Explicit inversion formulas for Toeplitz band matrices

William F. Trench, Trinity University

Available at: https://works.bepress.com/william_trench/43/
EXPLICIT INVERSION FORMULAS FOR TOEPLITZ BAND MATRICES*

WILLIAM F. TRENCH†

Abstract. Explicit formulas are given for the elements of \( T_n^{-1} \) and the solution of \( T_n X = Y \), where \( T_n \) is an \((n+1) \times (n+1)\) Toeplitz band matrix with bandwidth \( k \leq n \). The formulas involve \( k \times k \) determinants whose entries are powers of the zeros of a certain \( k \)th degree polynomial \( P(x) \) which is independent of \( n \), or simple related functions of these zeros if any are repeated. It is shown that \( T_n \) is invertible if and only if a certain \( k \times k \) determinant involving these zeros is nonvanishing.

AMS(MOS) classification numbers. 15, 65F

1. Introduction. We consider Toeplitz band matrices, i.e., matrices of the form

\[
T_n = (\phi_{\nu-\mu})_{\nu,\mu=0}^n,
\]

where there are nonnegative integers \( p \) and \( q \) such that

\[
\phi_{\nu} = 0 \quad \text{if} \quad \nu > p \quad \text{or} \quad \nu < -q.
\]

We use the notation of [13] and [15]. Notice that \( T_n \) is of order \( n+1 \), with rows and columns numbered from 0 to \( n \). We write

\[
T_n^{-1} = B_n = (b_{mn})_{m,n=0}^n.
\]

It is assumed throughout that

\[
\phi_p \phi_{-q} \neq 0 \quad \text{and} \quad p + q = k \leq n.
\]

Our main results are explicit formulas for the elements of \( T_n^{-1} \) and for the solution of the system

\[
T_n X = Y,
\]

in terms of the zeros of the polynomial

\[
P(x) = \sum_{\mu=-q}^{p} \phi_{\mu} x^{n+\mu}.
\]

These formulas involve determinants of order \( k \), the bandwidth of \( T_n \).

Many authors (e.g., [1], [3], [6], [7], [9], [10], [12], [15]) have given formulas and algorithms for inverting Toeplitz band matrices. Efficient methods have also been developed for solving (3) (e.g., [2], [4], [11], [14]). Since a survey of results along these lines is given in the introduction to the recent paper [9], there is no need to review earlier work here. We believe that the results presented here are new, and more general and explicit than others heretofore published. We treat the general Toeplitz band matrix, without assuming that \( T_n \) is symmetric or hermitian. Our formulas are explicit (i.e., not recursive with respect to \( n \)), and we do not have to assume that any matrix other than \( T_n \) is nonsingular; however, we do give a method for computing \( T_n^{-1} \) efficiently in the case where \( T_n^{-1} \) is also nonsingular.

The idea motivating our approach is that if \( n \) is large compared to \( k \), then \( T_n \) is "nearly triangular" in an obvious visual sense, which need not be defined precisely.

* Received by the editors December 20, 1983, and in revised form April 19, 1984.
† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104.
Therefore, it is not surprising that the elements of $T_n^{-1}$ are closely related to those of $(T^*_n)^{-1}$, where $T^*_n$ is the lower triangular Toeplitz matrix

$$T^*_n = (\phi_{r-s-q})_{r,s=0}^{n}.$$ 

The inverse of this matrix is the lower triangular Toeplitz matrix

$$\left(T^*_n\right)^{-1} = (\alpha_{r-s})_{r,s=0}^{n} \quad (\alpha_r = 0 \text{ if } r < 0),$$

with elements independent of $n$, defined by

$$\left(P(z)\right)^{-1} = \sum_{n=0}^{\infty} \alpha_n z^n.$$ 

It is easy to find an explicit formula for $\alpha_r$ in terms of the zeros of $P(z)$. Moreover, we will show that the differences

$$b_{rn} - \alpha_{r-s-q}, \quad 0 \leq r, s \leq n$$

can be found easily and explicitly in terms of the zeros of $P(z)$. This leads to explicit formulas for $T_n^{-1}$ and the solution of (3).

