

# On Swapping Diagonal Blocks in Real Schur Form\*

Zhaojun Bai<sup>†</sup> and James W. Demmel<sup>‡</sup>

## Abstract

We discuss a new version of an existing algorithm for reordering the eigenvalues on the diagonal of a matrix in real Schur form by performing an orthogonal similarity transformation. A detailed error analysis and software description are presented. Numerical examples show the superiority of our algorithm over previous algorithms.

## 1 Introduction

The problem of reordering the eigenvalues into a desired order along the (block) diagonal of a quasi-triangular real matrix arises in several applications: computing an invariant subspace corresponding to a given group of eigenvalues, estimating condition numbers for a cluster of eigenvalues or their associated invariant subspace [28], computing partial eigenvalues of a large nonsymmetric matrix by the simultaneous iteration method [14], computing matrix functions [41], solving the linear quadratic control problem [10] and so on. These problems can be solved in two phases: the first is to compute the Schur decomposition of the given matrix, and the second is to reorder a group of specified eigenvalues to appear at the upper left corner of the matrix. In this paper we describe an algorithm and its implementation for this reordering problem. The software is available in LAPACK [1], a public domain numerical linear algebra library.

Specifically, for a real matrix  $A$  there is a real orthogonal matrix  $Q$  such that

$$A = QTQ^T, \quad (1)$$

where  $T$  is a real upper quasi-triangular matrix, called the *real Schur form*. This means that  $T$  is block upper triangular with  $1 \times 1$  and  $2 \times 2$  blocks on the diagonal. The  $1 \times 1$  blocks contain the real eigenvalues of  $A$ . The eigenvalues of the  $2 \times 2$  diagonal blocks are the complex conjugate eigenvalues of  $A$ . The real Schur form may be computed using subroutine HQR from EISPACK [13] or subroutine SGEQR from LAPACK [1]). Here  $Q$  provides an orthonormal basis for the invariant subspaces of certain subsets of eigenvalues of the matrix  $A$ . If we partition  $Q$  and  $T$  conformally as

$$Q = [Q_1, Q_2], \quad T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

then from (1) we have

$$AQ_1 = Q_1T_{11} \quad (2)$$

---

\*This work was supported in part by NSF grant ASC-9005933 and DARPA grant DAAL03-91-C-0001, a subcontract from the University of Tennessee. The first author was also supported by ONR grant N00014-91-1-02963.

<sup>†</sup>Department of Mathematics, University of Kentucky, Lexington, KY 40506. (na.bai@na-net.ornl.gov)

<sup>‡</sup>Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720. (na.demmel@na-net.ornl.gov)

and hence  $Q_1$  gives an orthonormal basis for the invariant subspace of  $A$  corresponding to the eigenvalues contained in  $\eta_1 T$ .

Unfortunately, the  $T$  given by the QR algorithm will not generally contain the eigenvalues in which we are interested. We must therefore perform some further orthogonal similarities that preserve block triangular form but reorder the desired eigenvalues of  $A$  to the upper left corner of the Schur form  $T$ . The crux of such a reordering is to swap two adjacent  $1 \times 1$  or  $2 \times 2$  diagonal blocks by an orthogonal transformation. Formally, let  $A_{11}$  be a  $p \times p$  matrix,  $A_{22}$  be a  $q \times q$  matrix,  $p, q = 1$  or  $2$ ; we want to compute an orthogonal  $(p+q) \times (p+q)$  matrix  $Q$  such that

$$Q^T \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} Q = \begin{bmatrix} \tilde{A}_{22} & \tilde{A}_{12} \\ 0 & \tilde{A}_{11} \end{bmatrix}, \quad (3)$$

where  $\tilde{A}_{ii}$  is similar to  $A_{ii}$   $i = 1, 2$ , so that the eigenvalues are unchanged but their positions are exchanged along the (block) diagonal.

To this end, Stewart [15] has described an iterative algorithm for swapping consecutive  $1 \times 1$  and  $2 \times 2$  blocks of a quasi-triangular matrix, which we refer to as algorithm EXCHNG. In his method, the first block is used to determine an implicit QR shift. An arbitrary QR step is performed on both blocks to create a dense  $(p+q) \times (p+q)$  matrix. Then a sequence of QR steps using the previously determined shift is performed. Theoretically, after one step of QR iteration, the eigenvalues of the first block will emerge in the lower part. But in practice, two or even more QR iterations may still fail to reorder the eigenvalues for some hard problems. This use of QR iteration has been extended by Van Doren [19] to reordering the eigenvalues of a generalized eigenvalue problem using QZ iteration.

Another algorithm to be further developed in this paper is the so-called *direct swapping method* which was originally motivated by the work of Ruhe [12] and Dongarra, Hammarling and Wilkinson in 1983, although the paper was finished later (1991). [7] Ng and Parlett [1] also developed a program to implement the direct swapping algorithm. A similar idea has also been published by Cao and Zhang [6].

This previous work still does not solve the problems satisfactorily. The iterative swapping algorithm has the advantage of guaranteed backward stability, since it just multiplies the data by orthogonal matrices. But it may be inaccurate and even fail to reorder the eigenvalues in ill-conditioned cases. On the other hand, the direct swapping algorithm is simple and can better deal with ill-conditioned cases. But there are examples where these implementations fail to be stable.

In this paper, we further improve the direct swapping algorithm. Various strategies have been designed at each stage of the algorithm to improve its accuracy and robustness. A detailed analysis of the algorithm shows that backward instability is possible only in very ill-conditioned cases, so ill-conditioned in fact that we have been unable to construct a case where it fails. Our goal was to have an absolute stability guarantee, however; we achieved this by directly and cheaply testing for instability and rejecting a swap if it would have been unstable. This can occur only when the eigenvalues are so ill-conditioned as to be indistinguishable in a certain reasonable sense. Numerical experiments show the superiorities of our direct swapping algorithm over previous implementations.

The rest of the paper is organized as follows: §2 describes the direct swapping algorithm. The error analysis of the algorithm is carried out in §3. In §4, the software implementation and numerical experiments are reported in §5. §6 draws conclusions. All software including test software for the algorithms in this paper can be found in the LAPACK library [1].

We assume that any  $2 \times 2$  diagonal block in the quasi-triangular matrix is in standardized form

This means that its diagonal entries are equal and its off diagonal nonzero and of opposite sign:

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \alpha \end{bmatrix}, \quad \beta\gamma < 0. \quad (4)$$

For any  $2 \times 2$  block with complex conjugate eigenvalues, we can easily compute an orthogonal similarity transformation to standardize the block.

