

# On the Application of $\Sigma$ -orthogonal Transformations with Elementary Hyperbolic Rotations

Michael Stewart<sup>1</sup> and Paul Van Dooren<sup>2</sup>

## 1 Introduction

In block implementations of the generalized Schur algorithm for the Cholesky factorization of a block Toeplitz matrix it is necessary to compute

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

such that

$$H^T \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \hat{A} \\ 0 \end{bmatrix} \quad (1)$$

and

$$H^T \Sigma H = \Sigma \quad (2)$$

with

$$\Sigma = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

Here we take  $A$  to be square. Through most of this paper, without a significant loss of generality, we will restrict  $B$  to be square and of the same size. The reasons for this last restriction will be made clear.

Any transformation,  $H$ , satisfying (2) is referred to as  $\Sigma$ -orthogonal. Whenever  $A^T A - B^T B$  is positive definite, the existence of a  $\Sigma$ -orthogonal transformation satisfying (1) is guaranteed.

The positivity constraint is naturally satisfied in  $A$  and  $B$  which arise when applying the generalized Schur algorithm to a positive definite structured matrix. Although the existence of the transformation is guaranteed, there is no single approach to computing it. A common algorithm applies

---

<sup>1</sup>Computer Sciences Laboratory, RSISE, Australian National University, Canberra ACT 0200, Australia, email: [stewart@discus.anu.edu.au](mailto:stewart@discus.anu.edu.au)

<sup>2</sup>Dept. Mathematical Engineering, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, ([Vandooren@anma.ucl.ac.be](mailto:Vandooren@anma.ucl.ac.be)).

and computes  $H$  in a form which is factored into elementary orthogonal transformations (plane rotations or Householder transformations) and  $2 \times 2$  hyperbolic rotations.

It is easy to see that  $\Sigma$ -orthogonal transformations form a multiplicative group, i.e. the product  $H_1 H_2$  is  $\Sigma$ -orthogonal if  $H_1$  and  $H_2$  are  $\Sigma$ -orthogonal. It is then natural to decompose such matrices into a product of simpler  $\Sigma$ -orthogonal transformations. Typical choices are matrices of the form

$$H = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix}$$

and

$$\begin{bmatrix} I & & & & \\ & c_h & & s_h & \\ & & I & & \\ & s_h & & c_h & \\ & & & & I \end{bmatrix}$$

where  $c_h^2 - s_h^2 = 1$  and where the latter transformation acts on a single row of  $A$  together with a single row of  $B$ . It is possible to express  $c_h$  and  $s_h$  in terms of a single parameter by writing

$$c_h = \frac{1}{\sqrt{1 - \rho^2}}$$

and

$$s_h = \frac{\rho}{\sqrt{1 - \rho^2}}.$$

Thus there are two transformations from which to construct  $\Sigma$ -orthogonal transformations: orthogonal transformations which act solely on  $A$  or on  $B$  or hyperbolic transformations which act on a row of  $A$  together with a row of  $B$ .

A common approach for producing an  $H$  satisfying (1) follows a simple triangularization procedure. Suppose orthogonal  $U_A$  and  $U_B$  are computed so that

$$\begin{bmatrix} U_A^T & 0 \\ 0 & U_B^T \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = \begin{bmatrix} \hat{A}_1 \\ \hat{B}_1 \end{bmatrix}$$

where  $\hat{A}_1$  and  $\hat{B}_1$  are upper triangular. The procedure for introducing these zero elements in  $A_1$  and  $B_1$  through the use of plane rotations is well known. It is also well known that zeros can be introduced with hyperbolic rotations

$$\frac{1}{\sqrt{1 - \rho^2}} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \text{sign}(a)\sqrt{a^2 - b^2} \\ 0 \end{bmatrix},$$

for  $\rho = -b/a$ . If  $\hat{A}_1(:, 1) = ae_1$  and  $\hat{B}_1 = be_1$ , then this transformation can be applied to transform  $\hat{A}_1$  and  $\hat{B}_1$  to  $A_2$  and  $B_2$  with  $A_2(:, 1) = \text{sign}(a)\sqrt{a^2 - b^2}e_1$  and  $B_2(:, 1) = 0$ . This process can be repeated on the submatrices  $A_2(2 : n, 2 : n)$  and  $B_2(:, 2 : n)$  to successively zero the columns of  $B$ .

Thus, if  $A$  and  $B$  are  $n \times n$ , we have

$$H = U_1 H_1 U_2 H_2 \cdots U_n H_n \quad (3)$$

satisfying (1). This method has two properties which are characteristic of all practical methods for computing  $H$ . First, although each  $U_i$  may represent multiple plane rotations, each  $H_i$  is a single hyperbolic transformation represented by a single  $\rho_i$ , and thus the hyperbolic part of the transformation is represented by exactly  $n$  reflection coefficients. Second, the  $H_i$  and the  $U_i$  do not have any effect on  $A(j, :)$  for  $j < i$ . After hyperbolic transformation  $i$ , the  $i$  row of  $\hat{A}$  is determined. These two assumptions will be useful in characterizing the  $\rho_i$  which can occur in factorizations, of the form (3), of all  $\Sigma$ -orthogonal  $H$  which satisfy (1). The first step to this characterization involves the introduction of a canonical decomposition of  $H$ .

