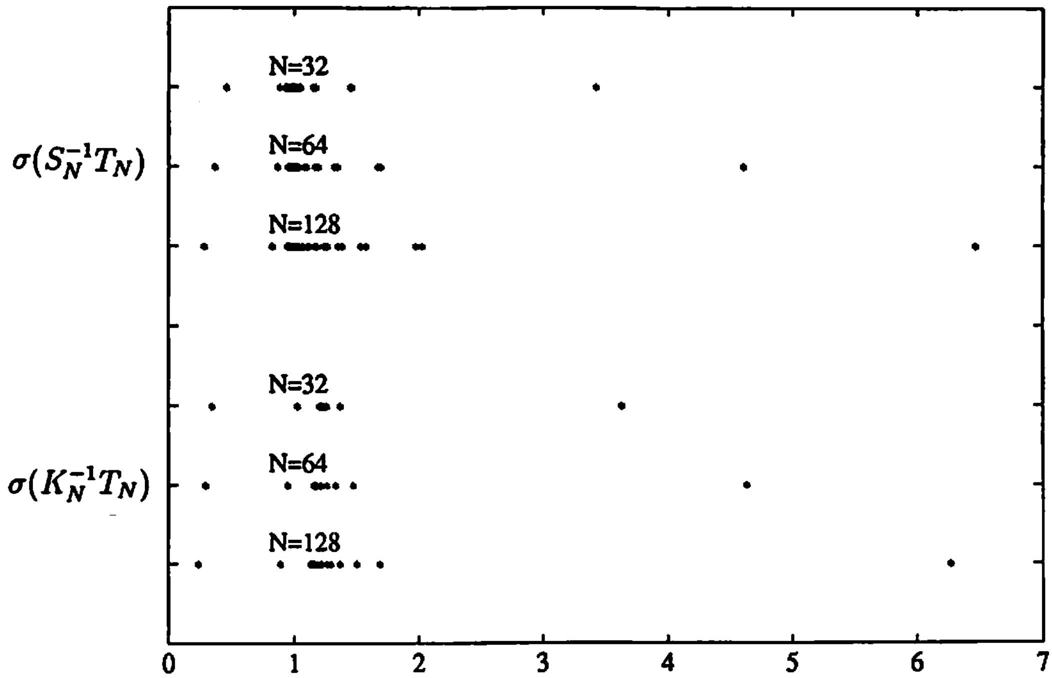
Figure 2: The eigenvalue distribution of (a) T_N, (b) $\tilde{S}_N^{-1}T_N$, (c) $S_N^{-1}T_N$ and (d) $K_N^{-1}T_N$ for Test Problem 2.

Figure 3: The convergence history of the CGS method for Test Problem 2.

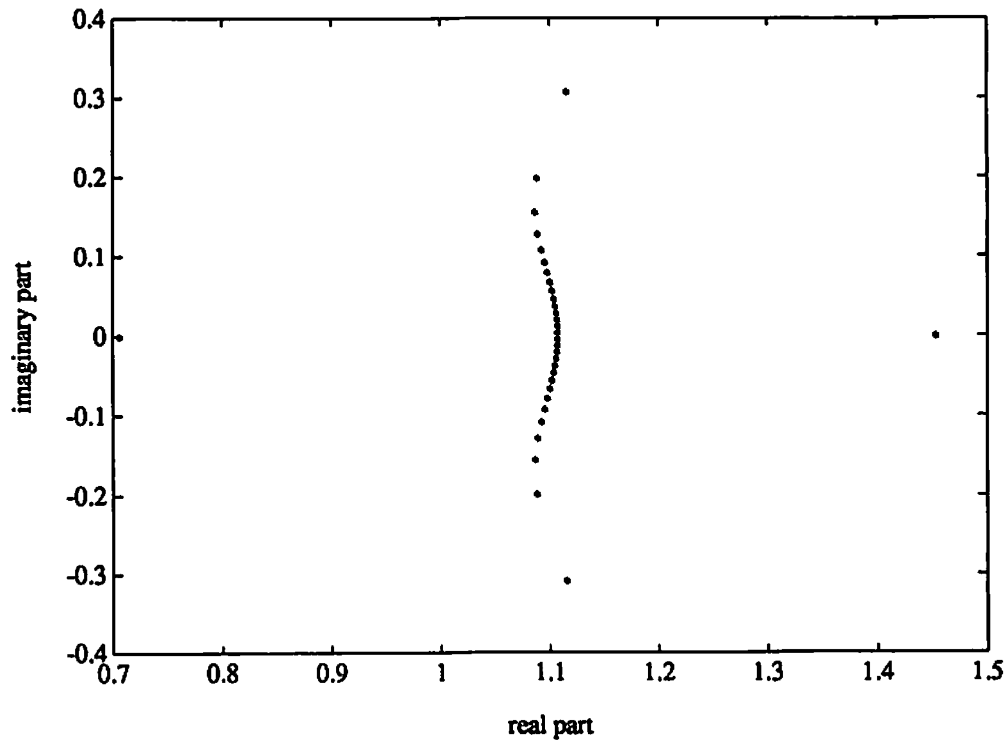
Figure 4: The eigenvalue distribution of (a) T_N, (b) $S_N^{-1}T_N$ and (c) $K_N^{-1}T_N$ for Test Problem 3.