## 2 Direct Swapping Algorithm

As we described in the introduction, the crux of reordering the diagonal blocks is to interchange the consecutive diagonal blocks  $A_{11}$  and  $A_{22}$  in the following block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad (5)$$

where  $A_{11}$  is  $p \times p$ ,  $A_{22}$  is  $q \times q$ ,  $p, q = 1$  or  $2$ . Throughout this paper, we assume that  $A_{11}$  and  $A_{22}$  have no eigenvalue in common, otherwise, they need not be exchanged. It is seen that the block matrix (5) can be block diagonalized as

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} = \begin{bmatrix} I_p & -X \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I_p & X \\ 0 & I_q \end{bmatrix},$$

where  $X$  is the solution of the Sylvester equation

$$A_{11}X - XA_{22} = A_{12}. \quad (6)$$

Since it is assumed that  $A_{11}$  and  $A_{22}$  have no eigenvalue in common, the solution  $X$  exists and is unique. If we choose an orthogonal matrix  $Q$  such that

$$Q^T \begin{bmatrix} -X \\ I_q \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

and conformally partition  $Q$  in the form

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

then

$$Q^T \begin{bmatrix} -X & I_p \\ I_q & 0 \end{bmatrix} = \begin{bmatrix} R & Q_{11}^T \\ 0 & Q_{12}^T \end{bmatrix}.$$

Since both matrices on the left are invertible so are  $R$  and  $Q$ .

$$\begin{aligned} Q^T \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} Q &= Q^T \begin{bmatrix} I_p & -X \\ 0 & I_q \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I_p & X \\ 0 & I_q \end{bmatrix} Q \\ &= \begin{bmatrix} R & Q_{11}^T \\ 0 & Q_{12}^T \end{bmatrix} \begin{bmatrix} A_{22} & 0 \\ 0 & A_{11} \end{bmatrix} \begin{bmatrix} R^{-1} & -R^{-1}Q_{11}^T Q_{12}^{-T} \\ 0 & Q_{12}^{-T} \end{bmatrix} \\ &= \begin{bmatrix} RA_{22}R^{-1} & -RA_{22}R^{-1}Q_{11}^T Q_{12}^{-T} + Q_{11}^T A_{11} Q_{12}^{-T} \\ 0 & Q_{12}^T A_{11} Q_{12}^{-T} \end{bmatrix} \\ &\equiv \begin{bmatrix} \tilde{A}_{22} & \tilde{A}_{12} \\ 0 & \tilde{A}_{11} \end{bmatrix} \end{aligned}$$

where  $\tilde{A}_{ii}$  is similar to  $A_{ii}$   $i=1,2$ , so that the eigenvalues are invariant, but their positions are exchanged. Furthermore, we have the following theorem to specify such orthogonal transformation:

**Theorem 1** (Ng and Parlett [11]). *An orthogonal  $(p+q) \times (p+q)$  matrix  $Q$  swaps  $A_{11}$  and  $A_{22}$  if and only if*

$$Q^T \begin{bmatrix} -X \\ I_q \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad (7)$$

for some invertible  $q \times q$  matrix  $R$  where  $X$  is defined in (6).

In the presence of rounding errors, the biggest concern is solving the Sylvester equation (6). It could possibly be ill-conditioned if  $A_{11}$  and  $A_{22}$  have close eigenvalues. In the extreme case, if  $A_{11}$  and  $A_{22}$  have the same eigenvalues, the Sylvester equation is singular and the solution  $X$  may be infinite. To prevent possible overflow, we instead solve the equation

$$A_{11}X - XA_{22} = \gamma A_{12} \quad (8)$$

or the corresponding linear system

$$Kx = \gamma b \quad (9)$$

where  $\gamma$  is a scaling factor ( $\gamma \leq 1$ ), and  $K = \begin{bmatrix} A_{11} & \\ & -A_{22}^T \end{bmatrix} \otimes I_p$ ,  $\otimes$  is the Kronecker product,  $x = \text{col}(X)$ ,  $b = \text{col}(A_{12})$ .  $\text{col}(W)$  denotes the column vector formed by taking columns of  $W$  and stacking them atop one another from left to right. Possible overflow of  $X$  is taken care of by choosing a small scaling factor  $\gamma$ . In the extreme case, when  $A_{11}$  and  $A_{22}$  have the same eigenvalues, we choose  $\gamma = 0$ . Because the linear system (9) can only be  $1 \times 1$ ,  $2 \times 2$  or  $4 \times 4$ , it does not cost too much to use Gaussian elimination with complete pivoting to solve it with better numerical properties (in particular, the pivots are within a modest factor of the singular values of the  $4$  by  $4$  matrix, so setting tiny pivots to a chosen tiny value controls the conditioning of the system and norm of the solution). Applying standard results from [20] straightforward analysis shows that for the computed solution  $\bar{X}$  of the Sylvester equation:

$$\frac{\|\bar{X} - X\|_F}{\|X\|_F} \leq \frac{\rho \varepsilon_M (\|A_{11}\|_F + \|A_{22}\|_F)}{\text{sep}(A_{11}, A_{22})}, \quad (10)$$

where  $E = X - \bar{X}$ ,  $\rho$  is a small constant of order  $O(\epsilon_M)$ ,  $\epsilon_M$  is machine precision, and  $\text{sep}(A_{11}, A_{22}) = \sigma_{\min}(Y)$  is called the separation of the matrices  $A_{11}$  and  $A_{22}$ .

In the following error analysis of the algorithm we will see that the numerical stability is essentially governed by the residual  $Y = A_{11}\bar{X} + \bar{X}A_{22} - A_{12}$ . Applying standard error analysis of Gaussian elimination [20] have

$$\|Y\|_F = \|A_{12} - A_{11}\bar{X} + \bar{X}A_{22}\|_F \leq \rho \varepsilon_M (\|A_{11}\|_F + \|A_{22}\|_F) \|X\|_F. \quad (11)$$

Note that the bound does not involve  $\text{sep}(A_{11}, A_{22})$ .