Before proceeding further, we examine the assumption that  $B$  is the same size as  $A$ . This really represents no significant loss of generality. If  $B$  is  $k \times n$  for  $k > n$ , then merely computing the  $QR$  decomposition of  $B$  suffices to reduce  $B$  to a square matrix. All of the quantities which we will examine in this paper are invariant under this transformation and they all apply immediately to the case of  $k \times n$   $B$  with  $k > n$ . We will make no assumption about the rank of  $B$ . Consequently, merely by adding zero rows to  $B$ , all results apply immediately to the case of  $k \times n$   $B$  for  $k < n$ .

However, stronger results may be proven for rank deficient  $B$ . In particular, if the rank of  $B$  is  $k$ , then it is possible to compute  $H$  using  $k$  hyperbolic rotations rather than  $n$ . The decomposition to be introduced in the next section provides an immediate proof of this assertion. We will take note of the applicable results for rank  $k$   $B$  while deriving our results for the rank  $n$  case.

## 2 Decomposition of a $\Sigma$ -Orthogonal Matrix

The main observation of this section is the following theorem, which gives a generalization of the CS decomposition of a partitioned orthogonal matrix,[5], to a partitioned  $\Sigma$ -orthogonal matrix. A more general form of the decomposition, in which  $A$  and  $B$  are not constrained to have the same number

of rows appears in [3]. For the purposes of this paper, it is not necessary to use this higher degree of generality.

**Theorem 1** *Every  $\Sigma$ -orthogonal  $H$  can be decomposed uniquely (except for sign changes and permutations in the orthogonal matrices and for equivalent orthogonal transformations operating on parts of the  $U$  and  $V$  matrices corresponding to equal values of  $P = \Sigma_B \Sigma_A^{-1}$ ) as*

$$H = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} (\Sigma_A^2 - \Sigma_B^2)^{-1/2} & 0 \\ 0 & (\Sigma_A^2 - \Sigma_B^2)^{-1/2} \end{bmatrix} \begin{bmatrix} \Sigma_A & \Sigma_B \\ \Sigma_B & \Sigma_A \end{bmatrix} \begin{bmatrix} V_A^T & 0 \\ 0 & V_B^T \end{bmatrix} \quad (4)$$

where  $U_A$ ,  $U_B$ ,  $V_A$  and  $V_B$  are orthogonal,  $\Sigma_A$  and  $\Sigma_B$  are diagonal with elements bounded in magnitude by one and the diagonal elements of  $\Sigma_A$  are strictly greater in magnitude than those of  $\Sigma_B$ .

**Proof:** Let the singular value decomposition of  $H_{11}$  be

$$H_{11} = U_A D_A V_A^T.$$

Equation (2) implies that

$$D_A^2 - V_A^T H_{21}^T H_{21} V_A = I$$

so that  $V_A^T H_{21}^T H_{21} V_A$  must be diagonal and the singular value decomposition of  $H_{21}$  must have the form

$$H_{21} = U_B (D_A^2 - I)^{1/2} V_A^T = U_B D_B V_A^T.$$

Similarly, since  $H^T$  will also be  $\Sigma$ -orthogonal,

$$D_A^2 - U_A H_{12} H_{12}^T U_A^T = I,$$

and the singular value decomposition of  $H_{12}$  will have the form

$$H_{12} = U_A D_B V_B^T.$$

In a similar manner, the two relations

$$D_B^2 - V_B^T H_{22}^T H_{22} V_B = -I$$

and

$$D_B^2 - U_B H_{22} H_{22}^T U_B^T = -I$$

imply that the singular value decomposition of  $H_{22}$  is

$$H_{22} = U_B(I + D_B^2)^{1/2}V_B^T = U_B D_A V_B^T.$$

Thus

$$H = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} D_A & D_B \\ D_B & D_A \end{bmatrix} \begin{bmatrix} V_A^T & 0 \\ 0 & V_B^T \end{bmatrix}.$$

The representation of  $D_A$  and  $D_B$  in terms of  $\Sigma_A$  and  $\Sigma_B$  with the stated properties follows from consideration of  $2 \times 2$  hyperbolic transformations and the fact that  $D_A^2 - I = D_B^2$ . The qualification of the uniqueness claim follows from the qualification of the uniqueness of the singular value decomposition.

■

Although the orthogonal analog is well known, it is worth dealing with a point which may be less familiar in the  $\Sigma$ -orthogonal context. In particular, we wish to understand the extent to which  $A$ ,  $B$  and (1) determine  $H$ . If the matrix  $\hat{A}$  is triangular with positive diagonal elements, then it is the Cholesky factor of  $A^T A - B^T B$  and it is unique. Without the triangularity, the only degree of freedom in choosing a symmetric factor of  $A^T A - B^T B$  is captured in the orthogonal transformation  $\hat{U}$ , for which  $A^T A - B^T B = (\hat{U}\hat{A})^T(\hat{U}\hat{A})$ .

The inverse of a  $\Sigma$ -orthogonal matrix,  $H^T$ , can easily be verified to be  $\Sigma H \Sigma$ . From this we find

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} H_{11} \\ -H_{21} \end{bmatrix} \hat{U}\hat{A}.$$

or

$$\begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} = \Sigma \begin{bmatrix} A \\ B \end{bmatrix} \hat{A}^{-1} \hat{U}^T.$$

Thus the only freedom in choosing  $H_{11}$  and  $H_{21}$  is in the transformation  $\hat{U}$ . It is also not difficult to verify that  $H_{12}$  and  $H_{22}$  are completely determined except for  $H_{12}\hat{V}^T$  and  $H_{22}\hat{V}^T$  for some orthogonal  $\hat{V}$ .

