MATHEMATICS OF COMPUTATION
Volume 65, Number 213
January 1996, Pages 137–150

WHICH CIRCULANT PRECONDITIONER IS BETTER?

V. V. STRELA AND E. E. TYRTYSHNIKOV

ABSTRACT. The eigenvalue clustering of matrices $S_n^{-1}A_n$ and $C_n^{-1}A_n$ is experimentally studied, where A_n, S_n and C_n respectively are Toeplitz matrices, Strang, and optimal circulant preconditioners generated by the Fourier expansion of a function $f(x)$. Some illustrations are given to show how the clustering depends on the smoothness of $f(x)$ and which preconditioner is preferable. An original technique for experimental exploration of the clustering rate is presented. This technique is based on the bisection idea and on the Toeplitz decomposition of a three-matrix product CAC, where A is a Toeplitz matrix and C is a circulant. In particular, it is proved that the Toeplitz (displacement) rank of CAC is not greater than 4, provided that C and A are symmetric.

1. INTRODUCTION

While solving systems of linear equations by an iterative method, there inevitably appears the problem of building up a preconditioner. It seems to be natural to choose as a preconditioner a matrix which, on the one hand, approximates the matrix of the system and, on the other hand, could be easily inverted.

An example of easily invertible matrices is the class of circulant matrices. $C_n = [c_{ij}]_{ij=0}^{n-1}$ is a circulant if $c_{ij} = c_{(i-j) \bmod n}$. By means of the fast Fourier transform (FFT), both C_n and C_n^{-1} can multiply a vector in $O(n \log n)$ arithmetic operations, which is rather fast.

The requirement of approximation of a matrix A can be expressed in different ways. For example, one could choose C minimizing the Frobenius norm $\| A - C \|_F$ over some set of matrices.

We consider the case of a linear algebraic system of equations with a Toeplitz matrix A. $A_n = [a_{ij}]_{ij=0}^{n-1}$ is Toeplitz if $a_{ij} = a_{i-j,0}$. For such matrices circulant preconditioners seem to be especially efficient.

There are several types of circulant preconditioners for systems with Toeplitz matrices. The Strang preconditioner S_n was the first one proposed [8]. The central $[n/2]$ diagonals of this preconditioner coincide with the central $[n/2]$ diagonals of the Toeplitz matrix A_n. The other diagonals of S_n are defined by the fact that it is a circulant.

In [5] T. Chan proposed an optimal circulant preconditioner C_n. It is constructed to minimize the functional $\| A_n - C_n \|_F$ over the set of circulant matrices. The explicit expression of C_n's elements in the case of a Toeplitz A_n is given in [5].

Received by the editor December 28, 1993 and, in revised form, August 3, 1994.

1991 *Mathematics Subject Classification.* Primary 15A18, 15A57, 65F15; Secondary 42A16, 15A23.

Key words and phrases. Preconditioning, eigenvalue clustering, circulants, Toeplitz matrices, Fourier series.

In [13] it is shown that optimal circulants can be naturally referred to as Cesàro circulants.

E. Tyrtyshnikov proposed in [12] a superoptimal circulant preconditioner T_n. It minimizes $\| I - T_n^{-1} A_n \|_F$. Some properties of C_n and T_n were also explored in [12] and it was proved that they are symmetric and positive definite if A_n is. Analogous preconditioners for Toeplitz matrices were independently studied by M. Tismenetsky in [10].

Toeplitz matrices are often associated with Fourier series of periodic functions. The sequence of Toeplitz matrices $\{A_n\}$ is said to be generated by $f(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx}$ if the first column of A_n is $(a_0, a_1, \ldots, a_{n-1})^T$ and its first row is $(a_0, a_{-1}, \ldots, a_{-(n-1)})$. For the given sequence $\{A_n\}$ we can construct sequences $\{C_n\}$, $\{T_n\}$, $\{S_n\}$ of optimal, super-optimal and Strang preconditioners, respectively. In Figures 1.1–1.3 are shown spectra of A_n, $S_n^{-1}A_n$, $C_n^{-1}A_n$ and $T_n^{-1}A_n$ $(n = 32)$ for three different functions $f(x)$.