We also give analogous formulas based on the inverse of the upper triangular Toeplitz matrix

$$T^U_n = (\phi_{r-s+p})_{r,s=0}^{n}.$$ 

The inverse of this matrix is

$$\left(T^U_n\right)^{-1} = (\beta_{r-s})_{r,s=0}^{n} \quad (\beta_r = 0 \text{ if } r < 0),$$

with elements independent of $n$, defined by

$$\left(z^p P\left(\frac{1}{z}\right)^{-1} = \sum_{n=0}^{\infty} \beta_n z^n.$$ 

### 2. Preliminary results

The following assumption applies throughout. (Recall (2) here.)

**Assumption A.** The distinct zeros of (4) are $z_1, \ldots, z_m$ with multiplicities $\mu_1, \ldots, \mu_m$; thus, $m \leq k$, $\mu_i \geq 1$, and

$$\mu_1 + \cdots + \mu_m = k.$$ 

**Definition 1.** If $j_1, \ldots, j_k$ are integers, let

$$C(z; j_1, \ldots, j_k) = \text{col}[z^{j_1}, \ldots, z^{j_k}],$$

and let $C^{(l)}(z; j_1, \ldots, j_k)$ denote the $l$th derivative of this column vector. Now define the $k \times k$ determinant $D(j_1, \ldots, j_k)$ as follows: Its first $\mu_1$ columns are $C^{(l)}(z; j_1, \ldots, j_k)(0 \leq l \leq \mu_1 - 1)$; it next $\mu_2$ columns are $C^{(l)}(z; j_1, \ldots, j_k)(0 \leq l \leq \mu_2 - 1)$; and so forth.

For example, if (4) has $k$ distinct roots, then

$$D(j_1, \ldots, j_k) = \text{det} (z_i^{j})_{i=1}^{k}.$$

There is an ambiguity in Definition 1, since the $m$ zeros of $P(z)$ may be numbered in any order; however, our results involve ratios of the form

$$D(j_1, \ldots, j_k)/D(j'_1, \ldots, j'_k),$$

which are left invariant if $z_1, \ldots, z_m$ are permuted. Because of (2), $z_j \neq 0$ (1 $\leq j \leq m$),
so \( D(j_1, \cdots, j_k) \) exists for all \( j_1, \cdots, j_k \). It can be shown that

\[
D(0, 1, \cdots, k-1) = K \prod_{1 \leq i < j \leq m} (z_j - z_i)^{y_{ji}},
\]

where \( K > 0 \) and the \( y_{ji} \)'s are positive integers. If \( m = k \), then \( D(0, 1, \cdots, k-1) \) is the Vandermonde determinant, so \( K = r_q = 1 \).

**Lemma 1.** Suppose \( 1 \leq i \leq k \) and \( j_1, \cdots, j_{i-1}, j_{i+1}, \cdots, j_k \) are fixed integers. Then the sequence

\[
e_r = D(j_1, \cdots, j_{i-1}, r, j_{i+1}, \cdots, j_k), \quad -\infty < r < \infty
\]

satisfies the difference equation

\[
\sum_{\nu=-q}^{p} \phi_r e_{r+r} = 0.
\]

**Proof.** Because of Definition 1, expanding the determinant in (9) in terms of cofactors of its \( i \)th row yields

\[
e_r = \sum_{j=1}^{m} \sum_{i=0}^{\mu_i - 1} a_{ji} (r)^{j} z_j^{-i},
\]

where

\[
(r)^{0} = 1, \quad (r)^{i} = r(r-1) \cdots (r-i+1), \quad i \geq 1,
\]

and the \( a_{ji} \)'s are constants. But if \( 1 \leq j \leq m \) and \( 0 \leq i \leq \mu_j - 1 \), then

\[
\sum_{\nu=-q}^{p} \phi_r (\nu+r)^{(i)} z_j^{\nu+r} = 0, \quad -\infty < r < \infty,
\]

since \( z_j \) is a zero of \( z^m + P(z) \), with multiplicity \( \mu_j \), for every \( r \). This and (11) imply (10).