### 3 The Orthogonal Case

Although the difference between the orthogonal and  $\Sigma$ -orthogonal cases is not great, the existence of a canonical decomposition is more widely recognized for orthogonal matrices. A goal of this paper is to develop an optimality result for the reflection coefficients associated with the canonical decomposition of a  $\Sigma$ -orthogonal matrix. However, for the purpose of clarity, it will be natural to develop the analogous results for the orthogonal case.

The fundamental decomposition is a special case of the CS decomposition and it has its roots in [2] and in [4].

Suppose we wish to compute an orthogonal  $Q$  such that

$$Q^T \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \hat{A} \\ 0 \end{bmatrix} \quad (5)$$

but that we wish to compute  $Q$  in such a manner that we use only  $n$  plane rotations acting between  $A$  and  $B$ . We assume that  $A^T A + B^T B$  has full rank. The situation is analogous to the factorization (3), but with each  $H$  replaced by a plane rotation  $G$ ,

$$Q = U_1 G_1 U_2 G_2 \cdots U_n G_n. \quad (6)$$

As before, the orthogonal  $U_i$  act on  $A$  and  $B$  separately while the  $G_i$  represent a single rotation between a row of  $A$  and a row of  $B$ .

As with the  $\Sigma$ -orthogonal case, it can be shown that the constraint in (5) for  $A$  and  $B$  for which  $A^T A + B^T B$  has full rank determines  $Q$  uniquely with the exception of a possible block orthogonal transformation

$$\hat{Q} = Q \begin{bmatrix} V_A^T & 0 \\ 0 & V_B^T \end{bmatrix}.$$

The transformation  $Q^T$  maps the subspace

$$\text{range} \begin{bmatrix} A \\ B \end{bmatrix} \quad (7)$$

to the subspace

$$\text{range} \begin{bmatrix} I \\ 0 \end{bmatrix}. \quad (8)$$

The notion of canonical angles between subspaces is well known, [2]. By introducing a set of canonical angles between these two subspaces, we gain some understanding of the representation of  $Q$  in terms of the angles associated with the  $G_i$ .

The fundamental theorem is the following special case of the CS decomposition, [5].

**Theorem 2** *A partitioned orthogonal matrix  $Q$  can be decomposed as*

$$Q = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} C & -S \\ S & C \end{bmatrix} \begin{bmatrix} V_A^T & 0 \\ 0 & V_B^T \end{bmatrix}$$

for orthogonal  $U_A$ ,  $U_B$ ,  $V_A$  and  $V_B$  and where  $C$  and  $S$  are real and diagonal, satisfying  $C^2 + S^2 = I$ .

If (5) is satisfied, then the angles for which the diagonal elements of  $C$  are the cosines are known as the *canonical angles* between the subspaces given by (7) and (8). The decomposition given in Theorem 2 is a special case of a decomposition of the form (6) in which  $U_2 = U_3 = \dots = U_n = I$ . In this special case, the  $G_i$  all commute and can be ordered so that  $G_i$  acts on the  $i$ th row of  $A$  and the  $i$ th row of  $B$ . The angles associated with the CS decomposition have an optimality property among a specific class of decompositions of  $Q$  into the form (6). The property is similar to the following theorem, taken from [2].

**Theorem 3** *If  $W^T$  is an orthogonal transformation mapping the subspace (7) to the subspace (8) and if the  $s_i$  are the sines of the canonical angles associated with the two subspaces and if  $w_i$  is an orthonormal basis for the subspace (7) then*

$$\sum_{k=1}^n \sin^2 \angle(w_k, W^T w_k) \geq \sum_{k=1}^n s_k^2.$$

The transformation,  $W$ , which achieves optimality is the *direct rotation*,

$$W = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} C & -S \\ S & C \end{bmatrix} \begin{bmatrix} U_A^T & 0 \\ 0 & U_B^T \end{bmatrix}$$

and the vectors  $w_i$  are given by

$$w_i = \begin{bmatrix} U_A & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} -s_i e_i \\ c_i e_i \end{bmatrix}.$$

The result we would like to prove is that an inequality similar to that of Theorem 3 holds for the angles associated with the factorization (6). Unfortunately, without further constraints on how the factorization is computed, this will not generally be true. The missing condition is that the rows of  $\hat{A}$  in (5) be computed sequentially as described in Section 1. As already noted, this constraint is satisfied by the triangularization procedure. Further, there is no real loss of generality in the assumption about the order in which the rows of  $\hat{A}$  are computed—all that is required is that a new row of  $\hat{A}$  be established with each  $G_i$  and that later transformations do not act on this row. The theorem is as follows.

**Theorem 4** *Assume that  $A^T A + B^T B$  has full rank. If an orthogonal  $Q$ , satisfying (5) and factored as (6), is computed using a scheme which computes the rows of  $\hat{A}$  sequentially as described in Section 1, but with the hyperbolic transformations replaced by plane rotations, then the angles associated*

with the plane rotations will have sines,  $\hat{s}_k$ , which satisfy

$$\sum_{k=1}^n \hat{s}_k^2 \geq \sum_{k=1}^n s_k^2.$$

The values  $s_k$  are sines associated with the canonical angles between the subspaces given in (7) and (8).

