SOLVING ILL-CONDITIONED TOEPLITZ SYSTEMS VIA INDEX CANCELLATION*

GEORG HEINIG†, PETER KRAVANJA , AND MARC VAN BAREL‡

Abstract. We investigate to which extent the principle of index cancellation, which is well-known in the theory of Toeplitz operators, can be used to solve large ill-conditioned Toeplitz systems.

1. Introduction. Subject of the paper are large systems of linear equations with an ill-conditioned Toeplitz coefficient matrix:

$$T_n x = b, \quad T_n := [a_{j-k}]_{j,k=0}^{n-1} . \tag{1.1}$$

We consider the matrix T_n as a finite section of an infinite Toeplitz matrix and assume that the symbol of this infinite Toeplitz matrix does not vanish on the unit circle. Then, according to the theory of Toeplitz operators (see, e.g., [3]), the ill-conditioning of T_n can only be due to the fact that the index of the symbol is different from zero. The principle of *index cancellation* provides a way to reduce the noninvertible case to the invertible one. In this paper we investigate to which extent this tool can also be applied to reduce an ill-conditioned Toeplitz system to a well-conditioned one plus a small unstructured system.

To be more precise, let us explain the principle of index cancellation for Toeplitz operators. For proofs and more details we refer to [3]. Suppose that an infinite Toeplitz matrix $T = [a_{j-k}]_{j,k=0}^{\infty}$ with $\sum_{k=-\infty}^{\infty} |a_k| < \infty$ is given. The matrix T generates a bounded linear operator in the space ℓ^2. We will identify the matrix and the operator and denote both by T. The function

$$a(t) := \sum_{k=-\infty}^{\infty} a_k t^k \qquad (|t| = 1)$$

is called the *symbol* of T. We assume that

$$a(t) \neq 0 \tag{1.2}$$

for all $|t| = 1$. The integer

$$w = \operatorname{wind} a(t) := \frac{1}{2\pi} \int_0^{2\pi} d \arg a(e^{i\theta})$$

is called the *winding number* or the *index* of $a(t)$ (with respect to the unit circle). The operator T is invertible if and only if (1.2) is satisfied and $w = 0$. The principle of index cancellation is based on the fact that if (1.2) is satisfied, then the matrix $T^w := [a_{j-k+w}]_{j,k=0}^{\infty}$, which can be identified as a submatrix of T, generates an

*This research was partially supported by the Fund for Scientific Research–Flanders (FWO–V), project "Orthogonal systems and their applications," grant #G.0278.97 and by Kuwait University Research Project SM–190.

†Kuwait University, Department of Mathematics, POB 5969, Safat 13060, Kuwait (georg@mcs.sci.kuniv.edu.kw)

‡Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan 200 A, B-3001 Heverlee, Belgium (Peter.Kravanja@na-net.ornl.gov, Marc.VanBarel@cs.kuleuven.ac.be)

invertible operator in ℓ^2. By using this fact, the kernel structure and one-sided inverses of T can be described.

Let us now consider large finite Toeplitz systems (1.1), which are thought of as a finite section of an infinite system T with a symbol $a(t)$ satisfying (1.2). Let $\kappa(T_n)$ denote the condition number of T_n. We set $\kappa(T_n) := \infty$ if T_n is singular. The following is well-known (see [3]):

THEOREM 1.1. *If T is invertible (i.e., if (1.2) is satisfied and if the winding number of $a(t)$ is equal to zero), then*

$$\overline{\lim_{n\to\infty}}\ \kappa(T_n) < \infty.$$

Vice versa, if

$$\underline{\lim_{n\to\infty}}\ \kappa(T_n) < \infty,$$

then T is invertible.

NOTE. It has been shown more recently in [1] that for rational symbols $a(t)$ the limit $\lim_{n\to\infty} \kappa(T_n)$ actually exists (finite or infinite).

The previous theorem indicates that a Toeplitz matrix T_n can be ill-conditioned for two reasons. First, because the symbol takes small values. In this paper we do not consider this case and assume that the values of $a(t)$ are well away from zero. Second, because the winding number is different from zero. We assume that this is the case. A nonzero winding number causes the matrix T_n to have a large condition number. In fact, it is shown in [2] that in the case of rational symbols the condition number of T_n grows exponentially with n.

Standard fast algorithms for solving such systems of equations will fail in most cases because of stability problems. One idea to overcome these stability problems is to divide the problem into two parts: a large well-conditioned structured subproblem and a small ill-conditioned problem that will be treated as an unstructured ill-conditioned problem via a "slow" but stable method such as the singular value decomposition.