**Lemma 2.** a) The sequence \( \{a_r\} \) defined by

\[
\alpha_r = \begin{cases} 
0, & r < 0, \\
\frac{1}{\phi_{r-q}} \frac{D(-r, 1, \cdots, k-1)}{D(0, 1, \cdots, k-1)}, & r \geq -k+1,
\end{cases}
\]

satisfies

\[
\sum_{\nu=-q}^{p} \phi_r \alpha_{r+q} = \delta_{j0}, \quad -\infty < j < \infty.
\]

b) The sequence \( \{\beta_r\} \) defined by

\[
\beta_r = \begin{cases} 
0, & r < 0, \\
\frac{1}{\phi_{r-p}} \frac{D(0, 1, \cdots, k-2, r+k-1)}{D(0, 1, \cdots, k-1)}, & r \geq -k+1,
\end{cases}
\]

satisfies

\[
\sum_{\nu=-q}^{p} \phi_r \beta_{r+p+j} = \delta_{j0}, \quad -\infty < j < \infty.
\]

(Note. The definitions (12) and (14) are redundant, but consistent, for \(-k+1 \leq j \leq -1\). They are stated this way for convenience.)
Proof. (a) If \( j < 0 \), then (12a) implies (13). If \( j = 0 \), then (13) reduces to
\[
\phi_{-q} a_0 = 1,
\]
again because of (12a). This is consistent with (12b) with \( r = 0 \). If \( j \geq 1 \), then (12b) and Lemma 1 imply (13).

(b) Similar proof.

Notice that the sequences \( \{a_r\} \) and \( \{\beta_r\} \) can be computed recursively from (13) and (15), or explicitly from (12) and (14).

Lemma 2 implies (5), (6), (7), and (8). Therefore, (12) provides an explicit formula for \( T_n^{-1} \) if \( q = 0 \) (\( T_n \) is lower triangular), while (14) serves the same purpose if \( p = 0 \) (\( T_n \) is upper triangular). Of course, the inversion of triangular Toeplitz matrices—banded or not—is very simple, as was observed in [15]. We assume henceforth that
\[
q \geq 1 \quad \text{and} \quad p \geq 1.
\]

3. The main results. The next theorem follows from a result in [16] concerning the eigenvalues of Toeplitz band matrices; however, since the proof in [16] utilizes a more involved argument than is needed here, it is convenient to prove Theorem 1 directly.

**Theorem 1.** If (2) and (16) hold, then \( T_n \) is invertible if and only if
\[
D(0, 1, \cdots, q - 1, n + q + 1, \cdots, n + k) \neq 0.
\]

**Proof.** We prove the equivalent assertion that the system
\[
T_n X = 0 \quad (\text{transpose})
\]
has a nontrivial solution \( X = \text{col}[x_0, \cdots, x_n] \) if and only if
\[
D(0, 1, \cdots, q - 1, n + q + 1, \cdots, n + k) = 0.
\]

Easy manipulations show that (18) holds if and only if the finite sequence
\[
x_{-q}, \cdots, x_{-1}, x_0, \cdots, x_m, x_{m+1}, \cdots, x_{n+p}
\]
satisfies the boundary value problem
\[
(a) \quad \sum_{\nu=-q}^p \phi_{\nu} x_{r+\nu} = 0, \quad 0 \leq r \leq n,
\]
\[
(b) \quad x_r = 0 \quad \text{if} \quad -q \leq r \leq -1 \text{ or } n + 1 \leq r \leq n + p.
\]

However, because of Assumption A and the fact that \( z_j \neq 0 \) \((1 \leq j \leq m)\), the elementary theory of constant coefficient difference equations implies that a solution of (20a) must be of the form
\[
x_r = \sum_{j=1}^{m} \sum_{i=0}^{\mu-1} a_i (q + r)^{i+j} z_j, \quad -q \leq r \leq n + p,
\]
where
\[
A = \text{col}[a_{01}, \cdots, a_{m-1,1}, \cdots, a_{0m}, \cdots, a_{m-1,m}]
\]
is a constant vector. On recalling Definition 1, it can be seen that (21) is consistent with (20b) if and only if \( A \) satisfies the \( k \times k \) system
\[
HA = 0,
\]
where
\[ \det H = D(0, 1, \ldots, q-1, n+q+1, \ldots, n+k). \]

Therefore (22) has a nontrivial solution, and the same is true of (18), if and only if (19) holds. This completes the proof.