**Proof:** The proof is inductive. Assume without loss of generality that each  $G_i$  acts on row  $i$  of  $A$  and row 1 of  $B$ . The uniqueness up to block diagonal orthogonal transformations imposed by the full rank condition together with the CS decomposition and the fact that the Frobenius norm is unitarily invariant imply that for any  $Q$  satisfying (5),

$$\|Q_{21}\|_F^2 = \sum_{j=1}^n s_j^2.$$

In the interest of finding a basis which clearly illustrates the actions of  $Q$ , we can choose invertible  $X$  so that

$$Q^T \begin{bmatrix} A \\ B \end{bmatrix} X = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

In fact, for the rest of the proof, we will assume that the transformation  $X$  has been applied to  $A$  and  $B$  so that

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix}$$

and we can assume that  $\|B\|_F^2 = \sum_j s_j^2$ . We will use  $Q_{i1}$  to refer to the initial matrices, while using  $A$  and  $B$  more loosely to describe various stages after transformations have been applied.

To allow the application of induction, we will prove the more general assertion that if  $Q$  is  $(m+n) \times (m+n)$ ,

$$Q = U_1 G_1 U_2 G_2 \cdots U_n G_n.$$

and  $Q_{11}$  is  $n \times n$ , then

$$\sum_{j=1}^n \hat{s}_j^2 \geq \|Q_{21}\|_F^2.$$

The induction is on  $n$ . The case  $n = 1$  is obvious. The assumption that  $G_j$  and  $U_j$  do not act on row  $i$  of  $A$  for  $j > i$  implies that after the application

of  $U_1^T$  and  $G_1^T$  the first column of  $A$  will be  $e_1$  and the first column of  $B$  will be zero. Similarly, after the application of  $U_1^T$  only, the first columns of  $A$  and  $B$  will be multiples of  $e_1$ . Thus, the cosine and sine associated with  $G_1$  will be computed to satisfy

$$\begin{bmatrix} \hat{c}_1 & -\hat{s}_1 \\ \hat{s}_1 & \hat{c}_1 \end{bmatrix} \begin{bmatrix} \|A(:,1)\| \\ \|B(:,1)\| \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

This gives

$$\hat{s}_1^2 = \|B(:,1)\|^2 = \|Q_{21}(:,1)\|^2.$$

Since orthogonality guarantees that  $G_1^T$  introduces zeros into the first row of  $A(:,2:n)$ , the application of  $G_1^T$  will only increase  $\|B(:,2:n)\|$ . Since later transformations will not act on the first row of  $A$ , the induction hypothesis guarantees that the sum of the squares of the later sines will be greater than this new, increased  $\|B(:,2:n)\|_F^2$  and consequently

$$\sum_{j=1}^n s_j^2 \geq \|Q_{21}\|_F^2.$$

This completes the proof. ■

It is fairly simple to check, even by constructing random matrices from factorizations of the form (3), that this result depends on the sequential computation of the rows of  $\hat{A}$  in an essential way. The significance of the theorem is that it gives a lower bound on the amount of “action” required by Givens rotations acting between  $A$  and  $B$  to zero  $B$  through a triangularization procedure, with or without any transformation from the right.

To put things in perspective, it is worth drawing a comparison with Theorem 3. Despite similarities, the sines which the two results bound in terms of the canonical sines are distinct. Suppose that  $A$  and  $B$  are such that we are applying the triangularization procedure described in Section 1 to a matrix with orthonormal columns. Theorem 3 and Theorem 4 can both be applied without difficulty, but they provide bounds on different quantities. In the former case, it is a bound on the sines associated with angles between the columns of the original orthogonal matrix and its transformed version. In the latter it is a bound on the sines computed by the triangularization procedure. In effect, we have shifted a theorem which describes the action of  $Q$  on a particular basis for a subspace to a theorem which describes the representation of  $Q$ .

The sines associated with the representation in Equation (3) are quite different from those involved in Theorem 3. The following theorem illus-

trates a property which holds for sines from Equation (3) but which does not hold for the sines involved in Theorem 3.

**Theorem 5** *Along with the assumptions in the statement of Theorem 4 if we also assume that*

$$|s_1| \geq |s_2| \geq \cdots \geq |s_n|$$

then

$$|s_1| \geq |\hat{s}_k| \geq |s_n|$$

for  $k = 1, 2, \dots, n$ .

**Proof:** Without loss of generality, we assume that  $G_j^T$  acts on row  $j$  of  $A$  and row  $j$  of  $B$ . Let  $X$  be defined as in the proof of Theorem 4 so that

$$Q^T \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix}$$

and, consequently,

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix}.$$

The orthogonality introduced by  $X$  and the fact that none of the other transformations after  $G_1^T$  will act on row 1 of  $A$  imply that

$$U_1^T \begin{bmatrix} A \\ B \end{bmatrix} X = \begin{bmatrix} a_{11} & a_{12}^T \\ 0 & A_{22} \\ b_{11} & b_{12}^T \\ 0 & B_{22} \end{bmatrix}$$

with  $a_{11}^2 + b_{11}^2 = 1$ . The sine and cosine associated with  $G_1^T$  must be computed to satisfy

$$\begin{bmatrix} \hat{c}_1 & -\hat{s}_1 \\ \hat{s}_1 & \hat{c}_1 \end{bmatrix} \begin{bmatrix} a_{11} \\ b_{11} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ 0 \end{bmatrix}.$$