The principle of index cancellation tells us how to find a well-conditioned subproblem. We assume that the winding number w is small compared to n and we define $T_n^w := [a_{j-k+w}]_{j,k=0}^{n-w}$. Then T_n^w is a principal subsection of an invertible operator and therefore, cf. Theorem 1.1, this matrix can be expected to be well-conditioned. We now solve a subproblem with coefficient matrix T_n^w by using a fast Toeplitz solver and we tackle the remaining part of the problem by using a slow but safe algorithm such as the singular value decomposition. In the next section we are going to explain the details of this approach.

2. The algorithm. First we assume that the winding number w is positive. We decompose the matrix T_n, the right-hand side b and the solution vector x as

$$T_n =: \begin{bmatrix} F_1 & F_2 \\ T_n^w & G \end{bmatrix}, \quad b =: \begin{bmatrix} \beta \\ b_0 \end{bmatrix}, \quad x =: \begin{bmatrix} x_0 \\ \xi \end{bmatrix}. \tag{2.1}$$

Observe that $b_0, x_0 \in \mathbb{C}^{n-w}$ and $\beta, \xi \in \mathbb{C}^w$. Let $u := [T_n^w]^{-1} b_0$ and $U := [T_n^w]^{-1} G$.

LEMMA 2.1. *Assume that the matrices T_n and T_n^w are nonsingular. Then the $w \times w$ matrix $\Gamma := F_1 U - F_2$ is nonsingular and the solution vector x to $T_n x = b$ is obtained from*

$$\xi = \Gamma^{-1}(F_1 u - \beta) \qquad \text{and} \qquad x_0 = u - U\xi.$$

This lemma can be proved by immediate verification.

Now our algorithm can be described as follows:

ALGORITHM

1. Compute the winding number w of $a_n(t) := \sum_{k=1-n}^{n-1} a_k t^k$ (cf. Section 4) and decompose T_n, b and x as in (2.1).
2. Compute the solutions U of $T_n^w U = G$ and u of $T_n^w u = b_0$ by any fast Toeplitz solver and form the $w \times w$ matrix $\Gamma = F_1 U - F_2$.
3. Solve the linear system

$$\Gamma \xi = F_1 u - \beta.$$

   This gives the second part ξ of the solution vector x.
4. Compute the vector

$$x_0 = u - U\xi,$$

   which is the first part of the solution vector x.

The algorithm is particularly easy if $w = 1$. In this case Γ is only a scalar. We consider an instructive model case. Let m be a nonnegative integer and assume that $a_1 = 1$, $a_k = 0$ for $k \notin \{-m, \ldots, 0, 1\}$ and $\sum_{k=-m}^{0} |a_k| < 1$. Then the winding number w of $a(t) := \sum_k a_k t^k$ is equal to one and the function $a(t)t^{-1}$ is analytic and nonzero both on the unit circle and outside the unit disk. Therefore, its reciprocal admits a representation

$$\left(a(t)t^{-1}\right)^{-1} = \sum_{k=0}^{\infty} c_k t^{-k}$$

with $c_0 = 1$ and $\sum_{k=0}^{\infty} |c_k| < \infty$. We consider the matrix $T_n = [\, a_{j-k} \,]_{j,k=0}^{n-1}$ for $n > m$. Then the matrix T_n^1 is a nonsingular triangular Toeplitz matrix and its inverse is given by

$$C := \left(T_n^1\right)^{-1} = [\, c_{-(j-k)} \,]_{j,k=0}^{n-2}$$

where $c_k = 0$ for $k < 0$. We obtain that $U = -[\, c_{n-1-k} \,]_{k=0}^{n-2}$ and

$$\Gamma = -(a_0 c_{n-1} + \cdots + a_{2-n} c_1) - a_{1-n} = c_n.$$

Hence T_n is nonsingular if and only if $c_n \neq 0$. Since $F_1 [T_n^1]^{-1} = -\left[\begin{array}{ccc} c_1 & \cdots & c_{n-1} \end{array}\right]$, we obtain that

$$\xi = -\frac{1}{c_n}\, c^T b$$

where $c := [\, c_k \,]_{k=0}^{n-1}$. For the solution x of $T_n x = b$ we find that

$$x = \left[\begin{array}{c} Cb_0 \\ 0 \end{array}\right] - \frac{1}{c_n}\, \widehat{c}\, c^T b \tag{2.2}$$

where $\widehat{c} := [\, c_{n-1-k} \,]_{k=0}^{n-1}$. This formula shows that x can be computed in a stable way, provided that the zeros of $a(t)$ are not too close to the unit circle.