Figure 5: The convergence history of the CGS method for Test Problem 3.

Figure 6: (a) The eigenvalue distribution of $K_N^{-1}T_N$, and (b) the convergence history of the CGS method for Test Problem 4.

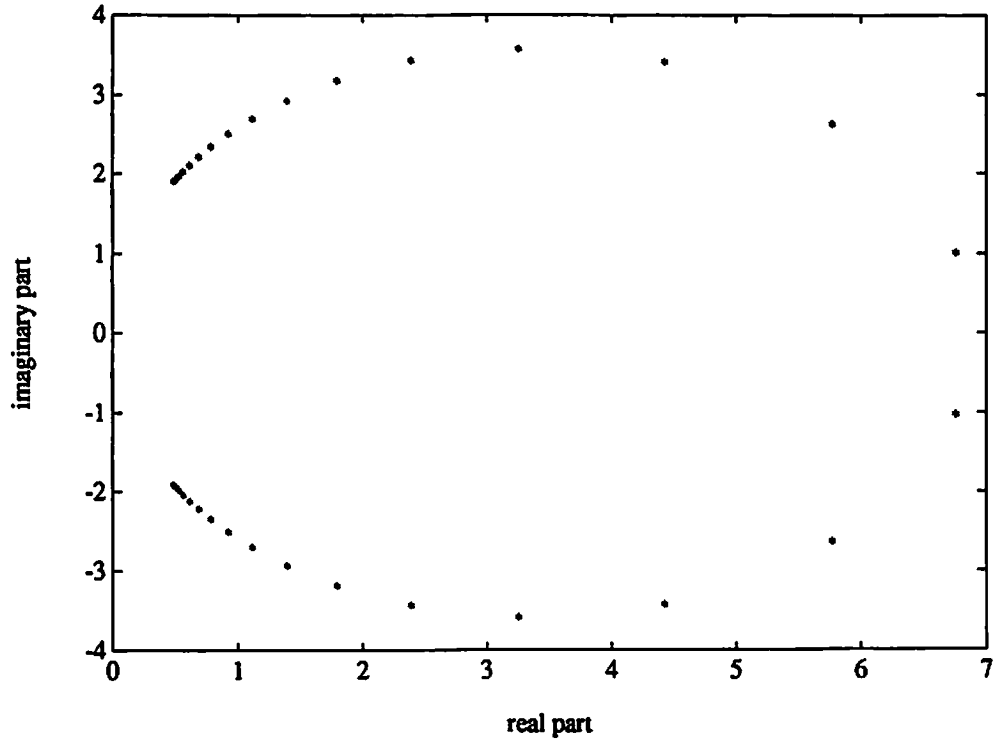


(a)