Next we form the QR factorization of the matrix  $(\bar{X}^T, \gamma I)^T$  by Householder elementary reflectors, so that

$$\begin{bmatrix} -\bar{X} \\ \gamma I \end{bmatrix} = \bar{Q} \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix}, \quad (12)$$

where  $\bar{Q} = Q + \delta Q$ ,  $\|\delta Q\| \approx \varepsilon_M$ ,  $Q^T Q = I$ . In other words, the computed matrix  $\bar{Q}$  is orthogonal to machine precision [20]

In the next section, we will show that in some pathological cases, the norm of the  $(2, 1)$  (block) entry of  $\bar{Q}^T A \bar{Q}$  may be larger than  $O(\epsilon_M \|A\|_M)$ , i.e., it may be backward unstable if we are forced to treat  $\bar{Q}^T A \bar{Q}$  as block upper triangular by setting the  $(2, 1)$  entry to zero. Therefore we propose to perform adjacent blocks swapping tentatively; if the norm of the  $(2, 1)$  (block) entry of  $\bar{Q}^T A \bar{Q}$  is less than or equal to  $O(\epsilon_M \|A\|_M)$ , we swap the blocks, otherwise we return without performing the swap. This gives an absolute guarantee of backward stability. We can fail to swap only if the eigenvalues  $A_{11}$  and  $A_{22}$  are so close that a small perturbation of the matrix could make them identical. If  $p = q = 1$ , then swapping will always succeed.

If the two blocks are exchanged, then an orthogonal similarity transformation is performed on the  $2 \times 2$  blocks (if any exist) to return them to standard form.

Finally, since the nonsymmetric eigenvalue problem is an ill-conditioned problem, a small perturbation to a  $2 \times 2$  block (complex conjugate eigenpair) could cause a large perturbation of its eigenvalues. In the extreme case, a  $2 \times 2$  block could split into two  $1 \times 1$  blocks if its complex conjugate eigenvalues become real. Carefully designed standardization steps will detect and report such phenomena. All above considerations are summed up in the following algorithm.

### Direct Swapping Algorithm SLAEXC

1. Copy  $A$  to  $T$ :  $T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \leftarrow A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$ ;
2. Use Gaussian elimination with complete pivoting to solve

$$T_{11}X - \gamma T_{12} = \gamma T_{22},$$

where  $\gamma$  is a scaling factor to prevent overflow. If there is a small diagonal element during Gaussian elimination, set it to roughly machine precision (times the norm of the matrix).

3. Compute the QR factorization  $G = (-X\gamma I)^T = QR$  by Householder transformations.
4. Perform swapping tentatively: if the norm of the  $(2, 1)$  (block) entry of  $TQQ^T$  is less than  $O(\epsilon_M \|T\|_M)$ , go to the next step, and otherwise exit;
5. If the swap is accepted, replace  $A$  by  $\bar{Q}A\bar{Q}$  and set the  $(2, 1)$  (block) entry of  $\bar{Q}A\bar{Q}$  to zero.
6. Standardize  $2 \times 2$  diagonal block(s) if any exist.

In our implementation of SLAEXC in LAPACK, we have chosen  $10\epsilon_M \|A\|_M$  as the stability criterion in step 4, where  $\|A\|_M = \max_{i,j} |a_{ij}|$ . Finally, we note that we also provide a subroutine STREXC in LAPACK which calls SLAEXC to reorder all the eigenvalues into a user selected order. In particular, the user may select any subset of the spectrum which will be reordered to appear at the top left of the matrix using the fewest possible calls to SLAEXC.

## 3 Error Analysis

In this section, we give an error analysis of the direct swapping algorithm SLAEXC described in the last section. We assume that  $p = q = 2$ , i.e., we only consider swapping two  $2 \times 2$  blocks, the hardest case of the problem. In addition, for the sake of exposition, we also assume that the computation of QR factorization and the similarity transformation  $\bar{Q}A\bar{Q}$  are exact, and the scaling

factor  $\gamma = 1$ . Including these rounding errors does not change the conclusion of the analysis, but makes the exposition appear more complicated.

Let  $\bar{X}$  be the computed solution of the Sylvester equation, where  $\bar{X} = X + E$ ,  $X$  is the exact solution, and  $E$  is an error matrix. By the argument of (12), and a result of Stewart on the perturbation of the QR factorization, we know that under mild conditions (such as  $\|E\|_2 \|Q\|_F < 1$ ), the QR factorization of  $(\bar{X}^T, I)^T$  can be written as

$$\begin{bmatrix} -X \\ I \end{bmatrix} + \begin{bmatrix} -E \\ 0 \end{bmatrix} = G + \begin{bmatrix} -E \\ 0 \end{bmatrix} = \hat{Q}\hat{R} = (Q+W) \begin{bmatrix} R+F \\ 0 \end{bmatrix}, \quad (13)$$

where  $W$  and  $F$  are the perturbations of the orthogonal matrix  $Q$  and the triangular matrix  $R$ , respectively, and  $\hat{Q} = Q+W$  is orthogonal.  $\|W\|_2$  and  $\|F\|_F$  are essentially bounded by the terms of order  $\|E\|_2 \|Q\|_F$ . From  $(Q+W)^T(Q+W) = I$ , up to the first order, we have  $\hat{Q}^T W = -W^T Q$ . When  $\hat{Q} = Q+W$  transforms  $A$  ignoring the second order perturbations we have

$$\begin{aligned} \hat{Q}^T A \hat{Q} &= (Q+W)^T A (Q+W) \\ &= Q^T A Q + W^T A Q + Q^T A W + W^T A W \\ &= \tilde{A} + W^T Q \cdot Q^T A Q + Q^T A Q \cdot Q^T W \\ &= \tilde{A} + \tilde{A} Q^T W - Q^T W \tilde{A}. \end{aligned}$$

Defining  $Z = Q^T W$  and partitioning it conformally with  $\tilde{A}$  in the form

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix},$$

we have

$$\hat{Q}^T A \hat{Q} = \begin{bmatrix} \tilde{A}_{22} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} + \begin{bmatrix} E_{22} & E_{12} \\ E_{21} & E_{11} \end{bmatrix}, \quad (14)$$

where

$$\begin{aligned} E_{11} &= \tilde{A}_{11} Z_{22} - Z_{22} \tilde{A}_{11} - Z_{21} \tilde{A}_{12}, \\ E_{22} &= \tilde{A}_{22} Z_{11} - Z_{11} \tilde{A}_{22} + \tilde{A}_{12} Z_{21}, \\ E_{21} &= \tilde{A}_{11} Z_{21} - Z_{21} \tilde{A}_{22}. \end{aligned}$$

$E_{11}$  and  $E_{22}$  perturb the eigenvalues directly and do not affect stability of interest because it measures the numerical stability of swapping in the error in the block. It is not of interest since it neither affects the numerical stability of the algorithm nor perturbs the eigenvalues. The task is to give bounds on the norms of  $E_{11}$ ,  $E_{22}$  and  $E_{21}$ . To do so, let us first express  $Z$  in terms of the blocks  $Q$  of  $Q$ ,  $E$ ,  $F$  and  $R$ . From (13), we have

$$(I + Q^T W) \begin{bmatrix} R+F \\ 0 \end{bmatrix} = Q^T \begin{bmatrix} -X \\ I \end{bmatrix} + Q^T \begin{bmatrix} -E \\ 0 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix} + \begin{bmatrix} -Q_{11}^T E \\ -Q_{12}^T E \end{bmatrix}.$$