Formula (2.2) also leads to the inequality $\kappa(T_n) \geq |c_n|^{-1}$. If $a(t)$ has a simple zero t_0 such that $|t_0| < |t_j|$ for all other zeros t_j of $a(t)$, then we have that $|c_n|^{-1} = \mathcal{O}(|t_0|^{-n})$. This implies that the condition number of T_n grows at least exponentially as $|t_0|^{-n}$.

For the sake of completeness let us mention how to handle the case that $w < 0$. We now decompose the matrix T_n, the right-hand side b and the solution vector x as

$$T_n =: \begin{bmatrix} G & T_n^w \\ F_2 & F_1 \end{bmatrix}, \quad b =: \begin{bmatrix} b_0 \\ \beta \end{bmatrix}, \quad x =: \begin{bmatrix} \xi \\ x_0 \end{bmatrix}.$$

Now $b_0, x_0 \in \mathbb{C}^{n+w}$ and $\beta, \xi \in \mathbb{C}^{-w}$. Define again $u := [T_n^w]^{-1} b_0$ and $U := [T_n^w]^{-1} G$. The following lemma is the analogue of Lemma 2.1.

LEMMA 2.2. *Assume that the matrices T_n and T_n^w are nonsingular. Then the $(-w)\times(-w)$ matrix $\Gamma := F_1 U - F_2$ is nonsingular and the solution vector x to $T_n x = b$ is obtained from*

$$\xi = \Gamma^{-1}(F_1 u - \beta) \qquad \textit{and} \qquad x_0 = u - U\xi.$$

3. Numerical Experiments. We have implemented our algorithm in Matlab version 5.3 (floating point relative accuracy approximately $2.2\,10^{-16}$). In the following numerical experiments we have chosen the solution vector x as $x := [\ 1 \quad \cdots \quad 1\ ]^T$ and, for a given Toeplitz matrix T_n, we have computed the right-hand side vector b as the product $b := T_n \cdot x$, which we have evaluated by using FFT. We will compare the results obtained via our algorithm with the results obtained via the superfast solver for nonsymmetric Toeplitz systems that we have presented in [6] (for more information concerning this superfast solver, see also [5]). The computed approximations have not been improved via iterative refinement.

3.1. Band matrices. We have considered 100 Toeplitz band matrices of size 500×500. Each matrix contains 7 nonzero diagonals (there are 3 diagonals above the main diagonal and 3 diagonals below the main diagonal). The nonzero entries of the first row and the first column are random uniformly distributed in the interval $[0, 1]$. The winding numbers w of the generating functions of these matrices lie between -3 and 3. The following table gives for each value of w the number of corresponding matrices.

-3	-2	-1	0	1	2	3
5	2	31	14	38	4	6

In Figures 3.1–3.3 we plot, for the cases $w = 1, 2, 3$, the 2-condition numbers of T_n (dashed line) and T_n^w (+ signs) and the relative residual

$$\frac{\|b - T_n \hat{x}\|_2}{\|T_n\|_2 \|\hat{x}\|_2}$$

where $\hat{x}$ denotes the solution computed via the algorithm that we have presented in this paper (open dots) or via our more general superfast Toeplitz solver (solid line).

3.2. Exponential decay. For our next numerical experiment we have considered 100 Toeplitz matrices of size 500×500 whose entries are determined by $a_k := \eta_k / 2^{|k|}$ where the η_k's are random uniformly distributed in $[0, 1]$. The winding numbers w of the generating functions of these matrices lie between -2 and 1. The following table gives for each value of w the number of corresponding matrices.

-2	-1	0	1
1	20	65	14

In Figure 3.4 we plot, for the case $w = 1$, the 2-condition numbers and the relative residuals as before.

3.3. Polynomial decay. We have also considered 100 Toeplitz matrices of size 500×500 whose entries are determined by $a_k := \eta_k/k^2$ where the η_k's are again random uniformly distributed in $[0, 1]$. The winding numbers w of the generating functions of these matrices lie between -2 and 1. The following table gives for each value of w the number of corresponding matrices.

-2	-1	0	1
1	43	20	36

In Figure 3.5 we plot, for the case $w = 1$, the 2-condition numbers and the relative residuals as before.