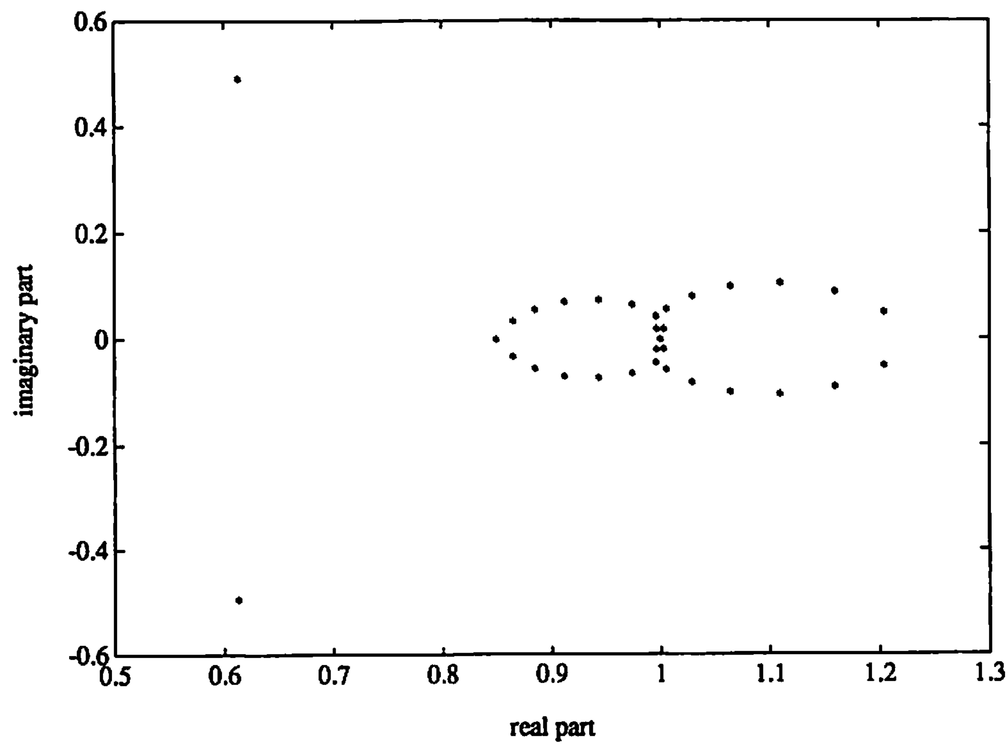


(b)

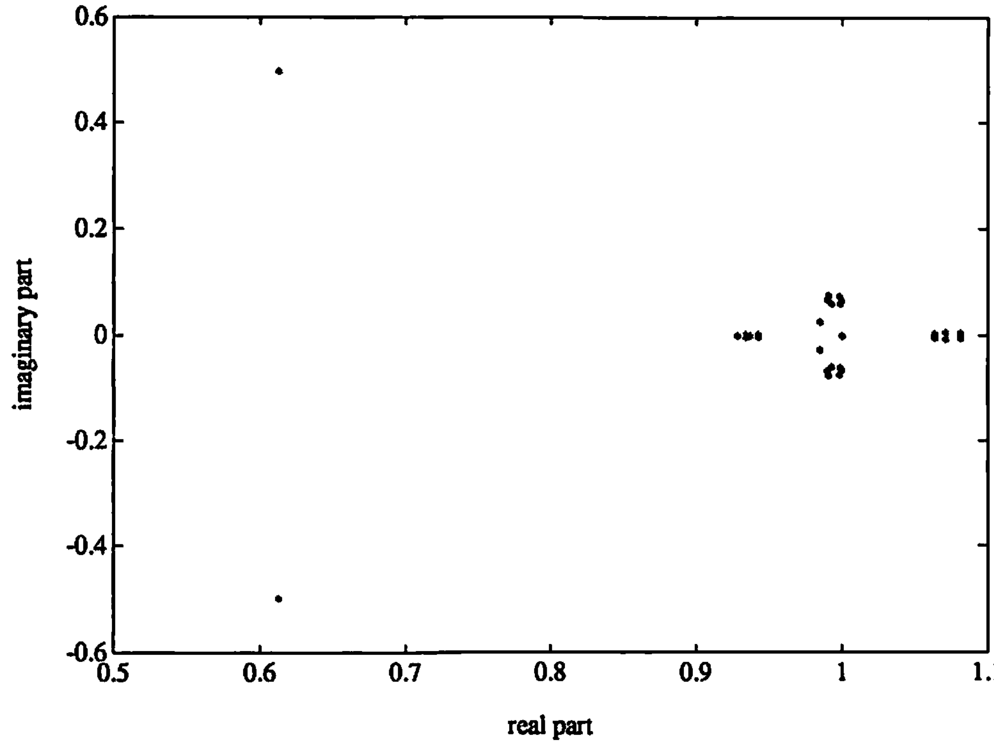
FIG. 1. *(a) The singular value distribution of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, and (b) the eigenvalue distribution of $K_N^{-1}T_N$ for Test Problem 1.*


imaginary part
real part
4
3
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5
6
7