If we let  $\hat{A}X$  and  $\hat{B}X$  represent the matrices after  $G_1^T$  is applied, then, using orthogonality,

$$\begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix} X = \begin{bmatrix} 1 & 0 \\ 0 & A_{22} \\ 0 & b_{12}^T/\hat{c}_1 \\ 0 & B_{22} \end{bmatrix}.$$

We have

$$\sigma_n(BX) \leq |\hat{s}_1| = |b_{11}| \leq \sigma_1(BX)$$

But the singular values of  $BX$  are the  $s_k$  so the desired inequalities clearly hold for  $\hat{s}_1$ . The proof can be completed by induction by showing that

$$\sigma_{n-1}(\hat{B}X) \geq \sigma_n(BX)$$

and

$$\sigma_1(\hat{B}X) \leq \sigma_1(BX).$$

The first inequality follows easily from the fact that deleting a column only increases the smallest singular value and from the fact that  $1/c_1 \geq 1$ . To prove the second inequality note that

$$(\hat{A}X)^T(\hat{A}X) + (\hat{B}X)^T(\hat{B}X) = I$$

and

$$(AX)^T(AX) + (BX)^T(BX) = I$$

so that

$$(\hat{B}X)^T(\hat{B}X) = \begin{bmatrix} 0 & 0 \\ 0 & I - A_{22}^T A_{22} \end{bmatrix}$$

and

$$(BX)^T(BX) = I - (AX)^T(AX).$$

But since  $\sigma_1(AX) \geq \sigma_1(A_{22})$  we have  $\sigma_{n-1}(\hat{B}X) \geq \sigma_n(BX)$ . ■

The results of Theorem 5 clearly show that the sort of optimality expressed in Theorem 4 is different from that of Theorem 3. The sines associated with Theorem 3 do not have to satisfy the bounds of Theorem 5. Consequently, Theorem 4 can be viewed as a distinct way in which the CS-decomposition provides an optimal mapping. Suppose  $U_1$  and  $V_1$  have orthonormal columns and span two  $n$  dimensional subspaces. Let  $U_2$  have orthonormal columns with  $U_2^T U_1 = 0$  and let

$$Q^T \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} V_1 = \begin{bmatrix} \hat{A} \\ 0 \end{bmatrix}.$$

Then the theorems of this section give bounds on the sines of transformations which combine vectors between the two parts of  $U$  to produce vectors in  $\text{span}(V_1)$ . In a sense, this can be viewed as dual to the result in Theorem 3, which assumes that  $Q$  maps  $U_1$  onto  $\text{span}(V_1)$  and then bounds the sines associated with the angle between a column of  $U_1$  and its image under  $Q$ .

## 4 The $\Sigma$ -Orthogonal Case

The  $\Sigma$ -orthogonal case is somewhat less intuitive, if only because of a lesser degree of familiarity. However, the proof of an optimality result for reflection coefficients associated with the canonical decomposition is directly analogous to the proof of Theorem 4. The result is the following theorem.

**Theorem 6** *Assume that  $A$  and  $B$  are such that  $A^T A - B^T B$  has full rank. If a  $\Sigma$ -orthogonal  $H$ , satisfying (1) and factored as (3), is computed using a scheme which computes the rows of  $\hat{A}$  sequentially as described in Section 1, then the reflection coefficients,  $\hat{\rho}_k$  associated with the hyperbolic transformations,  $H_k$ , will satisfy*

$$\sum_{k=1}^n \hat{\rho}_k^2 \geq \sum_{k=1}^n \rho_k^2.$$

*The values  $\rho_k$  are reflection coefficients associated with the canonical decomposition of  $H$ , and are equal to the diagonal elements of  $\Sigma_B \Sigma_A^{-1}$ .*

**Proof:** The proof is inductive and is very similar to the proof of Theorem 4. Assume without loss of generality that each  $H_i$  acts on row  $i$  of  $A$  and row 1 of  $B$ . The uniqueness up to block orthogonal transformations imposed by the full rank condition together with the canonical decomposition and the fact that the Frobenius norm is unitarily invariant imply that for any  $H$  satisfying (1),

$$\|H_{21} H_{11}^{-1}\|_F^2 = \sum_{j=1}^n \rho_j^2.$$

In the interest of finding a basis which clearly illustrates the actions of  $H$ , we can choose invertible  $X$  so that

$$H^T \begin{bmatrix} A \\ B \end{bmatrix} X = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

As before, for the rest of the proof, we will assume that the transformation  $X$  has been applied to  $A$  and  $B$  so that

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix}$$

Thus we can assume that initially  $\|BA^{-1}\|_F^2 = \sum_j \rho_j^2$ . We will use  $H_{i1}$  to refer to the initial matrices, while using  $A$  and  $B$  more loosely to describe stages after transformations have been applied to  $A$  and  $B$ .

Again, to allow the application of induction, we will prove the more general assertion that if  $H$  is  $(m+n) \times (m+n)$ ,

$$H = U_1 H_1 U_2 H_2 \cdots U_n H_n.$$

and  $H_{11}$  is  $n \times n$ , then

$$\sum_{j=1}^n \hat{\rho}_j^2 \geq \|H_{21} H_{11}^{-1}\|_F^2$$

still holds.