Finally, we have also considered 100 Toeplitz matrices of size 500×500 whose entries are determined by $a_k := \eta_k/k$ where the η_k's are random uniformly distributed in $[0, 1]$. The winding numbers w of the generating functions of these matrices lie between -3 and 2. The following table gives for each value of w the number of corresponding matrices.

-3	-2	-1	0	1	2
1	5	35	30	26	3

In Figure 3.6 we plot, for the case $w = 1$, the 2-condition numbers and the relative residuals as before.

4. Appendix: Fast Winding Number Calculation. The winding number of a polynomial can be evaluated in many ways. The Schur–Cohn test is a classical algorithm (see [4]). It has complexity $\mathcal{O}(n^2)$. An iterative method is given in [7] (see also [3]). Below we describe another possibility. It is based on FFT and hence requires only $\mathcal{O}(n \log n)$ flops.

Let $a(t) := a_0 + \cdots + a_n t^n$ be a polynomial that does not vanish on the unit circle. We want to compute the winding number $\operatorname{wind} a(t)$. Define

$$M_1 := \sum_{k=1}^{n} k|a_k|.$$

Note that $|a'(t)| \le M_1$ for all $|t| = 1$.

PROCEDURE. Let $\omega_N^1, \ldots, \omega_N^N$ denote the Nth roots of unity in their classical counterclockwise configuration. Start with some $N \ge n$, preferably a power of 2, and compute $a(\omega_N^1), \ldots, a(\omega_N^N)$ by using FFT. Then check whether

$$N > \frac{\pi M_1}{\min_k |a(\omega_N^k)|} \tag{4.1}$$

holds. If not, then replace N by $2N$ and calculate the values $a(\omega_{2N}^k)$ that have not been computed yet. Since $\min_{|t|=1} |a(t)| > 0$, the inequality (4.1) will be satisfied for sufficiently large N and our procedure is guaranteed to end.

If N is such that (4.1) is satisfied, then the winding number is obtained in the following way. Consider only those values $a(\omega_N^k)$ that have a positive real part and

look for the sign changes in the imaginary parts. Let ν_+ denote the number of changes from $-$ to $+$ and ν_- the number of changes from $+$ to $-$. Then

$$\text{wind}\, a(t) = \nu_+ - \nu_-.$$

JUSTIFICATION. We have computed the winding number of the polygon with vertices $a(\omega_N^k)$. We have to show that this number is equal to the winding number of $a(t)$. If these two integers are different, then there has to exist an index k such that the closed contour Γ_k consisting of the line segment from $a(\omega_N^k)$ to $a(\omega_N^{k+1})$ ($k+1$ is understood modulo N) and the arc γ_k of $a(t)$ going from $a(\omega_N^k)$ to $a(\omega_N^{k+1})$ contains the point 0 in its interior. The length $\ell(\gamma_k)$ of the arc γ_k can be estimated as

$$\ell(\gamma_k) = \int_{2\pi k/N}^{2\pi(k+1)/N} |a'(e^{\mathrm{i}\theta})|\, d\theta \;\le\; 2\pi\, M_1/N.$$

On the other hand, the length of any arc γ_k such that the point 0 lies in the interior of Γ_k is at least $|a(\omega_N^k)| + |a(\omega_N^{k+1})|$. Thus, if (4.1) is satisfied, then this cannot be the case.

Concluding remarks. In this paper we have presented an algorithm for solving large ill-conditioned Toeplitz systems. Our approach is based on the principle of index cancellation and reduces the problem to that of solving a well-conditioned Toeplitz system plus a small unstructured system. We have done a large number of experiments on several types of ill-conditioned Toeplitz matrices (band matrices, exponential decay matrices, polynomial decay matrices) of size 500×500. Our numerical results indicate that index cancellation leads to a significant increase in accuracy in case the decay of the elements a_k is at least as $1/k^2$. This reflects the fact that in this case the matrix can be considered as a finite section of a Toeplitz operator with a continuous symbol. In the case of a decay of $1/k$ or less, index cancellation does not improve the accuracy very much. Surprisingly in this case the standard algorithm already gives accurate results in the case of a nonzero index.

REFERENCES

[1] A. BÖTTCHER, *Pseudospectra and singular values of large convolution operators*, J. Integral Equations Appl., 6 (1994), pp. 267–301.

[2] A. BÖTTCHER AND S. GRUDSKY, *Toeplitz band matrices with exponentially growing condition number*. To appear.

[3] I. C. GOHBERG AND I. A. FEL'DMAN, *Convolution Equations and Projection Methods for their Solution*, vol. 41 of Translations of Mathematical Monographs, American Mathematical Society, Providence, Rhode Island, 1974.