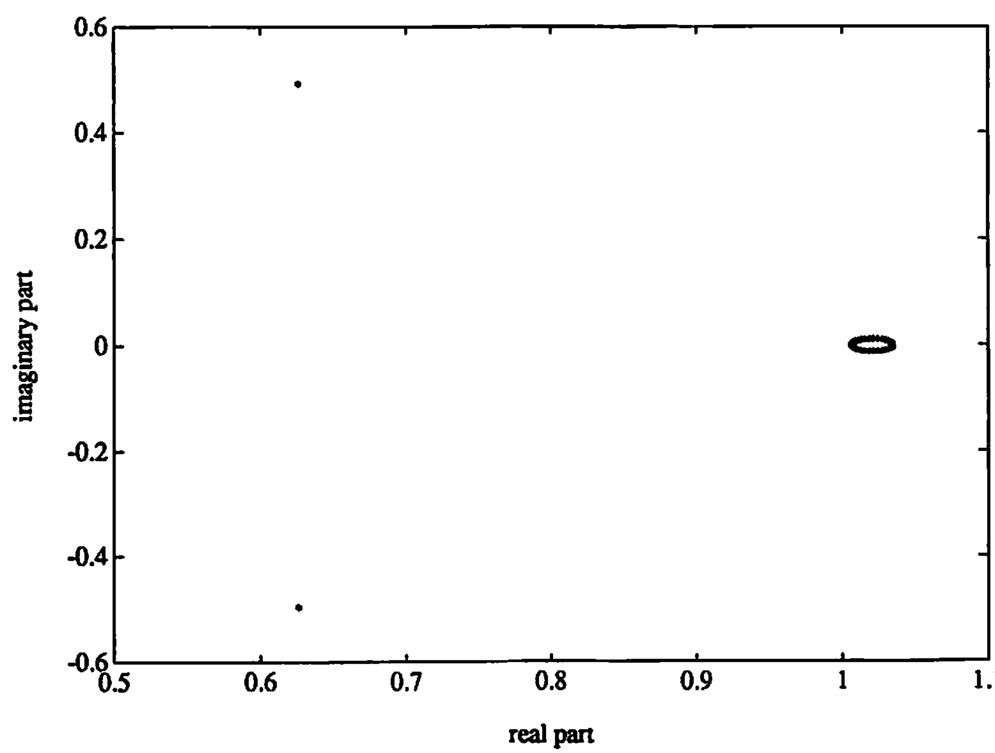
(a)


imaginary part
real part
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3

(b)



(c)



(d)

FIG. 2. *The eigenvalue distribution of (a)* T_N, *(b)* $\tilde{S}_N^{-1}T_N$, *(c)* $S_N^{-1}T_N$ *and (d)* $K_N^{-1}T_N$ *for Test Problem 2.*