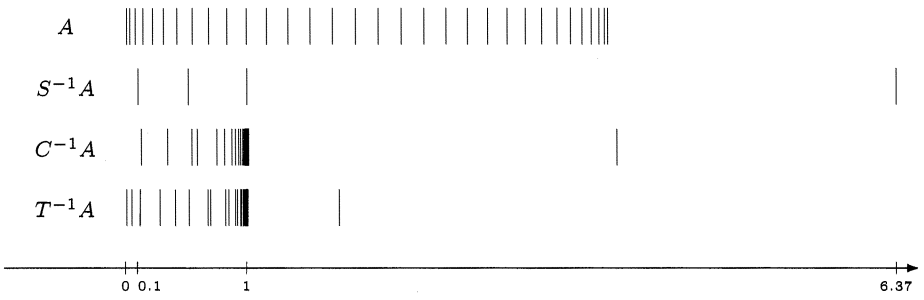

FIGURE 1.1. $f(x) = 2 - 2\cos x$

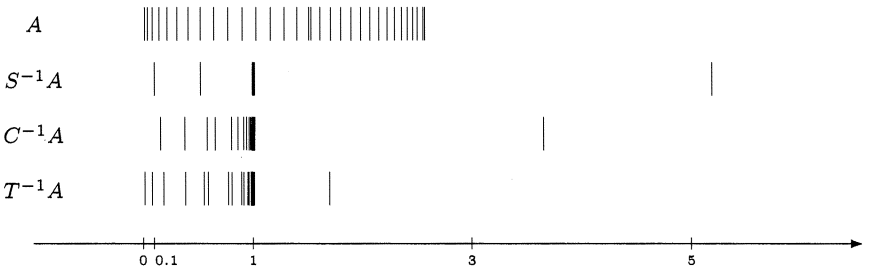

FIGURE 1.2. $f(x) = \begin{cases} x \sin x, & -\frac{\pi}{2} \le x \le \frac{\pi}{2}; \\ \frac{\pi}{2} - \cos x, & -\pi \le x < -\frac{\pi}{2};\ \frac{\pi}{2} < x \le \pi \end{cases}$

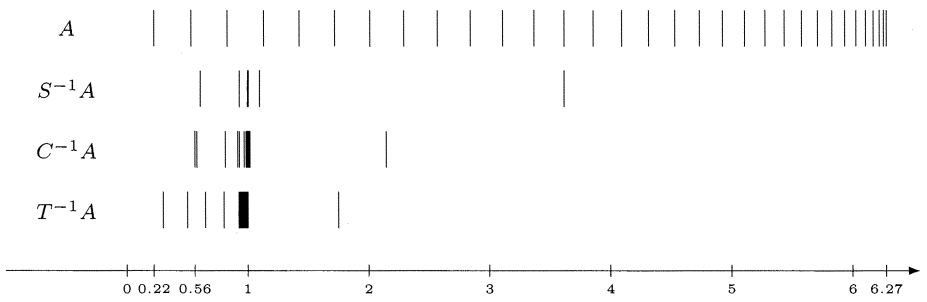

FIGURE 1.3. $f(x) = 2\pi \mid \sin \frac{x}{2} \mid$

One can see that the eigenvalues of all preconditioned matrices are clustered. $S_n^{-1}A_n$ always has the sharpest cluster while $T_n^{-1}A_n$ always has the widest one. The number of the eigenvalues cutside the cluster seems to depend on the smoothness of $f(x)$. And finally, in all three cases, $S_n^{-1}A_n$ has a condition number (ratio of the largest and smallest eigenvalue) larger than the condition number of $C_n^{-1}A_n$.

The exploration of the connection between spectral properties of $C_n^{-1}A_n$, $S_n^{-1}A_n$, $T_n^{-1}A_n$ and the properties of the generating function $f(x)$ is very important for practical application of circulant preconditioners, because the "better" the clustering rate, the "faster" the conjugate gradient method converges [1].

In [4, 2] it was proved that if $f(x) > 0$ belongs to the Wiener class, then the spectra of matrix sequences $S_n^{-1}A_n$, $C_n^{-1}A_n$ generated by $f(x)$ are clustered at 1. An analogous statement for sequence $T_n^{-1}A_n$ was proved in [3].

A unifying approach to the theorems on distribution and clustering for Toeplitz and circulant matrices is proposed in [13]. On the basis of this approach the theorems from [2] were generalized to the case of $f(x) \in L_2$.

Finally, the properties of $\gamma_n(\varepsilon)$, the number of eigenvalues which fall outside an ε-neighborhood of the clustering point, are studied in [14]. Only on the basis of this investigation does it become possible to decide which type of preconditioner is "better". The results of [14] give the impression that with increasing smoothness of $f(x)$, Strang's circulant is "better" than the optimal one (from the view point of spectrum clustering). However, we do not know how sharp the estimates are and what the constants are. So, it is useful to carry out a numerical investigation of the spectrum clustering.

Attempts to perform such experiments were already made in [9]. There were given the spectrum distributions of $S_n^{-1}A_n$, $C_n^{-1}A_n$, $T_n^{-1}A_n$ $(n = 32)$ for various functions $f(x)$. The large amount of arithmetic operations required did not allow to increase sufficiently the matrix order, which would be essential for an exploration of the asymptotics of clustering.

One relatively fast way to get some information about the spectrum of a large matrix is to apply to it the conjugate gradient iterations. However, this method allows only to investigate the extreme eigenvalues and tells almost nothing about the spectrum distribution. To acquire more information about the spectrum we need to increase considerably the number of iterations. This is impracticable in our case, so we choose another way.

On the basis of the bisection method we succeeded in designing a rather fast,

$O(n^2)$ algorithm, using the Toeplitz decomposition of matrices. This algorithm gave us the opportunity to study the behavior of $\gamma_n(\varepsilon)$ in the case of large n.