Henceforth, we assume that (17) holds.

**Definition 2.** Let
\[ U_n = \{0, 1, \ldots, q-1, n+q+1, \ldots, n+k\}. \]

If \( \mu \in U_n \) and \( l \) is an arbitrary integer, define
\[
a_{n,\mu}(l) = \frac{D(j_0, \ldots, j_{q-1}, j_{n+q+1}, \ldots, j_{n+k})}{D(0, 1, \ldots, q-1, n+q+1, \ldots, n+k)},
\]
where
\[ j_i = \begin{cases} i & \text{if } i \notin U_n - \{\mu\}, \\ l & \text{if } i = \mu. \end{cases} \]

For example,
\[ a_{n,0}(0) = \frac{D(l, 1, \ldots, q-1, n+q+1, \ldots, n+k)}{D(0, 1, \ldots, q-1, n+q+1, \ldots, n+k)} \]
and
\[ a_{n,0}(n+q+1) = \frac{D(0, 1, \ldots, q-1, l, n+q+2, \ldots, n+k)}{D(0, 1, \ldots, q-1, n+q+1, \ldots, n+k)}. \]

Lemma 1 and Definition 2 imply the following lemma.

**Lemma 3.** If \( \mu \) is a fixed integer in \( U_n \), then
\[ \sum_{\nu=-q}^{q} \phi_{\nu} a_{n,\mu}(\nu + r) = 0, \quad -\infty < r < \infty, \]
and
\[ a_{n,\mu}(r) = \delta_{\mu r}, \quad r \in U_n; \]

i.e., \( e = a_{n,\mu}(r) \) is the unique solution of (10) which satisfies the boundary conditions
\[ e = \delta_{\mu r}, \quad 0 \leq r \leq q-1, \quad n+q+1 \leq r \leq n+k. \]

The uniqueness assertion of Lemma 3 follows from (17), as can be seen from the proof of Theorem 1, since the difference of two solutions would satisfy (20).

The next two theorems give explicit formulas for \( b_{r,s} \), the general element of \( T_n^{-1} \). The formula in Theorem 2 is more convenient if \( q < p \), while the formula in Theorem 3 is more convenient if \( p > q \).

**Theorem 2.** The general element \( b_{r,s} \) of \( B_n = T_n^{-1} \) is given by
\[ b_{r,s} = a_{r-s-q} - \sum_{l=0}^{q-1} a_{r-s-l}(q+s), \]
where \( \{a_{r}\} \) is as in (12).

**Proof.** The condition \( B_n T_n = I_n \) is equivalent to
\[ \sum_{j=0}^{n} b_{r,s} \phi_{j} = \delta_{r,s}, \quad 0 \leq r, s \leq n. \]
which can be rewritten as
\[(24) \quad \sum_{j=-q}^{n+p} b_{j,s} \phi_{j-s} = \delta_{rs}, \quad 0 \leq r, s \leq n,\]

if we define
\[(25) \quad b_{rs} = 0 \quad \text{when} \quad -q \leq s \leq -1 \text{ or } n+1 \leq s \leq n+p.\]

Shifting the index of summation in (24) and recalling (1) yields
\[(26) \quad \sum_{s=-q}^{p} \phi_{s} b_{n+r+s,n} = \delta_{rs}, \quad 0 \leq r, s \leq n.\]