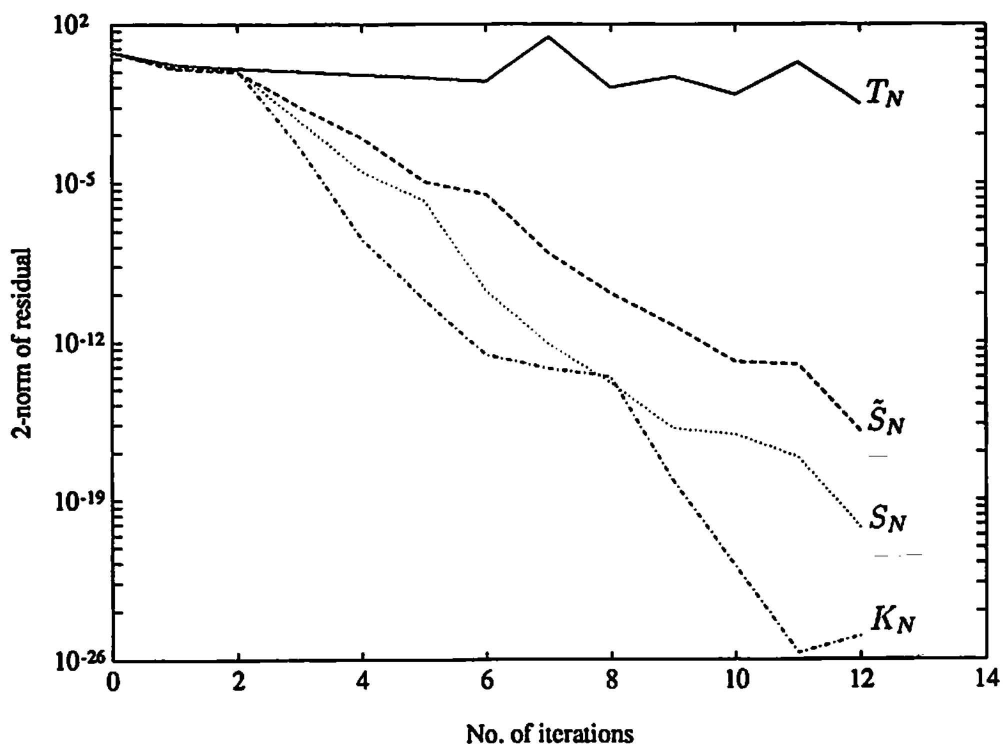
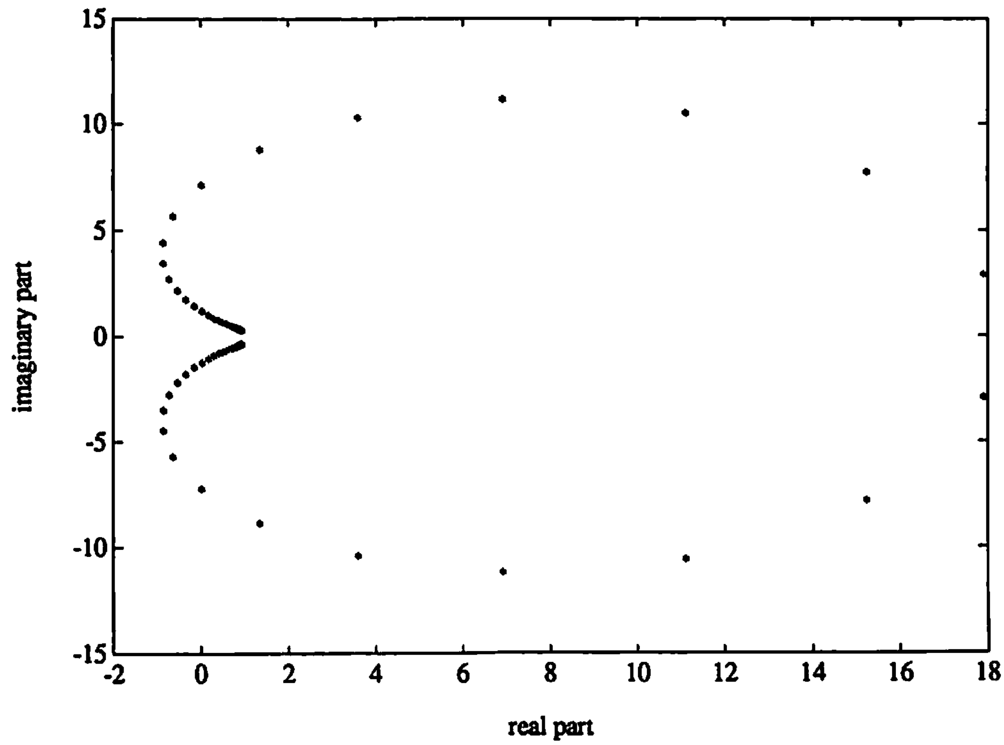
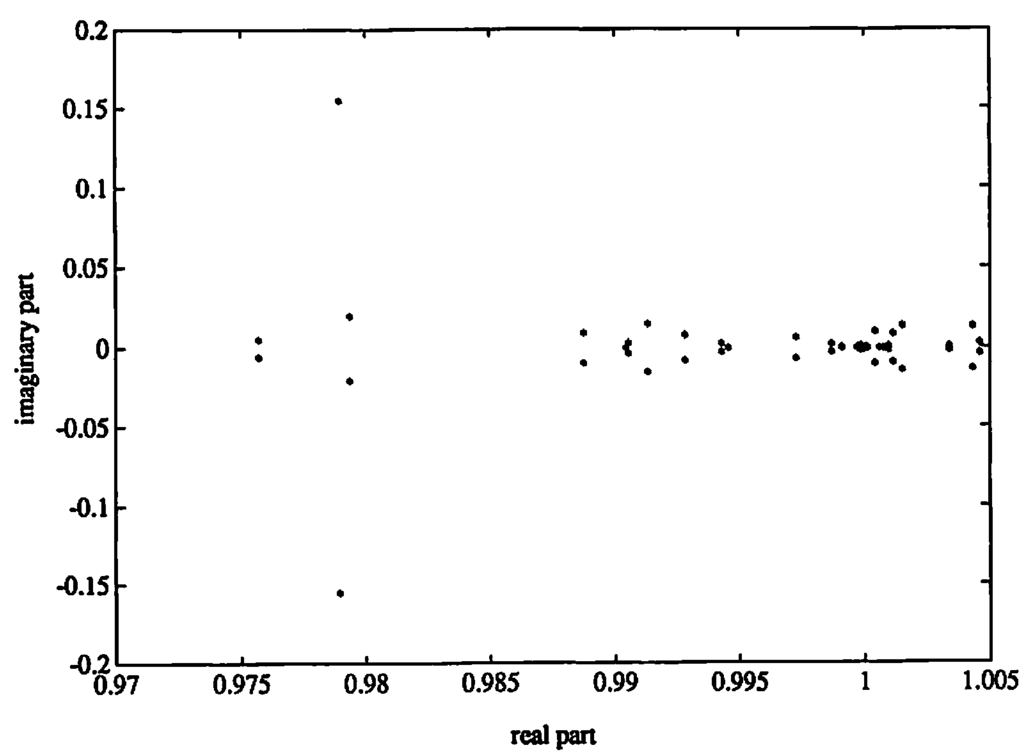