Thus, this paper is devoted to an experimental investigation and comparison of the clustering properties of $S_n^{-1}A_n$ and $C_n^{-1}A_n$. (The results of [9] showed T_n to be not as efficient as was expected, both for clusterization and the number of iterations in the conjugate gradient method. So we decided to concentrate only on C_n and S_n.) The paper is organized as follows. In §2 the main definitions and theoretical results to be experimentally studied, are given. Section 3 is devoted to the description of the tools. In particular, we prove that the Toeplitz decomposition of CAC (C is a symmetric circulant, A is a symmetric Toeplitz matrix) contains no more than 4 terms. In §4 the experimental results are discussed and some conclusions are drawn.

2. Definitions and theorems

Suppose we are given a function $f(x)$ and its Fourier coefficients a_k,

$$f(x) = \sum_{k=-\infty}^{\infty} a_k \exp(ikx).$$

We put in correspondence to $f(x)$ three families of matrices A_n, C_n, S_n:

$A_n = [a_{i-j}]_{ij=0}^{n-1}$ is a Toeplitz matrix,
$C_n = [c_{(i-j) \bmod n}]_{ij=0}^{n-1}$ is a circulant matrix,
$c_k = \frac{1}{n}((n-k)a_k + ka_{-(n-k)})$,
$S_n = [s_{(i-j) \bmod n}]_{ij=0}^{n-1}$ is a circulant matrix,

$$s_k = \begin{cases} a_k, & 0 \le k < \frac{n-1}{2}, \\ a_{-(n-k)}, & \frac{n-1}{2} < k \le n-1, \\ 0, & k = \frac{n-2}{2}. \end{cases}$$

It is conventional to call C_n an optimal (Cesàro) circulant, and S_n a simple circulant (it differs from the Strang construction only by the diagonal with number $[n/2]$).

Below we consider only real 2π-periodic functions $f(x)$. It is easy to see that these conditions on $f(x)$ make the matrices A_n, C_n, S_n Hermitian.

Let $\gamma_n(\varepsilon)$ be the number of $\lambda_k^{(n)} \notin (\mu - \varepsilon, \mu + \varepsilon)$. Then the point μ is called a cluster of the sequence $\{\lambda_k^{(n)}\}_{k=1}^n$ if

$$\lim_{n\to\infty} \frac{\gamma_n(\varepsilon)}{n} = 0.$$

A cluster is called proper if $\gamma_n(\varepsilon) \le c(\varepsilon)$, where $c(\varepsilon)$ does not depend on n.

The first results on the eigenvalue clustering of $C_n^{-1}A_n$ and $S_n^{-1}A_n$ are given in [4, 2]:

Theorem 2.1 ([2]). *If $f(x) \ge m > 0$ belongs to the Wiener class ($\sum_{k=-\infty}^{\infty} | a_k | < +\infty$), then optimal circulants C_n, and for sufficiently large n simple circulants S_n, are positive definite and the point 1 is a proper cluster for the eigenvalues of $C_n^{-1}A_n$ and $S_n^{-1}A_n$.*

The unifying approach, put forward in [13], allowed to generalize Theorem 2.1 to the case of $f(x) \in L_2$:

Theorem 2.2 ([13]). *Let $f(x) \in L_2$ and $f(x) \geq m > 0$. Then optimal circulants C_n are positive definite, and the eigenvalues of the matrices $C_n^{-1}A_n$ are clustered at 1. If $\lambda(S_n) \geq \delta > 0$, then the same is valid for $S_n^{-1}A_n$.*

In [14] it was shown that the condition $f(x) \geq m > 0$ is not essential, and Theorem 2.2 was generalized to the case where $f(x)$ is a slightly vanishing function. A 2π-periodic Lebesgue–integrable function $f(x)$ is called slightly vanishing if

$$\lim_{\varepsilon \to +0} \int_{-\pi}^{\pi} \varphi_\varepsilon(|f(x)|)dx = 0,$$

where

$$\varphi_\varepsilon(x) = \begin{cases} 1, & 0 \leq x \leq \varepsilon, \\ 0, & x < 0,\ x > \varepsilon. \end{cases}$$

Theorem 2.3 ([14]). *Let $f(x)$ be a nonnegative slightly vanishing function from L_2, and A_n, C_n, S_n are the associated Toeplitz matrices, optimal and simple circulants, respectively. Assume also that the matrices C_n, S_n are positive definite. Then the eigenvalues of $C_n^{-1}A_n$ and $S_n^{-1}A_n$ are real and have a cluster at 1.*

The drawback of Theorem 2.2 is the condition of positive definiteness. It can be removed by the construction of improved optimal and simple circulants $\hat{C}_n$ and $\tilde{S}_n$. The eigenvalues of these circulants coincide with the eigenvalues of C_n and S_n except for the nonpositive ones, which are changed to $\delta > 0$. In [14] it is shown that Theorem 2.3 still holds for $\hat{C}_n$ and $\tilde{S}_n$.

The theorems given above say almost nothing about the properties of the function $\gamma_n(\varepsilon)$, so they do not allow a comparison of C_n and S_n. The results of [14] provide us with such an opportunity. But before citing these theorems, we define the class of functions for which they are valid.

Suppose a 2π-periodic function $f(x)$ is such that its mth derivative $f^{(m)}(x)$ is piecewise continuous and has a bounded derivative on each continuity interval. Let K_m denote the set of all such functions.