[4] P. HENRICI, *Applied and Computational Complex Analysis: I. Power Series—Integration—Conformal Mapping—Location of Zeros*, Wiley, 1974.

[5] P. KRAVANJA, *On Computing Zeros of Analytic Functions and Related Problems in Structured Numerical Linear Algebra*, PhD thesis, Katholieke Universiteit Leuven, Department of Computer Science, Mar. 1999.

[6] M. VAN BAREL, G. HEINIG, AND P. KRAVANJA, *A stabilized superfast solver for nonsymmetric Toeplitz systems*, Report TW 293, K.U.Leuven, Dept. Computer Science, Oct. 1999.

[7] V. L. ZAGUSKIN AND A. V. CHARITONOV, *An iteration method for the solution of problems concerning stability*, Zhurnal vysch. matema. i matem. fiz., 3 (1963), pp. 361–364. (In Russian).

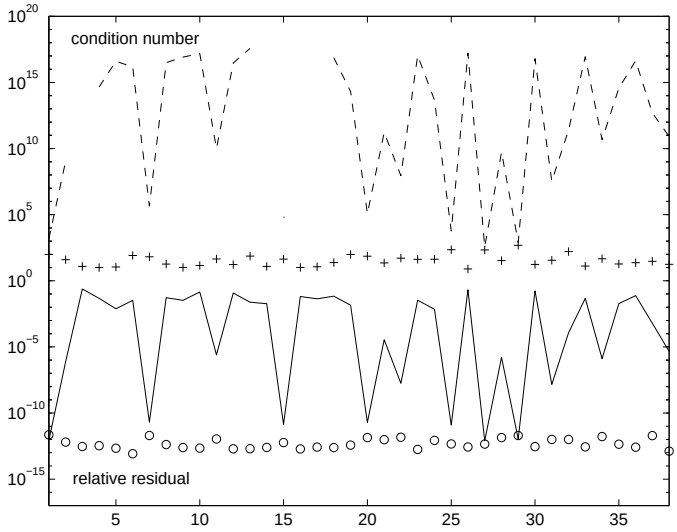

FIG. 3.1. *Condition numbers and relative residuals in case* $w = 1$ *(band matrices of size* 500×500*).*

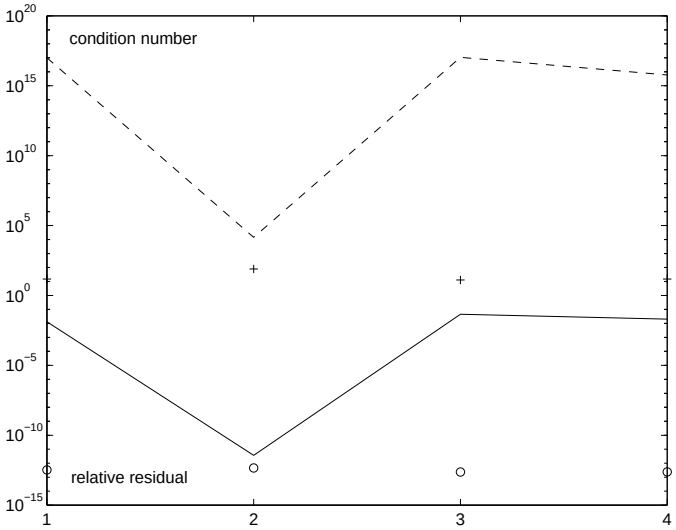

FIG. 3.2. *Condition numbers and relative residuals in case* $w = 2$ *(band matrices of size* 500×500*).*

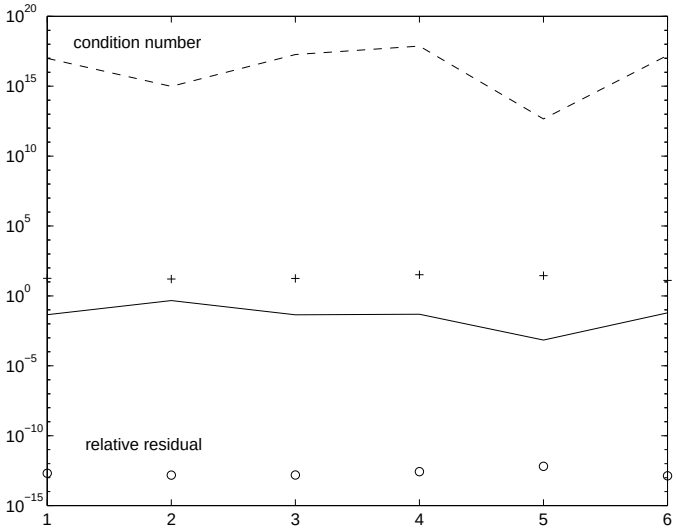

FIG. 3.3. *Condition numbers and relative residuals in case* $w = 3$ *(band matrices of size* 500×500*).*