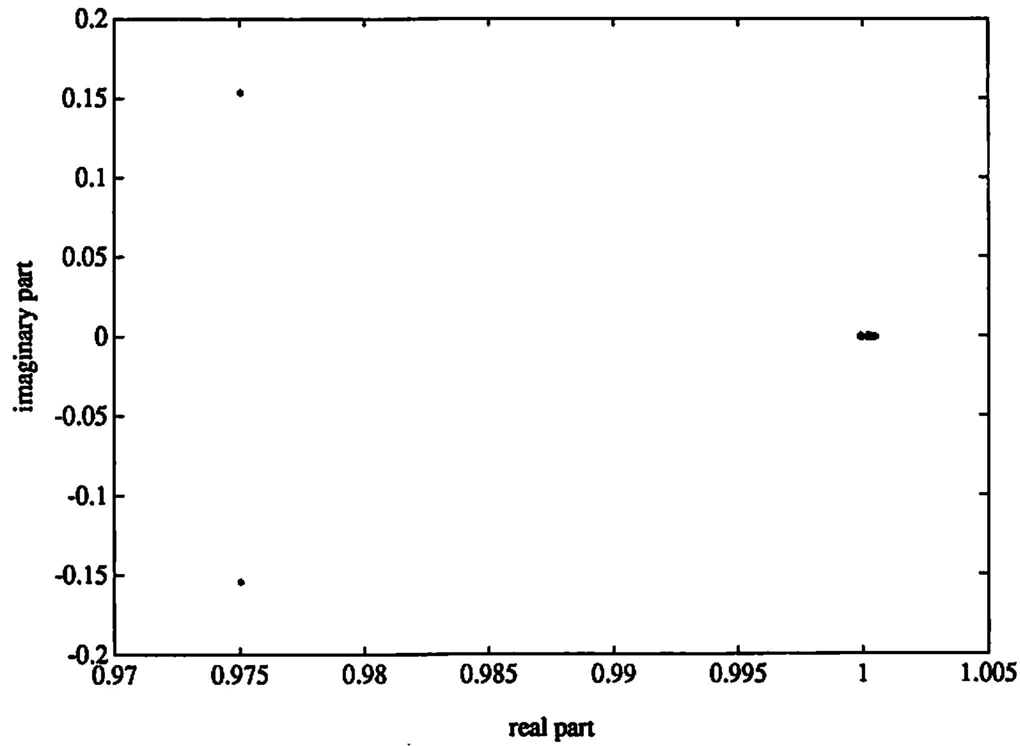
FIG. 3. *The convergence history of the CGS method for Test Problem 2.*