Further, assume that there is a finite number of points $x_j \in [-\pi, \pi]$, $j = 1, \ldots, t$, such that

$$f(x_j) = 0, \qquad j = 1, \ldots, t. \tag{2.1}$$

Assume that at every x_j both left and right derivatives of some order are distinct from zero. Denote by $p_j^{\pm}$ the orders of the first such derivatives, that is,

$$\begin{aligned} &f^{(1)}(x_j + 0) = \cdots = f^{(p_j^+ - 1)}(x_j + 0) = 0, f^{(p_j^+)}(x_j + 0) \neq 0, \\ &f^{(1)}(x_j - 0) = \cdots = f^{(p_j^- - 1)}(x_j - 0) = 0, f^{(p_j^-)}(x_j - 0) \neq 0, \end{aligned} \tag{2.2}$$

and set

$$p = \max\{p_j^{\pm} : j = 1, \ldots, t\}. \tag{2.3}$$

Let $K_m^{(p)}$ denote the set of those $f(x) \in K_m$ which are characterized by the relationships (2.1)–(2.3).

Theorem 2.4 ([14]). *Assume that $f(x) \in K_m^{(p)}$ is nonnegative. Then the eigenvalues of $\tilde{S}_n^{-1}A_n$ are real and clustered at 1 so that*

$$\gamma_n^S(\varepsilon) = O(n^{\frac{p}{p+m}}). \tag{2.4}$$

Theorem 2.5 ([14]). *Assume that $f(x) \in K_m^{(p)}$ is nonnegative. Then the eigenvalues of $\hat{C}_n^{-1}A_n$ are real and clustered at 1 so that*

$$\gamma_n^C(\varepsilon) = \begin{cases} O(n^{\frac{p}{p+1}}), & m > 1, \\ O(n^{\frac{p}{p+1}} \ln n), & m = 1. \end{cases} \tag{2.5}$$

If $f(x)$ can be expressed in the form

$$f(x) = \sum_{k=-\nu}^{\nu} a_k e^{ikx}, \tag{2.6}$$

then the following theorem holds.

Theorem 2.6 ([14]). *Assume that $f(x)$ is a nonnegative function of the form* (2.6), *and not everywhere zero. Then the eigenvalues of $\tilde{S}_n^{-1}A_n$ are clustered at 1 so that*

$$\gamma_n^S(\varepsilon) = O(1) \tag{2.7}$$

(*the proper cluster*).

From the above results it follows that if the function is smooth, the simple circulant S_n should be "better" (from the viewpoint of spectrum clustering). But we do not know how sharp the estimates (2.4), (2.5) are and what the constants in them are. It appears that this question should be clarified experimentally. This is the aim of §4 of our paper.

3. Tools of research

In this section we describe a very efficient method for solving the following problem. Given a symmetric Toeplitz matrix A and a symmetric positive definite circulant matrix C, both of order n, find the number of eigenvalues of $P \equiv CAC$ lying in an ε-neighborhood of 1.

As we want to deal with large n, we resist the use of the classical methods for dense matrices. We try to apply to our problem the idea underlying the bisection method. In other words, we will rely upon the following well-known fact.

Lemma 3.1. *Suppose a symmetric matrix P of order n has nonsingular leading submatrices P_k of order k for all $k = 1, \dots, n$. Then the number of negative eigenvalues of $P \equiv P_n$ coincides with the number of sign changes in the sequence*

$$1, \det P_1, \dots, \det P_n.$$

The bordering method certainly is one of the natural methods for the leading minors computation. However, we would like to use the specific structure of P. As

was shown in [6], the bordering method can be successfully applied to certain types of matrices, for example matrices which are a sum of products of triangular Toeplitz matrices. The algorithm from [6] for such matrices with q terms can be easily modified ([11]) to an algorithm requiring $\frac{5q-4}{2}n^2$ multiplications and additions. In essence, it is the algorithm already proposed in [7] but within the confines of a block method.

However, we will use a more efficient algorithm ([15]) for computation of the inertia of matrices defined by their Toeplitz decomposition. This algorithm needs only $(q-1)n^2$ multiplications and additions. It also has a parallel structure.

So, the main thing we have to do is to find a fast way of performing the Toeplitz decomposition of P. We call

$$T = \sum_{s=1}^{q} L(\alpha_s)L^T(\beta_s) \tag{3.1}$$

the Toeplitz decomposition of the matrix T, provided that

$$L(\alpha_s)L^T(\beta_s) = \begin{bmatrix} \alpha_{s,0} & & & 0 \\ \alpha_{s,1} & \alpha_{s,0} & & \\ \vdots & \vdots & \ddots & \\ \alpha_{s,n-1} & \alpha_{s,n-2} & \cdots & \alpha_{s,0} \end{bmatrix} \begin{bmatrix} \beta_{s,0} & \beta_{s,1} & \cdots & \beta_{s,n-1} \\ & \beta_{s,0} & \cdots & \beta_{s,n-2} \\ & & \ddots & \vdots \\ 0 & & & \beta_{s,0} \end{bmatrix}.$$