The induction is on  $n$ . The case  $n = 1$  is obvious. The assumption that  $H_j$  and  $U_j$  do not act on row  $i$  of  $A$  for  $j > i$  implies that after the application of  $U_1^T$  and  $H_1^T$  the first column of  $A$  will be  $e_1$  and the first column of  $B$  will be zero. Similarly, after the application of  $U_1^T$  only, the first columns of  $A$  and  $B$  will be multiples of  $e_1$ . Assume that after  $U_1^T$  has been applied, we have

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12}^T \\ 0 & A_{22} \\ b_{22} & b_{12}^T \\ 0 & B_{22} \end{bmatrix}.$$

The application of  $U_1^T$  does not change  $\|BA^{-1}\|_F^2$  and after  $U_1^T$  has been applied

$$BA^{-1} = \begin{bmatrix} \frac{b_{11}}{a_{11}} & \left(b_{12}^T - \frac{b_{11}}{a_{11}} a_{12}^T\right) A_{22}^{-1} \\ 0 & B_{22} A_{22}^{-1} \end{bmatrix} = \begin{bmatrix} -\hat{\rho}_1 & (b_{12}^T + \hat{\rho}_1 a_{12}^T) A_{22}^{-1} \\ 0 & B_{22} A_{22}^{-1} \end{bmatrix}.$$

After  $H_1^T$  is applied,

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}^2 - b_{11}^2} & 0 \\ 0 & A_{22} \\ 0 & \frac{(b_{12}^T + \hat{\rho}_1 a_{12}^T)}{\sqrt{1 - \hat{\rho}_1^2}} \\ 0 & B_{22} \end{bmatrix}$$

Clearly,

$$\left\| \begin{bmatrix} \frac{1}{\sqrt{1 - \hat{\rho}_1^2}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} b_{12}^T + \hat{\rho}_1 a_{12}^T \\ B_{22} \end{bmatrix} A_{22}^{-1} \right\|_F^2 \geq \left\| \begin{bmatrix} b_{12}^T + \hat{\rho}_1 a_{12}^T \\ B_{22} \end{bmatrix} A_{22}^{-1} \right\|_F^2.$$

Consequently,

$$\hat{\rho}_1^2 + \left\| \begin{bmatrix} \frac{b_{12}^T + \hat{\rho}_1 a_{12}^T}{\sqrt{1 - \hat{\rho}_1^2}} \\ B_{22} \end{bmatrix} A_{22}^{-1} \right\|_F^2 \geq \|BA^{-1}\|_F^2$$

and the theorem follows from the induction hypothesis upon noting that we are left with the smaller problem of computing reflection coefficients from  $A_{22}$  and

$$\begin{bmatrix} \frac{b_{12}^T + \hat{\rho}_1 a_{12}^T}{\sqrt{1 - \hat{\rho}_1^2}} \\ B_{22} \end{bmatrix}.$$

This completes the induction step and the proof. ■

It is possible to prove other properties of the reflection coefficients associated with the canonical form. Of particular interest is the following result.

**Theorem 7** *Along with the assumptions in the statement of Theorem 6 if we also assume that*

$$|\rho_1| \geq |\rho_2| \geq \cdots \geq |\rho_n|.$$

then

$$|\rho_1| \geq |\hat{\rho}_k| \geq |\rho_n|$$

for  $k = 1, 2, \dots, n$ .

**Proof:** Without loss of generality, we assume  $A$  and  $B$  have been multiplied from the right by  $X$  as in the proof of Theorem 6 and that they have been multiplied from the left by the appropriate orthogonal transformations for the first stage of triangularization. Clearly  $|\rho_1| = \sigma_1(BA^{-1})$  and  $|\rho_n| = \sigma_n(BA^{-1})$ . Since

$$BA^{-1} = \begin{bmatrix} \frac{b_{11}}{a_{11}} & \left( b_{12}^T - \frac{b_{11}}{a_{11}} a_{12}^T \right) A_{22}^{-1} \\ 0 & B_{22} A_{22}^{-1} \end{bmatrix} = \begin{bmatrix} -\hat{\rho}_1 & (b_{12}^T + \hat{\rho}_1 a_{12}^T) A_{22}^{-1} \\ 0 & B_{22} A_{22}^{-1} \end{bmatrix}$$

we see that  $-\hat{\rho}_1$  is an eigenvalue of  $BA^{-1}$ . Consequently  $|\hat{\rho}_1| \leq |\rho_1|$ . This is the starting point for an inductive proof. If  $\hat{A}$  and  $\hat{B}$  are the matrices after the hyperbolic transformation has been applied, then  $A^T A - B^T B = \hat{A}^T \hat{A} - \hat{B}^T \hat{B} = I$ . From this we see that

$$(BA^{-1})^T (BA^{-1}) = I - A^{-T} A^{-1}$$

and

$$(\hat{B}\hat{A}^{-1})^T(\hat{B}\hat{A}^{-1}) = I - \hat{A}^{-T}\hat{A}^{-1}.$$

Since the smallest eigenvalue of  $\hat{A}^{-T}\hat{A}^{-1}$  is  $1/\|\hat{A}\|^2$  and since a similar equality holds for  $A$ , it is easily verified from

$$\hat{A} = \begin{bmatrix} \sqrt{1-\rho^2}a_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

that the smallest eigenvalue of  $\hat{A}^{-T}\hat{A}^{-1}$  is larger than the smallest eigenvalue of  $A^{-T}A^{-1}$ . This implies  $\|\hat{B}\hat{A}^{-1}\| < \|BA^{-1}\|$  and consequently, by induction,  $|\hat{\rho}_k| \leq |\rho_1|$ .