15
10
5
0
-5
-10
-15
imaginary part
-2
0
2
4
6
8
10
12
14
16
18
real part

(a)


0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
imaginary part
0.97
0.975
0.98
0.985
0.99
0.995
1
1.005
real part

(b)



(c)

FIG. 4. *The eigenvalue distribution of (a) T_N, (b) $S_N^{-1}T_N$ and (c) $K_N^{-1}T_N$ for Test Problem 3.*

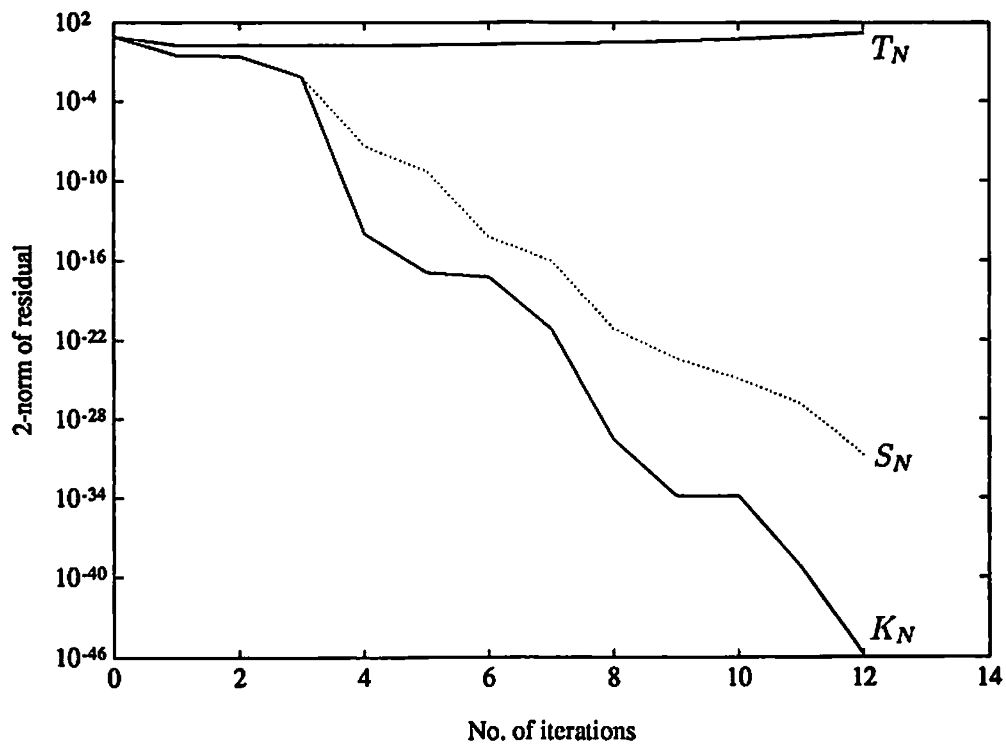
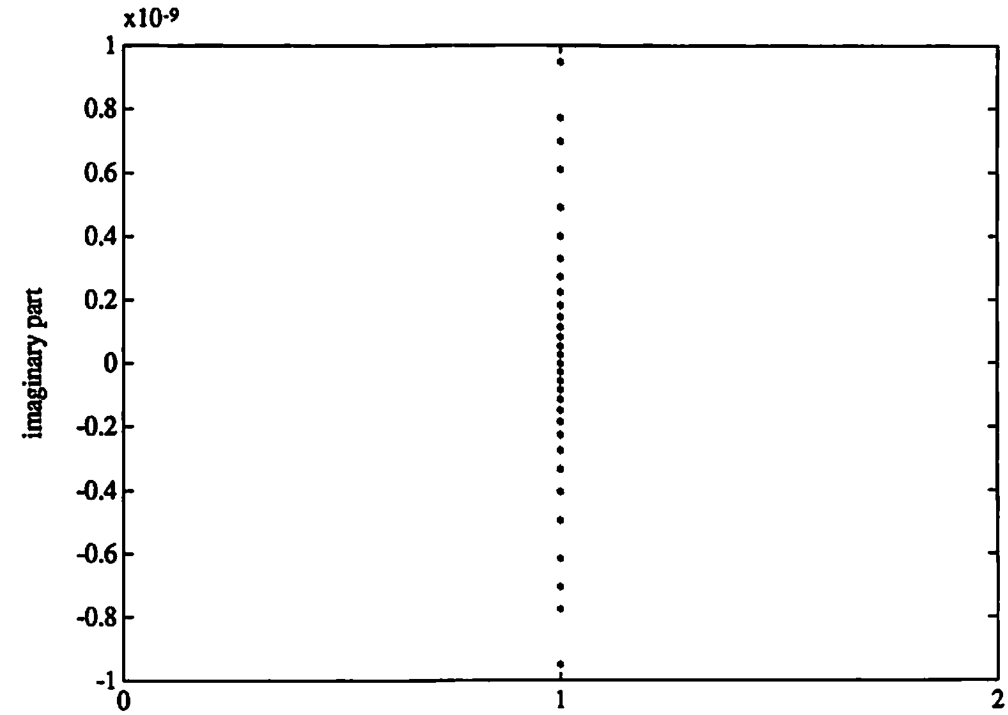
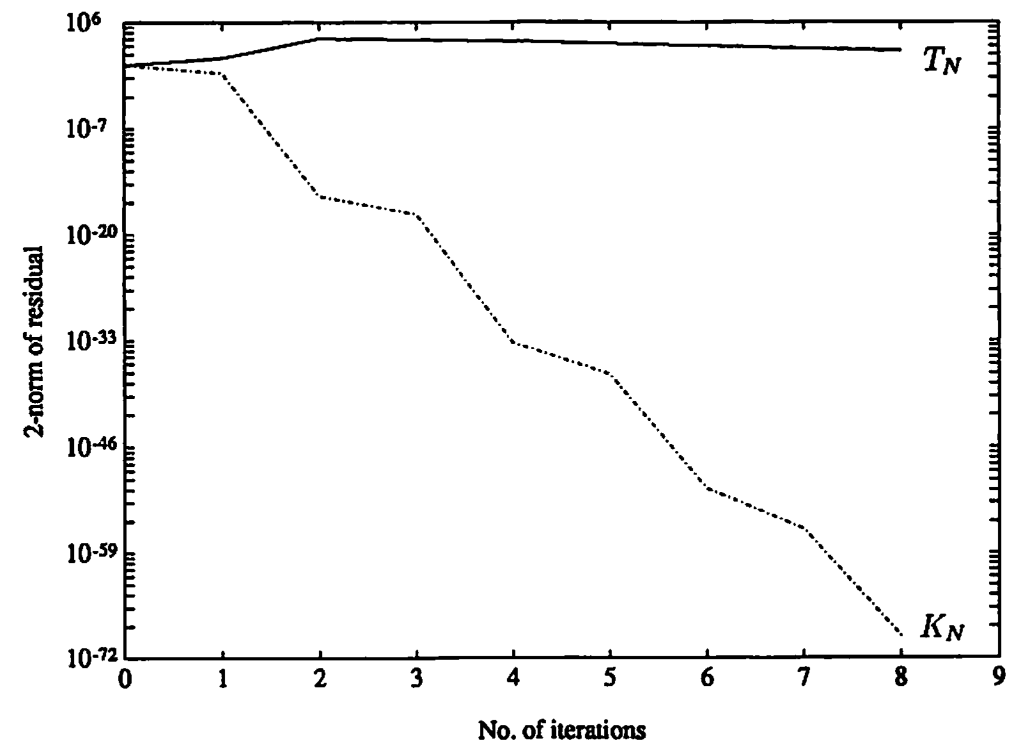


FIG. 5. *The convergence history of the CGS method for Test Problem 3.*



(a)



(b)

FIG. 6. *(a) The eigenvalue distribution of $K_N^{-1}T_N$, and (b) the convergence history of the CGS method for Test Problem 4.*