It is easy to see that

$$\sum_{s=1}^{q} \begin{bmatrix} \alpha_{s,0} \\ \vdots \\ \alpha_{s,n-1} \end{bmatrix} [\beta_{s,0}, \dots, \beta_{s,n-1}] = \Delta T = [\Delta t_{ij}]_{ij=0}^{n-1}, \tag{3.2}$$

where

$$T = [t_{ij}]_{ij=0}^{n-1},$$

$$\Delta t_{ij} = t_{ij} - t_{i-1,j-1}, \qquad 0 < i, j \le n-1, \tag{3.3}$$

$$\Delta t_{i0} = t_{i0}, \quad \Delta t_{0i} = t_{0i}, \qquad i = 0, \dots, n-1.$$

The smallest q is called the Toeplitz rank (displacement rank in the terminology of [7]) of T.

Toeplitz ranks were introduced in [7]. In [11] it was proved that the Toeplitz rank of a product of two matrices is not greater than the sum of the factors' ranks increased by 1. It is easy to see that the Toeplitz rank of a Toeplitz matrix is 2. The matrix P contains three Toeplitz factors, so the Toeplitz rank of P is not greater than 8. However, we show that it does not exceed 4.

Theorem 3.1. *The Toeplitz rank of $P = CAC$, where C is a symmetric circulant and A is a symmetric Toeplitz matrix, is not greater than* 4.

Proof. We have

$$p_{ij} = \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} c_{i-l}a_{l-k}c_{k-j}$$
$$= c_{-j}\sum_{l=0}^{n-1} c_{i-l}a_l + c_i\sum_{k=1}^{n-1} a_{-k}c_{k-j} + \sum_{k=1}^{n-1}\sum_{l=1}^{n-1} c_{i-l}a_{l-k}c_{k-j}, \qquad 0 \le i,j \le n-1;$$

$$p_{i-1,j-1} = \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} c_{i-1-l}a_{l-k}c_{k-j+1} = \sum_{k=1}^{n}\sum_{l=1}^{n} c_{i-l}a_{l-k}c_{k-j}$$
$$= c_{n-j}\sum_{l=1}^{n} c_{i-l}a_{l-n} + c_{i-n}\sum_{k=1}^{n-1} a_{n-k}c_{k-j}$$
$$+ \sum_{k=1}^{n-1}\sum_{l=1}^{n-1} c_{i-l}a_{l-k}c_{k-j}, \qquad 1 \le i,j \le n-1;$$

$$\Delta p_{ij} = p_{ij} - p_{i-1,j-1}$$
$$= c_i\sum_{k=1}^{n-1} a_{-k}c_{k-j} - c_{i-n}\sum_{k=1}^{n-1} a_{n-k}c_{k-j}$$
$$+ c_{-j}\sum_{l=0}^{n-1} c_{i-l}a_l - c_{n-j}\sum_{l=1}^{n} c_{i-l}a_{l-n}, \qquad 1 \le i,j \le n-1.$$

Here we used that C, A are Toeplitz matrices, and so $a_{ij} = a_{i-j}$, $c_{ij} = c_{i-j}$.

Since A, C are symmetric and C is a circulant, we know that

$$a_k = a_{-k},\ a_n = a_0, \quad c_k = c_{-k}, \quad c_n = c_0, \quad c_i = c_{n-i}.$$

So, in our case (for $1 \le i,j \le n-1$),

(3.4)

$$\Delta p_{ij} = c_i\sum_{k=0}^{n-1} a_k c_{k-j} - c_i\sum_{k=0}^{n-1} a_{n-k}c_{k-j} + c_j\sum_{k=0}^{n-1} c_{i-k}a_k - c_j\sum_{k=0}^{n-1} c_{i-k}a_{k-n}$$
$$= c_i\sum_{k=0}^{n-1} c_{j-k}(a_k - a_{n-k}) + c_j\sum_{k=0}^{n-1} c_{i-k}(a_k - a_{n-k}) = c_i d_j + d_i c_j,$$

where

(3.5)

$$d_i = \sum_{k=0}^{n-1} c_{i-k}(a_k - a_{n-k}),$$

$$\Delta p_{i0} = \Delta p_{0i} = p_{0i} = p_{i0} = \sum_{k,l=0}^{n-1} c_{i-l}a_{l-k}c_k, \qquad i = 0,\dots,n-1.$$

In accordance with (3.1)–(3.2) and (3.4)–(3.5), we get the Toeplitz decomposition of P with $q = 4$:

(3.6)

$$P = \begin{bmatrix} \frac{1}{2}p_{00} & & & 0 \\ p_{1,0} & \frac{1}{2}p_{00} & & \\ \vdots & \vdots & \ddots & \\ p_{n-1,0} & p_{n-2,0} & \cdots & \frac{1}{2}p_{00} \end{bmatrix} + \begin{bmatrix} \frac{1}{2}p_{00} & p_{1,0} & \cdots & p_{n-1,0} \\ & \frac{1}{2}p_{00} & \cdots & p_{n-2,0} \\ & & \ddots & \vdots \\ 0 & & & \frac{1}{2}p_{00} \end{bmatrix}$$
$$+ \begin{bmatrix} 0 & & & \\ c_1 & 0 & 0 & \\ \vdots & & \ddots & \\ c_{n-1} & \cdots & c_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & d_1 & \cdots & d_{n-1} \\ & 0 & & \vdots \\ & 0 & \ddots & d_1 \\ & & & 0 \end{bmatrix}$$
$$+ \begin{bmatrix} 0 & & & \\ d_1 & 0 & 0 & \\ \vdots & & \ddots & \\ d_{n-1} & \cdots & d_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & c_1 & \cdots & c_{n-1} \\ & 0 & & \vdots \\ & 0 & \ddots & c_1 \\ & & & 0 \end{bmatrix}.$$