Postmultiply by  $(R+F)^{-1}$  on both sides of the above equation, and noting that  $Z = Q^T W$ , we get

$$(I+Z) \begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} R - Q_{11}^T E \\ -Q_{12}^T E \end{bmatrix} (R+F)^{-1},$$

so that

$$\begin{aligned} Z_{11} &= -I + (R - Q_{11}^T E)(R + F)^{-1}, \\ Z_{21} &= -Q_{12}^T E(R + F)^{-1}, \end{aligned}$$

and up to the first order perturbations, we have

$$Z_{11} = -Q_{11}^T E R^{-1} - F R^{-1}, \quad (15)$$

$$Z_{21} = -Q_{12}^T E R^{-1}. \quad (16)$$

To express  $Z_{22}$ , again from (13),

$$(I + Q W^T) \begin{bmatrix} -X - E \\ I \end{bmatrix} = Q \begin{bmatrix} R \\ 0 \end{bmatrix} + Q \begin{bmatrix} F \\ 0 \end{bmatrix} = \begin{bmatrix} -X \\ I \end{bmatrix} + Q \begin{bmatrix} F \\ 0 \end{bmatrix}.$$

By canceling  $(-X, I)^T$  from both sides of the equation, and premultiplying by  $Q$ , we obtain

$$W^T \begin{bmatrix} -X - E \\ I \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix} + Q^T \begin{bmatrix} E \\ 0 \end{bmatrix}.$$

By inserting  $Q = I$  in the left side of the above equation and noting that  $QW = -Q^T W = -Z$ , we have

$$Z \begin{bmatrix} R \\ 0 \end{bmatrix} - Z \begin{bmatrix} Q_{11}^T E \\ Q_{12}^T E \end{bmatrix} = - \begin{bmatrix} F \\ 0 \end{bmatrix} - \begin{bmatrix} Q_{11}^T E \\ Q_{12}^T E \end{bmatrix}.$$

Thus the ‘‘bottom’’ equation is

$$Z_{21} R - Z_{21} Q_{11}^T E - Z_{22} Q_{12}^T E = -Q_{12}^T E$$

by (16) and assuming that error matrix  $E$  is nonsingular, we get

$$Z_{22} = -Z_{21} Q_{11}^T Q_{12}^{-T} = Q_{12}^T E R^{-1} Q_{11}^T Q_{12}^{-T}. \quad (17)$$

From expressions (15), (16) and (17) of  $Z_{11}$ ,  $Z_{21}$  and  $Z_{22}$ , the  $E_{11}$ ,  $E_{22}$  and  $E_{21}$  are recast as

$$\begin{aligned} E_{11} &= Q_{12}^T A_{11} Q_{12}^{-T} Q_{12}^T E R^{-1} Q_{11}^T Q_{12}^{-T} - Q_{12}^T E R^{-1} Q_{11}^T Q_{12}^{-T} Q_{12}^T A_{11} Q_{12}^{-T} \\ &\quad + Q_{12}^T E R^{-1} (-R A_{22} R^{-1} Q_{11}^T Q_{12}^{-T} + Q_{11}^T A_{11} Q_{12}^{-T}) \\ &= Q_{12}^T (A_{11} E - E A_{22}) R^{-1} Q_{11}^T Q_{12}^{-T} \\ &= -Q_{12}^T Y R^{-1} Q_{11}^T Q_{12}^{-T}, \end{aligned}$$

and

$$\begin{aligned} E_{22} &= -R A_{22} R^{-1} (Q_{11}^T E R^{-1} + F R^{-1}) + (Q_{11}^T E R^{-1} + F R^{-1}) R A_{22} R^{-1} \\ &\quad - (-R A_{22} R^{-1} Q_{11}^T Q_{12}^{-T} + Q_{11}^T A_{11} Q_{12}^{-T}) Q_{12}^T E R^{-1} \\ &= Q_{11}^T (-A_{11} E + E A_{22}) R^{-1} - \tilde{A}_{22} F R^{-1} + F R^{-1} \tilde{A}_{22} \\ &= Q_{11}^T Y R^{-1} - \tilde{A}_{22} F R^{-1} + F R^{-1} \tilde{A}_{22}, \end{aligned}$$

and

$$\begin{aligned} E_{21} &= -Q_{12}^T A_{11} Q_{12}^{-T} Q_{12}^T E R^{-1} + Q_{12}^T E R^{-1} R A_{22} R^{-1} \\ &= -Q_{12}^T (-A_{11} E + E A_{22}) R^{-1} \\ &= Q_{12}^T Y R^{-1}. \end{aligned}$$

We see that  $E_{11}$ ,  $E_{22}$  and  $E_{21}$  are essentially related to the residual vector  $Y$  of the Sylvester equation solver,  $R$  and the subblocks  $Q_{11}$  and  $Q_{12}$  of  $Q$ . Furthermore, rewriting (7) as

$$\begin{bmatrix} -X \\ I \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}$$

we see that

$$Q_{21} = R^{-1}$$

and

$$R^T R = I + X^{-T} X$$

Let  $\alpha(C)$  denote the set of singular values of matrix  $C$  and  $\chi(C)$  denote the set of eigenvalues of matrix  $C$  then

$$\sigma^2(R) = \chi(R^T R) = \chi(I + X^{-T} X) = 1 + \chi(X^{-T} X) = 1 + \sigma^{-2}(X).$$

Therefore

$$\|Q_{21}\|_2 = \|R^{-1}\|_2 = \frac{1}{\sigma_2(R)} = \frac{1}{(1 + \sigma_2^2(X))^{1/2}}, \quad (18)$$

where  $\sigma_1(X) \geq \sigma_2(X) \geq 0$ . Now to estimate the norm of the blocks  $Q_i$  of  $Q$ , we use the following CS decomposition of a partitioned orthogonal matrix, which was introduced by Stewart. [16] A proof of the existence of the decomposition can be found in [18]

**CS Decomposition:** Let the orthogonal matrix  $Q \in \mathbb{R}^{2k \times 2k}$  be partitioned in the form

$$Q = \begin{matrix} & k & k \\ k & \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} & \end{matrix}.$$