The other side of the inequality is similarly easy to prove. In the same manner as before, we can conclude that  $|\hat{\rho}_1| \geq |\rho_n|$ . But

$$\hat{B}\hat{A}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{1-\hat{\rho}_1^2}} & 0 \\ 0 & I \end{bmatrix} [ 0 \quad (BA^{-1})(:, 2:n) ]$$

which implies that  $\sigma_{n-1}(\hat{B}\hat{A}^{-1}) > \sigma_n(BA^{-1})$ . Since we now effectively have a problem of size  $(n-1) \times (n-1)$ , this is the inductive step which proves that  $\hat{\rho}_k > \rho_n$ . ■

As suggested earlier, the significance of these result is broader than just providing an inequality which says that the canonical reflection coefficients are better in a sum-of-squares sense than those of the obvious triangularization procedure. The results in this section give a characterization of the range of possible reflection coefficients for schemes for computing  $H$  which are more general than triangularization. Since, for an appropriate transformation from the right of  $A$  and  $B$ , the triangularization procedure will compute the canonical reflection coefficients, we effectively have bounds on the the range of reflection coefficients which are possible when triangularizing after transforming from the right. As with the orthogonal case, the sequential computation of the rows of  $\hat{A}$  is essential for the result to hold.

The fact that we have decoupled the  $n$  reflection coefficients with orthogonal transformations shows that the canonical reflection coefficients reveal the inherent condition of this problem. Theorem 7 is reassuring in that it implies that it is not possible to introduce artificially ill-conditioned transformations through a poor choice of a method for computing  $H$ . However, it does not really go far enough: in most schemes for computing  $H$ , the hyperbolic transformations will not be decoupled as they are in the decomposition of  $H$ .

## 5 Condition Numbers of Block Toeplitz Matrices

For an  $n \times n$  symmetric positive definite Toeplitz matrix,  $T$ , it is shown in [1] that

$$\prod_{k=1}^n \frac{1 + \rho_k}{1 - \rho_k^2} \leq \|T^{-1}\|_1 \leq \prod_{k=1}^{n-1} \frac{(1 + |\rho_k|)^2}{1 - \rho_k^2}, \quad (9)$$

where  $\rho_k$  are the reflection coefficients of  $T$ . We assume that  $T$  is scaled to have ones on the diagonal, so that the bounds can be immediately translated into bounds on the condition number of  $T$ . Although these bounds can give a very broad range for  $\|T^{-1}\|$ , it is immediately clear that the presence of a reflection coefficient which is very close to 1 yields an ill-conditioned Toeplitz matrix. It is in the other direction that problems arise: a Toeplitz matrix can be very ill-conditioned without revealing it in the form of a single very bad reflection coefficient. The cumulative effect of a large number of moderately bad reflection coefficients can result in an extremely ill-conditioned Toeplitz matrix.

For block Toeplitz matrices, the situation is even worse. There are likely to be more reflection coefficients and their relation to the condition number is even more dubious. Without claiming to have put the notion of estimating condition numbers from reflection coefficients on a safe footing, the point of this section is to show that the decoupling provided by the canonical decomposition makes the canonical reflection coefficients more appropriate for this task.

Assume we are dealing with a symmetric positive definite Toeplitz matrix,

$$T_N = \begin{bmatrix} B_0 & B_1 & B_2 & \cdots & B_{N-1} \\ B_1 & B_0 & B_1 & \ddots & \vdots \\ B_2 & B_1 & \ddots & \ddots & B_2 \\ \vdots & \ddots & \ddots & \ddots & B_1 \\ B_{N-1} & \cdots & B_2 & B_1 & B_0 \end{bmatrix}$$

with symmetric blocks,  $B_k$ . If

$$R_N = [ B_1 \quad B_2 \quad \cdots \quad B_N ]^T$$

then triangular factors of  $T_N^{-1}$  can be obtained from the solutions  $A_k$  to  $T_k A_k = R_k$  for  $k = 1, 2, \dots, N - 1$ . In particular, if  $A$  is defined as a lower

triangular matrix by

$$A(:, (k-1)n+1 : kn) = \begin{bmatrix} 0_{(k-1)n \times n} \\ I_{n \times n} \\ -A_{N-k} \end{bmatrix}$$

for  $k = 1, \dots, N$  then  $T_k A_k = R_k$  implies that  $T_N A$  will be upper block triangular. Consequently, by symmetry, we see that it is possible to express  $T_N^{-1}$  as

$$T_N^{-1} = ADA^T$$

where  $D$  is a block diagonal matrix with  $n \times n$  symmetric positive definite diagonal blocks.

The algorithm which computes such a factorization is a block version of the Levinson algorithm. This is not a new idea and the derivation of the algorithm is so similar to the traditional Levinson algorithm that we will only summarize the results before applying them to the problem at hand.