This completes the proof. □

To use the algorithm from [15], the Toeplitz decomposition should be expressed in the following form:

$$T = \sum_{s=1}^{q} k_s L(\alpha_s) L^T(\alpha_s), \tag{3.7}$$

i.e.,

$$\Delta T = \sum_{s=1}^{q} k_s \begin{bmatrix} \alpha_{s,0} \\ \vdots \\ \alpha_{s,n-1} \end{bmatrix} [\alpha_{s,0}, \ldots, \alpha_{s,n-1}]. \tag{3.8}$$

The relations (3.4)–(3.6) can be transformed into the form (3.7)–(3.8) by using the following obvious result.

Lemma 3.2. *If*

$$u = \begin{bmatrix} u_0 \\ \vdots \\ u_{n-1} \end{bmatrix}, \quad v = \begin{bmatrix} v_0 \\ \vdots \\ v_{n-1} \end{bmatrix},$$

then

$$uv^T + vu^T = \frac{1}{2}(u+v)(u+v)^T - \frac{1}{2}(u-v)(u-v)^T.$$

Note that if we set $\hat{d}_0 = 0$ in the vector $\hat{d} = [\hat{d}_0, \hat{d}_1, \ldots, \hat{d}_{n-1}]^T$, which is a product of the matrix C by the vector $\hat{a} = [0, a_1 - a_{n-1}, \ldots, a_{n-1} - a_1]^T$, we get the vector $d = [0, d_1, \ldots, d_{n-1}]^T$ needed in (3.6). As C is a circulant, d can be computed in $O(n \log n)$ arithmetic operations. The vector $p = [p_{00}, \ldots, p_{n-1,0}]^T$ can be computed at equal cost (as p is the first column of CAC).

In our concrete case, P is of the following form:

$$P = C^{1/2} A C^{1/2},$$

where C is a circulant preconditioner and A is the Toeplitz matrix of the system. The spectrum of P is the same as the spectrum of $C^{-1}A$, and so we can compute $\gamma_n(\varepsilon)$ applying Lemma 3.1 and our algorithms to the matrices $P - (1+\varepsilon)I$, $P - (1-\varepsilon)I$.

The computation of the first column of $C^{1/2}$ can also be performed in $O(n \log n)$ arithmetic operations ([11]). Thus, computing the number of $C^{-1}A$'s eigenvalues which lie in the interval $(1-\varepsilon, 1+\varepsilon)$ costs $O(n^2)$ arithmetic operations, which allows us to perform experiments with matrices of rather large order.

All experiments presented in §4 were made on an IBM PC/AT-386. All programs were written in FORTRAN 77.

4. Results of experiments

In this section we experimentally study the function $\gamma_n(\varepsilon)$ — the number of the eigenvalues of the matrices $\hat{C}_n^{-1}A_n$ and $\tilde{S}_n^{-1}A_n$ which fall outside the interval $(1-\varepsilon, 1+\varepsilon)$. In all experiments we assume $\varepsilon = 0.1$.

We begin with an illustration of Theorem 2.1. Assume

$$f(x) = 0.1 + 2\pi |\sin \frac{x}{2}|.$$

It is easy to see that in this case $f(x)$ belongs to the Wiener class. Besides, $f(x) \geq 0.1 > 0$, i.e., $f(x)$ obeys the conditions of Theorem 2.1. This means that the spectra of $C_n^{-1}A_n$ and $S_n^{-1}A_n$ must have a proper cluster at 1 (here, $\hat{C}_n = C_n$, $\tilde{S}_n = S_n$ are positive definite matrices as $f(x) \geq m > 0$ [4, 2]). This is confirmed by the experimental results shown in Table 4.1. (Everywhere below, γ_n^C refers to $\hat{C}_n^{-1}A_n$ and γ_n^S to $\tilde{S}_n^{-1}A_n$.)

The dependence of $\gamma_n(\varepsilon)$ on n should appear if $f(x)$ is equal to 0 even at one point. The case of such functions is studied in Theorems 2.4–2.6. The next four examples are devoted to these theorems.

To begin with, let

$$f(x) = 2\pi |\sin \frac{x}{2}|.$$

This function is continuous and vanishes at $x = 2\pi k$, $k = \ldots, -1, 0, 1, \ldots$. However, even the first derivative of $f(x)$ is discontinuous and not equal to 0 at these points. So, in this case, $p = 1$, $m = 1$, and according to (2.4), (2.5) we have

$$\gamma_n^S(\varepsilon) = O(n^{1/2}), \qquad \gamma_n^C(\varepsilon) = O(n^{1/2} \ln n). \tag{4.1}$$

Comparing (4.1) with the experimental results, shown in Table 4.2, we can see that estimates (4.1) do not have a large margin and correctly reflect the behavior of $\gamma_n(\varepsilon)$.