Then there are orthogonal matrices  $U = \text{diag}(U_1, U_2)$  and  $V = \text{diag}(V_1, V_2)$  with  $U_i, V_i \in \mathbb{R}^{k \times k}$  such that

$$U^T Q V = \begin{matrix} & k & k \\ k & \begin{pmatrix} C & S \\ -S & C \end{pmatrix} & \end{matrix},$$

where

$$C = \text{diag}(c_1, c_2, \dots, c_k) \geq 0, \quad S = \text{diag}(s_1, s_2, \dots, s_k) \geq 0, \quad C^2 + S^2 = I.$$

By the CS decomposition of  $Q$  and (18), we have

$$\|Q_{11}\|_2 = \frac{\sigma_1(X)}{(1 + \sigma_1^2(X))^{1/2}}$$

and

$$\|Q_{22}\|_2 = \|Q_{21}\|_2; \quad \|Q_{22}\|_2 = \|Q_{11}\|_2;$$

Thus, for  $E_{11}$ , we have

$$\|E_{11}\|_2 \leq \|Q_{12}^T\|_2 \|Y\|_F \|R^{-1}\|_2 \|Q_{11}^T\|_2 \|Q_{12}^T\|_2 = \frac{\sigma_1(X)}{1 + \sigma_1^2(X)} \|Y\|_F.$$

Similarly, for  $E_2$  from [17], we have  $\|FR^{-1}\|_F \leq 2\|\tilde{G}\|_2\|\tilde{H}\|_F$ , therefore

$$\|E_{22}\|_2 \leq \|\tilde{Q}_{11}^T\|_2 \|Y\|_F \|R^{-1}\|_2 + 2\|\tilde{A}_{22}\|_2 \|FR^{-1}\|_F \leq \frac{\sigma_1(\tilde{X})}{1+\sigma_2^2(\tilde{X})} \|Y\|_F + 4\|\tilde{A}_{22}\|_2 \|\tilde{G}\|_2 \|\tilde{H}\|_F.$$

Finally, for  $E_1$  we have

$$\|E_{21}\|_2 \leq \|\tilde{Q}_{12}^T\|_2 \|Y\|_F \|R^{-1}\|_2 = \frac{1}{1+\sigma_2^2(\tilde{X})} \|Y\|_F.$$

Hence we have the following theorem

**Theorem 2.** Let  $Y = A_{12} - A_{11}\bar{X} + \bar{X}A_{22}$ , where  $\bar{X} = X + E$  is the computed solution of the Sylvester equation (6), assume that the error matrix  $E$  is nonsingular, let the QR factorization of  $(-\bar{X}^T, I)^T$  satisfies

$$\begin{bmatrix} -\bar{X} \\ I \end{bmatrix} = \hat{Q} \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix},$$

then

$$\hat{Q}^T A \hat{Q} = \begin{bmatrix} \tilde{A}_{22} & \tilde{A}_{12} \\ 0 & \tilde{A}_{11} \end{bmatrix} + \begin{bmatrix} E_{22} & E_{12} \\ E_{21} & E_{11} \end{bmatrix},$$

where  $\tilde{A}_{ii}$  is similar to  $A_{ii}$ ,  $i=1,2$ , and up to the first order perturbation  $O(\|E\|)$

$$\|E_{11}\|_2 \leq \frac{\sigma_1(\tilde{X})}{1+\sigma_2^2(\tilde{X})} \|Y\|_F, \quad (19)$$

$$\|E_{22}\|_2 \leq \frac{\sigma_1(\tilde{X})}{1+\sigma_2^2(\tilde{X})} \|Y\|_F + 4\|\tilde{A}_{22}\|_2 \|\tilde{G}\|_2 \|\tilde{H}\|_F \quad (20)$$

$$\|E_{21}\|_2 \leq \frac{1}{1+\sigma_2^2(\tilde{X})} \|Y\|_F. \quad (21)$$

Three remarks are in order:

*Remark 1.* From the theorem we see that the departure  $\|E\|_F$  from upper block-triangular form (the measure of numerical instability) is bounded by  $\|Y\|_F / (\|Y\|_F + \sigma_2^2(\tilde{X}))$ . It is easy to see that

$$\|Y\|_F \leq \frac{\|A_2\|_F}{\text{sep}(A_{11}, A_{22})}, \quad (22)$$

where the equality is attained when  $e_{12}$  is a left singular vector of  $K$  corresponding to the smallest singular value  $\sigma(K) = \text{sep}(A_{11}, A_{22})$ . Combining (22), (11) and (21), we have

$$\|E_{21}\|_2 \leq \frac{\rho \varepsilon_M (\|A_{11}\|_F + \|A_{22}\|_F) \|A_2\|_F}{(1+\sigma_2^2(\tilde{X})) \text{sep}(A_{11}, A_{22})}.$$

Logically, the above bound indicates that the numerical instability will occur if we have small  $\text{sep}(A_{11}, A_{22})$ . But in practice, numerical experiments show that this upper bound is very pessimistic. Small  $\text{sep}(A_{11}, A_{22})$  does not imply instability. We will discuss this further in the following section.

*Remark 2.* Iterative refinement applied to the Sylvester equation will improve the accuracy of computed  $\bar{X}$ , (unless the Sylvester equation is too close to singular), but it need not improve  $\|Y\|_F$  at least when Gaussian elimination with complete pivoting is used to solve the Sylvester equation.

*Remark 3.* The factor  $\sigma_1(X)/(1+\sigma_2(X))$  that affects  $\|E_{11}\|_2$  and  $\|E_{22}\|_2$  is interesting, since it warns that large and ill-conditioned  $X$  may endanger accuracy, because of (11) and

$$\frac{\sigma_1(X)}{1+\sigma_2(X)} = \frac{\kappa(X)}{\sigma_2(X)+\sigma_2^{-1}(X)},$$

where  $\kappa(X) = \sigma_1(X)/\sigma_2(X)$ . How  $\kappa(X)$ ,  $\text{sep}(A_{11}, A_{22})$ , and the accuracy of the swapped eigenvalues are related in practice needs further investigation.

## 4 Software Development and Numerical Experiments

In this section, we first discuss the development of software for the swapping algorithm SLAEXC. Then we discuss numerical experiments to show the capability of our software to deal with ill-conditioned cases, compare with Stewart's swapping algorithm EXCHNG and finally demonstrate the sharpness of our perturbation bounds.

### 4.1 Software development

A set of FORTRAN subroutines has been developed to implement the direct swapping algorithm described in Section 3. It is part of LAPACK project. [As with other LAPACK routines, this algorithm was designed for accuracy, robustness and portability.