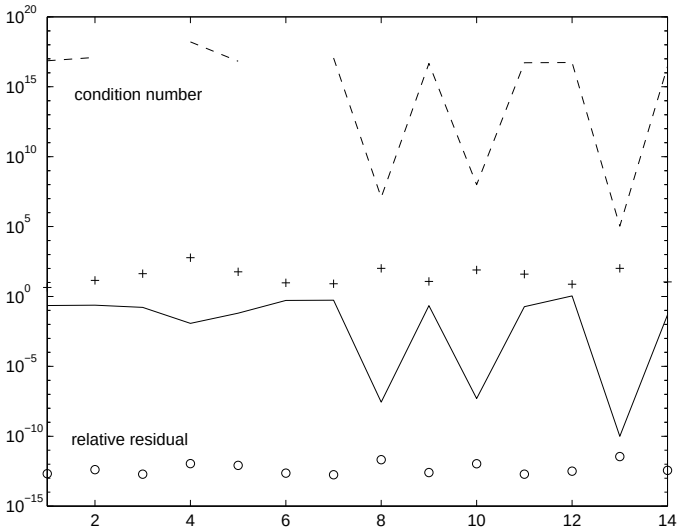

FIG. 3.4. *Condition numbers and relative residuals in case* $w = 1$ *(exponential decay matrices of size* 500×500*).*

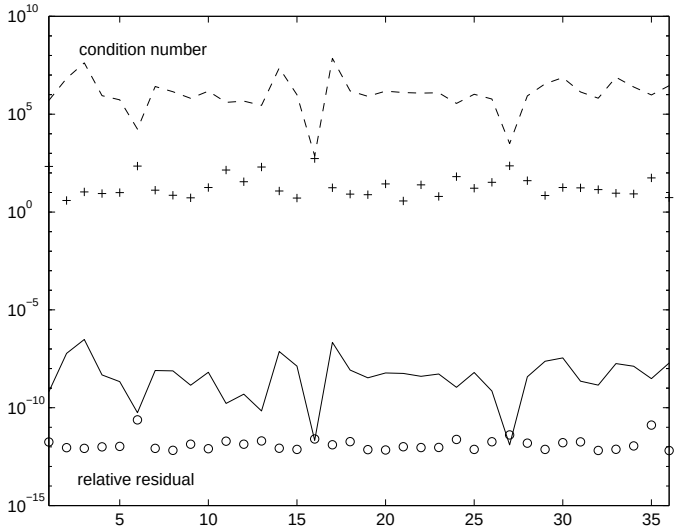

FIG. 3.5. *Condition numbers and relative residuals in case* $w = 1$ *(polynomial decay '*$1/k^2$*' matrices of size* 500×500*).*

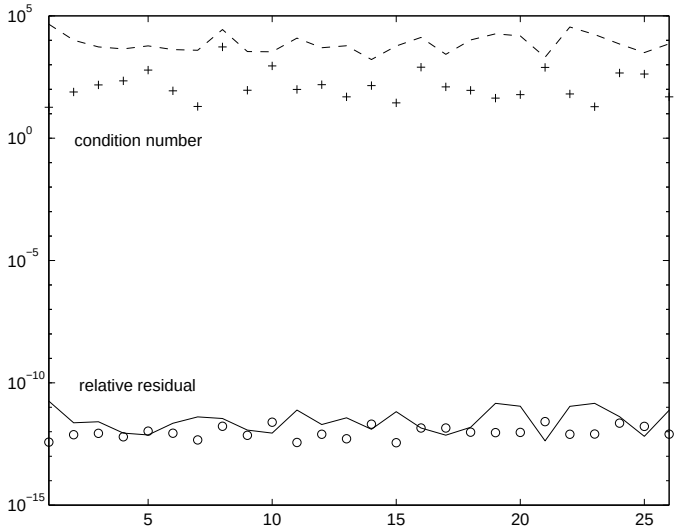

FIG. 3.6. *Condition numbers and relative residuals in case* $w = 1$ *(polynomial decay '*$1/k$*' matrices of size* 500×500*).*