Table 4.1

n	32	64	128	256	512	1024	2048	4096
γ_n^C	4	5	6	5	5	5	5	5
γ_n^S	2	3	4	5	5	5	5	5

TABLE 4.2

n	32	64	128	256	512	1024	2048	4096
γ_n^C	4	6	6	9	10	11	12	14
γ_n^S	2	4	4	4	4	6	7	8

TABLE 4.3

n	32	64	128	256	512	1024	2048	4096
γ_n^C	11	15	20	26	38	53	74	103
γ_n^S	4	4	4	5	6	9	12	17

TABLE 4.4

n	32	64	128	256	512	1024	2048	4096
γ_n^C	8	11	16	21	30	42	59	84
γ_n^S	4	4	4	5	5	6	7	8

We now consider a function which equals zero with its first derivative at a point of $[-\pi, \pi]$. One such function is

$$f(x) = x^2, \qquad f(0) = f'(0) = 0.$$

Since we consider only 2π-periodic functions and $f'(\pi) \neq f'(-\pi)$, $f'(x)$ is discontinuous. So $p = 2$, $m = 1$, and we get

$$\gamma_n^S(\varepsilon) = O(n^{2/3}), \qquad \gamma_n^C(\varepsilon) = O(n^{2/3} \ln n). \tag{4.2}$$

The results of the experiments (Table 4.3) again show that the estimates (2.4), (2.5) are rather adequate, though not perfectly sharp.

We now increase smoothness of the generating function and consider

$$f(x) = \begin{cases} x \sin x, & -\pi/2 \le x \le \pi/2, \\ \pi/2 - \cos x, & -\pi \le x < -\pi/2,\ \pi/2 < x \le \pi, \end{cases}$$

$$f(0) = f'(0) = 0.$$

Now, $f'(x)$ is continuous on the whole real axis, however $f''(x)$ is not, i.e., $p = 2$, $m = 2$, and

$$\gamma_n^S(\varepsilon) = O(n^{1/2}), \qquad \gamma_n^C(\varepsilon) = O(n^{2/3}). \tag{4.3}$$

According to (2.4), (2.5), an increase of the smoothness should entail improvement of the clustering. This is clearly observed for S and to smaller degree for C (Table 4.4).

The next function we consider is

$$f(x) = 2 - 2\cos x, \qquad p = 2,\ m = \infty.$$

TABLE 4.5

n	32	64	128	256	512	1024	2048	4096
γ_n^C	9	12	16	23	32	44	61	85
γ_n^S	3	3	3	4	4	4	4	4

Here, $f(x)$ has a finite Fourier series ($a_0 = 2$, $a_1 = -1$, $a_k = 0$, $k = 2, \dots, n-1$). This means that Theorem 2.6 can be applied to this function, and

$$\gamma_n^S(\varepsilon) = O(1). \tag{4.4}$$

According to Theorem 2.5, we have for C

$$\gamma_n^C(\varepsilon) = O(n^{2/3}). \tag{4.5}$$

(Note that in this case S_n is singular for all n and so $S_n \neq \tilde{S}_n$.) Looking at the results of the experiments shown in Table 4.5, we see that the estimate (4.4) is almost ideally true. As to the estimate (4.5), it apparently could be somewhat improved.

We note that in contrast to $\tilde{S}_n^{-1} A_n$, smoothness of the generating function only weakly influences the eigenvalue clustering of $\hat{C}_n^{-1} A_n$. This is not surprising because S_n's eigenvalues are the values of the partial sums of $f(x)$'s Fourier series, chosen on the uniform mesh:

$$\lambda_k(S_n) = f_{[n/2]}\left(\frac{2\pi k}{n}\right), \qquad k = 0, 1, \dots, n-1,$$

$$f_m(x) = \sum_{k=-m}^{m} a_k \exp(ikx),$$

while C_n's eigenvalues are the values of $f(x)$'s Cesàro sums, chosen on the uniform mesh:

$$\lambda_k(C_n) = \sigma_n\left(\frac{2\pi k}{n}\right), \qquad k = 0, 1, \dots, n-1,$$

$$\sigma_n(x) = \frac{1}{n+1} \sum_{m=0}^{n} f_m(x)$$

([13]). From the theory of Fourier series it is known that Cesàro sums converge uniformly to the function without dependence on its smoothness, while the convergence of partial Fourier sums essentially depends on the function's smoothness. So, we can expect that in the case of a discontinuous generating function, the spectrum clustering rate of $\hat{C}_n^{-1} A_n$ is better than that of $\tilde{S}_n^{-1} A_n$. This expectation is confirmed by the following example (Table 4.6):

$$f(x) = \begin{cases} x \sin x, & -\pi/2 \le x \le \pi/2, \\ -\cos x, & -\pi \le x < -\pi/2,\ \pi/2 < x \le \pi, \end{cases}$$

$$f(0) = f'(0) = 0.$$

Here, $f(x)$ is discontinuous at the point $x = \pi/2$.