The main subroutine is called STREXC. STREXC moves a given  $1 \times 1$  or  $2 \times 2$  diagonal block of a real quasi-triangular matrix to a user specified position. On return, parameter INFO reports whether the given block has moved to the desired position, or whether there are blocks too close to swap, and what is the current position of the given block. The subroutine STREXC is supported by subroutine SLAEXC, which exchanges adjacent blocks. The subroutine SLAEXC is an implementation of the algorithm SLAEXC described in Section 3, where the subproblem of solving the Sylvester equation (8) by Gaussian elimination with complete pivoting is implemented in subroutine SLASY2, and the subproblem of standardizing a  $2 \times 2$  block is implemented in subroutine SLANV2.

In the interest of simplicity, we also used some other subroutines from LAPACK and the BLAS to perform some basic linear algebra operations, such as generating Householder transformations, computing the 2-norm of a vector and so on.

Finally, a test subroutine has been written to automatically test the subroutine SLAEXC. There are nested loops over different block sizes, different numerical scales, and different conditionings of the problem.

### 4.2 Numerical experiments

**Backward stability test:** To measure the backward stability of a swapping algorithm we need to test (I) how close the matrix  $\tilde{Q}$  is to an orthogonal matrix, and (II) how close  $\tilde{Q}\tilde{A}\tilde{Q}^T$  is to the original matrix  $A$  where  $\tilde{A}$  is the computed  $\tilde{A}$ . In other words, we need to test whether the two quantities

$$E_Q = \frac{\|I - \tilde{Q}^T \tilde{Q}\|_1}{\varepsilon_M}, \quad E_A = \frac{\|A - \tilde{Q} \tilde{A} \tilde{Q}^T\|_1}{\varepsilon_M \|A\|_1}$$

are around 1, where  $\varepsilon_M$  is machine precision. To check the changes among eigenvalues is not required to judge the correctness of an algorithm since we know that there must have at least an order of  $O(\varepsilon_M \|A\|_1)$  perturbation to the original matrix after swapping, and the nonsymmetric

eigenvalue problems potentially ill-conditioned. However, for reasonably conditioned matrices, the changes in the eigenvalues do measure the accuracy of a swapping algorithm. For this reason, in the following numerical examples, we also compare the eigenvalues before and after swapping, besides checking quantities  $\kappa$  and  $E_A$ .

All numerical experiments were carried out on a SUN Sparc station 1+. The arithmetic is IEEE standard single precision, with machine precision  $\epsilon \approx 1.192 \times 10^{-7}$ .

We have done extensive testing on matrices with various mixtures of the block sizes, scales and closeness among eigenvalues. More specifically, we show the algorithm SLAEXC on the following four types of matrices:

Test Matrix 1: good separation of  $A_{11}$  and  $A_{22}$ , the eigenvalues before swapping are

$$\begin{aligned}\lambda_1 &= 0.2000000E+01 \pm i0.2085666E+02, \\ \lambda_2 &= 0.1000000E+01 \pm i0.2017424E+02,\end{aligned}$$

Test Matrix 2: moderate separation of  $A_{11}$  and  $A_{22}$ , the eigenvalues before swapping are:

$$\begin{aligned}\lambda_1 &= 0.1000000E+01 \pm i0.1732051E+01, \\ \lambda_2 &= 0.1001000E+01 \pm i0.1732916E+01.\end{aligned}$$

Test Matrix 3: close eigenvalues, the corresponding the Sylvester equation is very ill-conditioned, the eigenvalues before swapping are

$$\begin{aligned}\lambda_1 &= 0.1000000E+01 \pm i0.1000000E+01, \\ \lambda_2 &= 0.1001000E+01 \pm i0.1000000E+01.\end{aligned}$$

Test Matrix 4: the extreme case, where the eigenvalues of  $A_{11}$  and  $A_{22}$  are the same, and theoretically, the Sylvester equation solution is infinite. This matrix is used to test the robustness of our software against overflow,

$$\begin{aligned}\lambda_1 &= 0.1000000E+01 \pm i0.1732051E+01, \\ \lambda_2 &= 0.1000000E+01 \pm i0.1732051E+01.\end{aligned}$$

Table 1 summarizes the results of algorithm SLAEXC, where  $\text{sep}(A_{11}, A_{22})$  is computed by MATLAB, and included here for the sake of theoretical analysis. From Table 1, we see that both the backward stability and accuracy of the algorithm SLAEXC are satisfactory.

**Comparison with Stewart's algorithm EXCHNG:** We have done numerical comparisons between the direct swapping algorithm SLAEXC and Stewart's swapping algorithm EXCHNG [15], which uses QR iteration. Both algorithms perform well in most cases, but in certain cases, the algorithm EXCHNG is inferior to algorithm SLAEXC. For example, let

$$A(\tau) = \begin{pmatrix} 7.001 & -87 & 39.4\tau & 22.2\tau \\ & 5 & 7.001 & -12.2\tau \\ & 0 & 0 & 7.01 & -11.7567 \\ & 0 & 0 & 37 & 7.01 \end{pmatrix},$$

where  $\tau$  is a parameter. The matrix  $A(\tau)$  has the same eigenvalues for all  $\tau$ :

$$\begin{aligned}\lambda_1 &= 0.7001000E+01 \pm i0.2085666E+02, \\ \lambda_2 &= 0.7010000E+01 \pm i0.2085660E+02,\end{aligned}$$

Table 1: numerical tests of algorithmSLAEXC

Test	matrix	$\text{sep}(A_{11}, A_2)$	$E_Q$	$E_A$	eigenvalues after swapping
1	$\begin{pmatrix} 2 & -87 & -20000 & 10000 \\ 5 & 2 & -20000 & -10000 \\ 0 & 0 & 1 & -11 \\ 0 & 0 & 37 & 1 \end{pmatrix}$	$3.37 \times 10^{-1}$	0.260	0.197	$0.1000001E+01 \pm i0.2017424E+02$ $0.2000000E+01 \pm i0.2085665E+02$
2	$\begin{pmatrix} 1 & -3 & 3576 & 4888 \\ 1 & 1 & -88 & -1440 \\ 0 & 0 & 1.001 & -3 \\ 0 & 0 & 1.001 & 1.001 \end{pmatrix}$	$8.442 \times 10^{-4}$	0.625	0.423	$0.1001000E+01 \pm i0.1732917E+01$ $0.1000000E+01 \pm i0.1732051E+01$
3	$\begin{pmatrix} 1 & -100 & 400 & -1000 \\ 0.01 & 1 & 1200 & -10 \\ 0 & 0 & 1.001 & -0.01 \\ 0 & 0 & 100 & 1.001 \end{pmatrix}$	$2.000 \times 10^{-7}$	0.417	0.001	$0.1000996E+01 \pm i0.1000360E+01$ $0.1000003E+01 \pm i0.9995396E+00$
4	$\begin{pmatrix} 1 & -3 & 3 & 2 \\ 1 & 1 & 9 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 1 & 1 \end{pmatrix}$	$\infty$	0.687	0.241	$0.9999987E+00 \pm i0.1732051E+01$ $0.1000002E+01 \pm i0.1732051E+01$