We partition  $A_k$  as

$$\begin{bmatrix} \hat{A}_k \\ \hat{P}_k \end{bmatrix}$$

where  $\hat{P}_k$  is  $n \times n$  and let  $J$  be

$$J = \begin{bmatrix} & & & I \\ & & I & \\ & \dots & & \\ I & & & \end{bmatrix}.$$

Using the fact that  $T_N J = J T_N$  and introducing the partitioning

$$T_{k+1} = \begin{bmatrix} T_k & J R_k \\ R_k^T J & B_0 \end{bmatrix}$$

we find that it is possible to recursively compute  $A_{k+1}$  from  $A_k$ ,

$$\hat{A}_{k+1} = A_k - J A_k \hat{P}_{k+1}, \quad (10)$$

$$\hat{P}_{k+1} = E_k^{-1} (B_{k+1} - R_k^T J A_k), \quad (11)$$

where

$$E_k = (B_0 - R_{k-1}^T A_{k-1})$$

is easily computable from

$$E_{k+1} = E_k - \hat{P}_{k+1}^T E_k \hat{P}_{k+1}. \quad (12)$$

The proof of these results is directly analogous to the derivation of the recursions for the ordinary Levinson algorithm.

The matrices  $E_k$  are symmetric positive definite and

$$D = \text{diag}(E_{N-1}^{-1}, E_{N-2}^{-1}, \dots, E_0^{-1}).$$

Let  $E_k = C_k^T C_k$  for upper triangular  $C_k$ . Equation (12) is equivalent to

$$C_{k+1}^T C_{k+1} = C_k^T C_k - (C_k \hat{P}_{k+1})^T (C_k \hat{P}_{k+1}) = C_k^T (I - P_{k+1}^T P_{k+1}) C_k$$

where  $P_{k+1} = C_k \hat{P}_{k+1} C_k^{-1}$ . It is easily verified from (11) and

$$R_k^T J A_k = R_k^T J T_k^{-1} R_k = R_k^T T_k^{-1} J R_k = (R_k^T J A_k)^T$$

that  $P_{k+1}$  is symmetric. This is a downdating problem. Using the notation of (4) and letting  $U_{k+1}$  be the Cholesky factor of  $I - P_{k+1}^T P_{k+1}$  we have

$$U_{k+1} = U_A \Sigma_A^{-1} (\Sigma_A^2 - \Sigma_B^2)^{1/2} V_A^T$$

and

$$C_{k+1} = U_{k+1} C_k$$

where the reflection coefficients correspond to those from the  $\Sigma$ -orthogonal transformation computed by the Schur algorithm.

This provides a bound on the growth in the block diagonal elements of the triangular factor of  $T_N^{-1}$ , connecting the condition of  $T$  with the worst canonical reflection coefficient. We can tighten the connection by modifying (10). So far, the symmetry of  $P$  shows that for diagonal elements,  $E_k$ , the eigenvalues of  $\hat{P}$ , rather than the singular values, are what really matter. In (10) there is an apparent dependence on  $\|\hat{P}\|$ , but this can be dealt with easily. From (10)

$$\hat{A}_{k+1} C_k^{-1} = (A_k C_k^{-1}) - J(A_k C_k^{-1}) P_{k+1}.$$

But

$$C_{k+1}^{-1} = C_k U_{k+1}^{-1}$$

so

$$(\hat{A}_{k+1} C_{k+1}^{-1}) = (A_k C_k^{-1}) U_{k+1}^{-1} - J(A_k C_k^{-1}) P_{k+1} U_{k+1}^{-1}.$$

and

$$(\hat{P}_{k+1} C_{k+1}^{-1}) = C_k^{-1} P_{k+1} U_{k+1}^{-1}.$$

The effects of  $P_{k+1}$  and  $U_{k+1}$  in these equations can be easily observed by noting that

$$V_A^T U_{k+1}^{-1} U_k = (I - \Sigma_A^{-2} \Sigma_B^{-2})^{-1/2}$$

and

$$V_A^T P_{k+1} U_{k+1}^{-1} U_A = \Sigma_A^{-1} \Sigma_B (I - \Sigma_A^{-2} \Sigma_B^2)^{-1/2}.$$

Since  $A_k C_k^{-1}$  is just a column of the triangular factor of  $T_N^{-1}$ , we see that the growth in the norms of these columns will be determined by the largest diagonal element of  $\Sigma_A^{-1} \Sigma_B$ . From this result, we can derive bounds which are essentially identical to the upper bound in (9). Unfortunately, it is not obvious how best to do this; the most appropriate choice of norm is not completely clear. We will rest matters here with having revealed that the a single reflection coefficient can characterize the growth of an entire block of columns in the Levinson algorithm.

## 6 Summary

In this paper we have presented a decomposition of a  $\Sigma$ -orthogonal matrix and analyzed its properties which parallel those of the CS decomposition. In addition, we have addressed the issue of conditioning. The CS decomposition reveals canonical rotations which are always well conditioned, but the  $\Sigma$ -orthogonal equivalent reveals information which is relevant to understanding the conditioning of downdating problems and factorization of block Toeplitz matrices. Since the elementary hyperbolic transformations are completely decoupled by orthogonal transformations, the norms of these transformations reveal the inherent condition of the problem. In addition to this, we have shown that these reflection coefficients can be used to characterize the range of reflection coefficients which are achievable by other approaches to solving the problem.

## References

- [1] G. CYBENKO, *The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM Journal of Scientific and Statistical Computing, 1 (1980), pp. 303–319.
- [2] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation, III*, SIAM Journal of Numerical Analysis, 7 (1970), pp. 1–46.

- [3] E. J. GRIMME, D. C. SORENSON, AND P. VAN DOOREN, *Model reduction of state space systems via an implicitly restarted Lanczos method*, Numerical Algorithms, 12 (1995), pp. 1–31.
- [4] G. W. STEWART, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Review, 19 (1977), pp. 634–662.
- [5] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
- [6] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.