TABLE 4.6

n	32	64	128	256	512	1024	2048	4096
γ_n^C	12	18	25	34	45	63	86	120
γ_n^S	16	21	29	39	53	72	100	140

TABLE 4.7

n	32	64	128	256	512	1024	2048	4096
$\gamma_n^C(0.1)$	25	47	93	179	362	703	1397	2761
$\gamma_n^C(0.9)$	18	39	83	170	344	692	1385	2746
$\gamma_n^S(0.1)$	25	47	91	178	367	710	1408	2713
$\gamma_n^S(0.9)$	19	40	83	171	347	700	1389	2766

This experiment serves also as an illustration of Theorem 2 from [16]. This theorem states that in the case of piecewise continuous generating functions, γ_n^C grows at least as $O(\log n)$.

Consider now the slightly vanishing condition of the generating function in Theorem 2.3. The example of

$$f(x) = \begin{cases} 1, & -1 \le x \le 1, \\ 0, & -\pi \le x < -1,\ 1 < x \le \pi, \end{cases}$$

shows that it is essential for the clustering. Actually, this function is equal to 0 on a segment, and so the condition

$$\lim_{\varepsilon \to +0} \int_{-\pi}^{\pi} \varphi_\varepsilon(|f(x)|)dx = 0,$$

is not satisfied. The results of the experiments (Table 4.7) show that most of the eigenvalues of $\hat{C}_n^{-1}A_n$ and $\tilde{S}_n^{-1}A_n$ lie near 0.

In conclusion, let us say a few words about which circulant is "better", simple (Strang) or optimal. If $f(x) \ge m > 0$, then no essential difference between $\hat{C}_n$ and $\tilde{S}_n$ is observed. If $f(x)$ is continuous, slightly vanishing and vanishes at least at one point, then Theorems 2.3–2.6 affirm that $\tilde{S}_n$ should behave "better" than $\hat{C}_n$ (from the viewpoint of clustering). This is confirmed experimentally. If $f(x)$ is discontinuous, we can expect $\hat{C}_n$ to be "better".

However, the eigenvalue clustering is not the only feature of preconditioning. The other important feature is the condition number of the preconditioned matrix. Experiments show that if $f(x) = 0$ somewhere in $[-\pi, \pi]$, the condition number of $\tilde{S}_n^{-1}A_n$ is much larger than the condition number of $\hat{C}_n^{-1}A_n$. For example, if $f(x) = x^2$, then for $n = 1024$, $K_{\hat{C}_n^{-1}A_n} < 10^4$ while $K_{\tilde{S}_n^{-1}A_n} > 2 \cdot 10^6$ (K is the condition number). Thus, when f has a zero, we recommend the preconditioner $\hat{C}_n$.

REFERENCES

1. O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math. **48** (1986), 499–523. MR **88a:**65037b
2. R. H. Chan, *The spectrum of a family of circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal. **26** (1989), 503–506. MR **90f:**65048

3. R. H. Chan, X-Q. Jin, and M.-C. Yeung, *The spectra of super-optimal circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal. **28** (1991), 871–879. MR **92a:**65099
4. R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput. **10** (1989), 104–119. MR **90d:**65069
5. T. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput. **9** (1988), 766–771. MR **89e:**65046
6. I. Gohberg, T. Kailath, I. Koltracht, and T. Lancaster, *Efficient solution of linear systems of equations with recursive structure*, Linear Algebra Appl. **80** (1986), 80–113. MR **87i:**65058
7. T. Kailath, S. Kung, M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl. **68** (1979), 395–407. MR **80k:**65029
8. G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math. **74** (1986), 171–176.
9. V. Strela, *Exploration of circulant preconditioning properties*, Matrix Methods and Algorithms, IVM RAN, Moscow, 1993, pp. 9–46.
10. M. Tismenetsky, *A decomposition of Toeplitz matrices and optimal circulant preconditioning*, Linear Algebra Appl. **154/156** (1991), 105–121. MR **92c:**65056
11. E. Tyrtyshnikov, *Toeplitz matrices, some of their analogues and applications*, Dept. Numer. Math., USSR Acad. of Sci., Moscow (1989) (in Russian).
12. ———, *Optimal and superoptimal circulant preconditioners*, SIAM J. Matrix Anal. Appl. **13** (1992), 459–453. MR **92k:**65062
13. ———, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl. (to appear).
14. ———, *Circulant preconditioners with unbounded inverses*, Linear Algebra Appl. **216** (1995), 1–23.
15. ———, *Fast and parallel inertia finder for Toeplitz expanded matrices*, to appear.
16. M. Yeung and R. Chan, *Circulant preconditioners for Toeplitz matrices with piecewise continuous generating functions*, Math. Comp. **61** (1993), 701–718. MR **94a:**65024

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

E-mail address: strela@math.mit.edu

Institute of Numerical Mathematics, Russian Academy of Sciences, Leninskij Prosp., 32–A, 117334, Moscow, Russia

E-mail address: tee@adonis.iasnet.com