and  $\text{sep}(A_{11}, A_{22}) = 0.0024$ . When  $\tau = 1$ , the output matrix of the algorithmSLAEXC is

$$\tilde{A} = \begin{pmatrix} 0.70100012E+01 & -0.86993660E+02 & -0.39390938E+02 & -0.22241005E+02 \\ 0.50003409E+01 & 0.70100012E+01 & 0.12191071E+02 & -0.35999401E+02 \\ 0.00000000E+00 & 0.00000000E+00 & 0.70009995E+01 & -0.11755549E+02 \\ 0.00000000E+00 & 0.00000000E+00 & 0.37003792E+02 & 0.70009995E+01 \end{pmatrix}.$$

The eigenvalues after swapping are

$$\begin{aligned} \tilde{\lambda}_2 &= 0.7010001E+01 \pm i0.2085661E+02, \\ \tilde{\lambda}_1 &= 0.7000999E+01 \pm i0.2085665E+02, \end{aligned}$$

which are accurate to machine precision. However, the output of algorithmEXCHNG after 8 QR iterations<sup>1</sup> is

$$\tilde{A} = \begin{pmatrix} 0.28140299E+02 & -0.81122643E+02 & -0.39849255E+02 & -0.15834051E+02 \\ 0.10856283E+02 & -0.14087547E+02 & -0.23942078E+02 & 0.32877380E+02 \\ 0.00000000E+00 & 0.00000000E+00 & 0.19211971E+02 & 0.21227583E+02 \\ 0.00000000E+00 & 0.00000000E+00 & -0.27540298E+02 & -0.52427406E+01 \end{pmatrix},$$

which has eigenvalues

$$\begin{aligned} \tilde{\lambda}_2 &= 0.7026377E+01 \pm i0.2085408E+02 \\ \tilde{\lambda}_1 &= 0.6984615E+01 \pm i0.2085919E+02 \end{aligned}$$

They only have two decimal digits correct.

Table 2 shows the numerical results with different choices of parameter  $\tau$ , where when  $\tau = 10$ , it takes 17 QR iterations to converge. It clearly shows the superiority of algorithmSLAEXC. In particular, we note that algorithmEXCHNG is nonconvergent when  $\tau = 100$ . It means that the eigenvalues are not able to be exchanged by algorithmEXCHNG. But the algorithmSLAEXC has no difficulty. This convergence difficulty may reflect recent work of Battersby<sup>[13]</sup> who has discovered classes of nonsymmetric matrices where QR iteration fails to converge, or converges quite slowly.

<sup>1</sup>where the stopping criterion used in QR iteration is  $\text{eps} = 1.2 \times 10^{-16}$

Table 2: Comparison of algorithms SLAEX and EXNG

$\tau$	SLAEX	EXNG
1	$\tilde{\lambda}_2 = 0.700001E+01 \pm i0.2085661E+02$ $\tilde{\lambda}_1 = 0.700009E+01 \pm i0.208565E+02$	$\tilde{\lambda}_2 = 0.7026377E+01 \pm i0.2085408E+02$ $\tilde{\lambda}_1 = 0.6984615E+01 \pm i0.2085909E+02$
10	$\tilde{\lambda}_2 = 0.700000E+01 \pm i0.208560E+02$ $\tilde{\lambda}_1 = 0.700009E+01 \pm i0.208565E+02$	$\tilde{\lambda}_2 = 0.706305E+01 \pm i0.2086175E+02$ $\tilde{\lambda}_1 = 0.694970E+01 \pm i0.2085144E+02$
100	$\tilde{\lambda}_2 = 0.700009E+01 \pm i0.208560E+02$ $\tilde{\lambda}_1 = 0.700009E+01 \pm i0.208565E+02$	not convergent after 30 QR steps

**On the upper bound of  $\|E_{21}\|_2$ :** Finally in the interest of theoretical analysis, we discuss the sharpness of the bound on  $\|E_{21}\|_2$ , which controls the numerical stability of algorithm SLAEX. In most of the test examples, we see that the bound (21) of  $\|E_{21}\|_2$  is very pessimistic. However, we do find some examples indicating that the bound in (21) can roughly be attained. Let us consider the following example:<sup>2</sup>

$$A = \begin{pmatrix} 2 & 2 \\ A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} = \begin{pmatrix} 1.0000E+00 & -1.0000E+02 & 1.9900E+04 & 1.0201E+02 \\ 1.0000E-02 & 1.0000E+00 & 1.0000E+02 & -1.9800E+00 \\ & 0 & 0 & 1.0100E+00 \\ & 0 & 0 & 1.0000E+02 \end{pmatrix},$$

where  $\text{sep}(A_{11}, A_{22}) = 2 \times 10^{-6}$ . The  $A_{12}$  block of  $A$  is designed so that

$$X = \begin{pmatrix} 1.0000E+00 & -2.0000E+02 \\ 1.0000E+00 & -1.0000E+00 \end{pmatrix}$$

is the solution of the Sylvester equation. Note that  $\sigma_1(X) = 20.01$ ,  $\sigma_2(X) = 0.9998$ . We used MATLAB to compute the different quantities in the bound (where machine precision is doubled  $\varepsilon_M = 2.204 \times 10^{-16}$ ). First the norm of the residual matrix  $Y$  for computed solution  $\bar{X}$  of the Sylvester equation is

$$\|Y\|_F = \|A_{12} - A_{11}\bar{X} + \bar{X}A_{22}\|_F = 4.0272 \times 10^{-12},$$

which almost reaches the estimated bound (11) of  $Y$ :

$$\varepsilon_M (\|A_{11}\|_F + \|A_{22}\|_F) \|X\|_F = 8.8830 \times 10^{-12}.$$

Furthermore, the desired norm of (2,1) block  $\tilde{A}_{21}$  after swapping is

$$\|\tilde{A}_{21}\|_2 = 1.293 \times 10^{-12},$$

which also roughly attains the bound (21) for  $\|E_{21}\|_2$ :

$$\|E_{21}\|_2 \leq \frac{1}{1 + \sigma_2(X)} \|Y\|_F = 2.0237 \times 10^{-12},$$

Note that for this example, the algorithm is still backward stable, since

$$\|\tilde{A}_{21}\|_2 = 1.293 \times 10^{-12} \leq \varepsilon_M \|A\|_F = 4.4189 \times 10^{-12}.$$

<sup>2</sup>For brevity, only five digits are displayed for all the data in this section though we did run in double precision.