Now let
\[(27) \quad b_{rs} = \alpha_{r-s-n} + u_{rs}, \quad 0 \leq r \leq n, -q \leq s \leq n+p,\]

where \(u_{rs}\) is to be determined. From (13) with \(j = r-s\),
\[\sum_{s=-q}^{p} \phi_{s} \alpha_{r-1-s-n} = \delta_{0,r-1} = \delta_{rs}, \quad 0 \leq r, s \leq n;\]

hence, substituting (27) into (26) and recalling (25) shows that for each \(r\) in \(\{0, \cdots, n\}\),
the sequence \(u_{rs}\) satisfies the difference equation
\[\sum_{s=-q}^{p} \phi_{s} u_{r+s-n} = 0, \quad 0 \leq s \leq n,\]

and the boundary conditions
\[u_{rn} = -\alpha_{r-s-n} = 0, \quad n+1 \leq s \leq n+p \quad (\text{cf. (12a)}),\]
\[u_{rn} = -\alpha_{r-s-n}, \quad -q \leq s \leq -1.\]

This and Lemma 3 imply that
\[u_{rs} = \sum_{l=0}^{n+q} \alpha_{r-s-n}(l), \quad q+s,\]

which, with (27), implies (23).

**Theorem 3.** The general element \(b_{rs}\) of \(B_n = T_n^{-1}\) is given by
\[(28) \quad b_{rs} = \beta_{r-s-p} - \sum_{l=0}^{n+q} \beta_{r-p-l+1} \alpha_{r-s-n+q+l+1}, \quad n+q-r,\]

with \(\beta_{r}\) as defined by (14).

**Proof.** The condition \(T_n B_n = I_n\) is equivalent to
\[\sum_{j=0}^{n} \phi_{j-r} b_{jn} = \delta_{rn}, \quad 0 \leq r, s \leq n,\]

which can be rewritten as
\[(29) \quad \sum_{j=-p}^{n} \phi_{j-r} b_{jn} = \delta_{rn}, \quad 0 \leq r, s \leq n,\]

if we define
\[(30) \quad b_{rs} = 0 \quad \text{when} \quad -p \leq r \leq -1 \text{ or } n+1 \leq r \leq n+q.\]
Changing the index of summation in (29) and recalling (1) yields

$$\sum_{\nu = -q}^{\rho} \phi_{\nu} b_{r,\nu} = \delta_{rs}, \quad 0 \leq r, s \leq n. \quad (31)$$

Now let

$$b_{rs} = \beta_{s-r-p} + v_{rs}, \quad (32)$$

where \( v_{rs} \) is to be determined. From (15) with \( j = s - r \),

$$\sum_{\nu = -q}^{\rho} \phi_{\nu} \beta_{s-r-p+\nu} = \delta_{rs}, \quad 0 \leq r, s \leq n;$$

hence, substituting (32) into (31) and recalling (30) shows that for each \( s \) in \( \{0, \ldots, n\} \), the sequence \( \{v_{rs}\}_{r=0}^{\rho} \) satisfies the difference equation

$$\sum_{\nu = -q}^{\rho} \phi_{\nu} v_{r-\nu} = 0, \quad 0 \leq r \leq n,$$

and the boundary conditions

$$v_{rs} = -\beta_{s-r-p} = 0, \quad n+1 \leq r \leq n+q,$$

$$v_{rs} = -\beta_{s-r-p}, \quad -p \leq r \leq -1.$$

This and Lemma 3 imply that

$$v_{rs} = -\sum_{i=0}^{r-1} \beta_{s-r-p+i+1} a_{n}(n+q+l+1|n+q-r);$$

which, with (32), implies (28).

The next theorem provides explicit formulas for the solution of (3) when \( T_n \) is invertible. Here we write

$$X = \text{col} [x_0, \ldots, x_n] \quad \text{and} \quad Y = \text{col} [y_0, \ldots, y_n]$$

and adopt the convention that

$$\sum_{\nu = 0}^{\mu} = 0 \quad \text{if} \ \nu < \mu.$$

**Theorem 4.** If \( T_n \) is invertible, then the solution of (3) is given by

$$x_r = \sum_{s=0}^{r+q} \alpha_{r-s-q} y_s - \sum_{s=0}^{r+q-1} \alpha_{r-s} \sum_{l=0}^{n} y_l a_n(l|q+s), \quad 0 \leq r \leq n, \quad (33)$$

and by

$$x_r = \sum_{s=-p}^{n} \beta_{r-s-p} y_s - \sum_{i=0}^{p-1} \left( \sum_{s=0}^{n} \beta_{r-s-p+i+1} y_s \right) a_n(n+q+l+1|n+q-r), \quad 0 \leq r \leq n. \quad (34)$$

**Proof.** Since

$$x_r = \sum_{s=0}^{n} b_{rs} y_s \quad 0 \leq r \leq n,$$

(23) implies (33) and (28) implies (34).