After setting  $\check{A}_{21} = 0$ , then the measures of backward stability are  $E_Q = 2.3$  and  $E_A = 1.8$ . From Remark 1 after Theorem 2, we might worry that a huge  $\|X\|_F$  or tiny  $\text{sep}(A_{11}, A_{22})$  could cause numerical instability. However the following example illustrates how in practice a small separation of  $A_{11}$  and  $A_{22}$  does not necessarily lead to instability. Let

$$A_{11} = \begin{bmatrix} 1 & -10^{-6} \\ 1 & 1 \end{bmatrix}, \quad A_{22} = A_{11} + \sqrt{\varepsilon_M} I,$$

then the separation of  $A_{11}$  and  $A_{22}$  is tiny; that is  $\text{sep}(A_{11}, A_{22}) = 2.982 \times 10^{-14}$ . Let  $A_{12}$  be chosen such that  $\text{col}(A_{12})$  is the left singular vector of  $K$  corresponding to the smallest singular value  $\sigma_{\min}(K)$ , so that the norm of the solution  $X$  of the Sylvester equation  $A_{11}X - XA_{22} = A_{12}$  reaches its upper bound (2), that is

$$\|X\|_F = \frac{\|A_{12}\|_F}{\text{sep}(A_{11}, A_{22})} = 3.354 \times 10^{13}$$

and  $\kappa(X) = 10^6$ . Here the estimated bound of the norm of residual  $Y$  is

$$\varepsilon_M(\|A_{11}\|_F + \|A_{22}\|_F)\|X\|_F = 2.580 \times 10^{-2}.$$

However in practice, the observed residual  $\text{norm}\|Y\|_F = 3.723 \times 10^{-9}$ . After swapping it turns out that

$$\|\check{A}_{21}\|_F = 7.385 \times 10^{-24} \ll \varepsilon_M\|A\|_F = 5.874 \times 10^{-16}.$$

So the swapping is perfectly stable!

## 5 Conclusions

In this paper, we have developed a direct swapping algorithm which reorders the eigenvalues on the diagonal of a matrix in real Schur form by performing an orthogonal similarity transformation. A complete set of FORTRAN subroutines has been developed and included in the LAPACK library [1]. The algorithm is guaranteed to be numerically stable because we explicitly test for instability and do not reorder the eigenvalues if this would be unstable; this can only happen if the eigenvalues are so close as to be numerically indistinguishable. Unfortunately there is no proof of the backward stability of the algorithm without this explicit test, even though we have not seen an example where instability would occur. The detailed error analysis and numerical examples show how well it deals with ill-conditioned cases, whereas the alternative stable algorithm EXCHG may occasionally fail to converge.

**Acknowledgement:** The authors would like to thank K. C. Ng and B. Parlett for sharing their programming or initial work on the subject. The valuable comments of G. W. Stewart and B. Parlett during the development of software are gratefully acknowledged. The authors are also indebted to A. Edelman and N. Higham for their valuable comments on the subject.

## References

- [1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. Meney, S. Ostrouchov and D. Sorensen, *LAPACK Users' Guide, Release 1.0*, SIAM 1992.

- [2] Z. Bai, J. Demmel and A. McKenney, *On Computing Condition Numbers for the Nonsymmetric Eigenproblem*, to appear in *ACM TOMS*.
- [3] S. Bitterson, *Convergence of the QR algorithm on  $3 \times 3$  normal matrices*, *Num. Math.* 58:341-352 (1990).
- [4] C. Broyd and G. W. Stewart, *An algorithm for computing reducing subspaces by block diagonalization*, *SIAM. Num. Anal.*, 16:359-367 (1979).
- [5] R. S. Bartels and G. W. Stewart, *Solution of the matrix equation  $AX + XB = C$* , *Comm. ACM* 15:820-826 (1972).
- [6] Z. Ge and F. Zhang, *Direct methods for ordering eigenvalues of a real matrix*, *Chinese University J. of Comp. Math.* 1:27-36 (1981), in Chinese.
- [7] J. Dongarra, S. Hamming and J. Wilkinson, *Numerical considerations in computing invariant subspaces*, *SIAM. Math. Anal. Appl.* 13:145-161 (1982).
- [8] F. R. Gantmacher, *Theory of Matrices*, Vol. I, Chelsea, New York, 1959.
- [9] G. Golub, S. Nash and C. Van Loan, *A Hessenberg-Schur Method for the Problem  $AX + XB = C$* , *IEEE Trans. Automat. Contr.* AC-24: 909-913 (1979).
- [10] V. Mann, *A Symplectic orthogonal method for single input or single output discrete time optimal control problems*, in *Linear Algebra in Signals Systems and Control*, B. N. Datta et al. eds. SIAM Philadelphia, Penn pp 128-140 (1988).
- [11] K. C. Ng and B. N. Parlett, *Development of an accurate algorithm for  $EXP(Bt)$ , Part I, Programs to swap diagonal block, Part II*, *CAM-94*, University of California, Berkeley, 1988.
- [12] A. Rife, *An algorithm for numerical determination of the structure of a general matrix*, *BIT* 10:196-216 (1970).
- [13] B. T. Smith et al., *Matrix Eigensystem Routines - EISPACK guide*, second edition. Lecture Notes in Computer Science 19, Springer-Verlag, 1976.
- [14] G. W. Stewart, *Simultaneous Iteration for Computing Invariant Subspaces of Non-Hermitian Matrices*, *Numer. Math.* 25:12-36 (1976).
- [15] G. W. Stewart, *Algorithm 506 HQR3 and EXCHANG: Fortran subroutine for calculating and ordering the eigenvalues of a real upper Hessenberg matrix [F2]*, *ACM SIG2* 2:275-280 (1976).
- [16] G. W. Stewart, *On the perturbation of pseudo-Inverse, projections, and linear least squares problems*, *SIAM Rev.*, 19:634-662 (1977).
- [17] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, *SIAM. Num. Anal.* 14:309-318 (1977).
- [18] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Inc. New York, 1990.
- [19] P. Van Dooren, *Algorithm 590, DSUBSP and EXCHQZ: Fortran subroutines for computing deflating subspaces with specified spectrum*, *ACM SIG8* 8:376-382 (1982).
- [20] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford Univ. Press, 1965.