Since convolutions can be implemented efficiently by means of fast Fourier transforms, (33) provides an efficient computational method for solving (3). The
quantities

\[ M_l = \sum_{s=0}^{n} y_s a_n(l|q+s), \quad 0 \leq l \leq q-1 \]

(each of which can be expressed as the ratio of two \( k \times k \) determinants) would be computed first. Then, from (33),

\[ x_r = -\sum_{l=0}^{r} M_l a_{r-l} \quad 0 \leq r \leq q-1, \]

and

\[ x_r = \sum_{s=0}^{r-q} \alpha_{r-s-q} y_s \quad -\sum_{l=0}^{q-1} M_l a_{r-l} \quad q \leq r \leq n. \]

It is easily verified that (33) and (34) remain valid if \( p = 0 \) or \( q = 0 \).

The next lemma follows trivially from the last four equations of (13).

**Lemma 4.** Suppose \( T_n \) is an arbitrary (not necessarily banded) Toeplitz matrix, with inverse \( T_n^{-1} = (b_{kn})_{k,n=0}^{n} \), where

\[ b_{00} \neq 0. \]

Then the elements \( b_{rs} (1 \leq r, s \leq n) \) are determined in terms of \( b_{rn} (0 \leq r \leq n) \) and \( b_{0n} (0 \leq s \leq n) \) by the recursion formula

\[ b_{rs} = b_{r-1,s-1} + (b_{00})^{-1} (b_{rn} b_{0n} - b_{n-r+1,0} b_{0,n-r+1}), \quad 1 \leq r, s \leq n. \]

Since \( b_{00} = \det T_{n-1}/\det T_n \) (35) implies that \( T_{n-1} \) is also invertible. Since \( T_n^{-1} \) is persymmetric (i.e., symmetric about its secondary diagonal), it is only necessary to use (36) for \( r + s \leq n \), and then take

\[ b_{rs} = b_{n-r,n-s}, \quad 1 \leq r, s \leq n, \quad r + s > n. \]

Lemma 4 was rediscovered and presented in a useful matrix form by Gohberg and Semencul [5]. In most applications (e.g., [3], [8], [13], [15], [16]), it has been coupled with recursive procedures for obtaining the elements of the zeroth row and column of \( T_n^{-1} \); however, these methods usually require that other matrices in the sequence \( \{T_0, T_1, \cdots \} \) be nonsingular. Since Theorems 2 and 3 provide convenient explicit formulas for the zeroth row and column of \( T_n^{-1} \), we can dispense with this additional assumption here. Thus, from (12) and (23),

\[ b_{0s} = -(\phi_{-q})^{-1} a_k (0|q+s), \quad 0 \leq s \leq n, \]

while from (14) and (28),

\[ b_{0n} = -(\phi_n)^{-1} a_k (n+1|n+q-r), \quad 0 \leq r \leq n. \]

If evaluating the \( k \times k \) determinants in (37) is inconvenient, then it is only necessary to use (37) for \( 0 \leq s \leq p \); define

\[ b_{0s} = 0, \quad -q+1 \leq s \leq -1, \]

and compute recursively:

\[ b_{0s} = -(\phi_p)^{-1} \sum_{r=s}^{p-1} \phi_r b_{0, r-s-n}, \quad p+1 \leq s \leq n. \]
(See Lemma 3.) Similarly, we can use (38) for $0 \leq r \leq q$; define

$$b_{ron} = 0, \quad -p + 1 \leq r \leq -1,$$

and compute recursively:

$$b_{ron} = -\left(\phi_{-p}\right)^{-1} \sum_{r=-p}^{q-1} \phi_{r} b_{r+r-s, q, 0, n}, \quad q + 1 \leq r \leq n.$$

REFERENCES