SIAM J. SCI. STAT. COMPUT.
Vol. 2, No. 2, June 1981

0196-5204/81/0202-0001 $01.00/0

A GENERALIZED EIGENVALUE APPROACH FOR SOLVING RICCATI EQUATIONS*

P. VAN DOOREN†

Abstract. A numerically stable algorithm is derived to compute orthonormal bases for any deflating subspace of a regular pencil $\lambda B - A$. The method is based on an update of the QZ-algorithm, in order to obtain any desired ordering of eigenvalues in the quasitriangular forms constructed by this algorithm. As applications we discuss a new approach to solve Riccati equations arising in linear system theory. The computation of deflating subspaces with specified spectrum is shown to be of crucial importance here.

Key words. generalized eigenvalue problem, Riccati equation, optimal control, spectral factorization

1. Introduction. The computation of deflating subspaces with a specified spectrum has not received a great deal of attention until it was recently applied to the solution of the optimal control problem of a linear discrete time system [5], [15]. Before the development of reliable algorithms for the generalized eigenvalue problem [13], [16], these problems were often reduced to an equivalent standard eigenvalue problem and gave rise to the computation of invariant subspaces with a specified spectrum [8], [14], [17], [21]. The matrix involved in this standard eigenvalue problem does not consist of given data but has to be computed, which unfortunately requires inverses of possibly ill-conditioned matrices. In [5], [12], [15] the use of a generalized eigenvalue problem is recommended as a safer alternative, and attention is drawn to the absence of appropriate software for computing deflating subspaces of a regular pencil. In this paper we try to fill this gap, and we also exploit this new tool in a class of related problems arising in linear system theory. We thereby develop a new approach to tackle these problems in a numerically sound way.

In the rest of this section we briefly review some notions that we will need in later sections. The material covered here can be found, e.g., in [13], [18], [19], [20].

Notation will be as follows. We use uppercase for matrices and lowercase for vectors and scalars. $\mathbb{R}$ and $\mathbb{C}$ are the fields of real and complex numbers, respectively. We use A^* (resp. x^*) for the conjugate transpose of a complex matrix A (resp. vector x) and A' (resp. x') for the transpose of a real matrix A (resp. vector x). $\|\cdot\|_2$ denotes the spectral norm of a matrix and the Euclidean norm of a vector. A complex (real) square matrix A is called *unitary* (*orthogonal*) when $A^*A = I$ ($A'A = I$). When no explicit distinction is made between the complex and real case, we use the term unitary and the notation A^* for the real case as well.

Recently, more attention has been paid to the *generalized eigenvalue problem* (GEP):

$$Ax = \lambda Bx, \tag{1}$$

where B is not necessarily invertible but where the *pencil* $\lambda B - A$ is *regular*, i.e.,

$$\det(\lambda B - A) \not\equiv 0. \tag{2}$$

* Received by the editors July 15, 1980. This research was supported by the National Science Foundation under grant ENG78-10003 and by the U.S. Air Force under grant AFOSR-79-0094.

† Departments of Electrical Engineering and Computer Science, Stanford University, Stanford, California 94305. Present address, Philips Research Lab., Av. Van Becelaere 2, Box 8, B-1170 Brussels, Belgium.

When the coefficients of the matrices A and B belong to $\mathbb{C}$, there exist unitary transformations Q and Z reducing the $n \times n$ pencil $\lambda B - A$ to the upper triangular form

$$(3) \qquad Q^*(\lambda B - A)Z = \lambda \hat{B} - \hat{A} = \lambda \begin{bmatrix} \hat{b}_{11} & \cdots & * \\ & \ddots & \vdots \\ 0 & & \hat{b}_{nn} \end{bmatrix} - \begin{bmatrix} \hat{a}_{11} & \cdots & * \\ & \ddots & \vdots \\ 0 & & \hat{a}_{nn} \end{bmatrix}.$$

The ratios $\lambda_i = \hat{a}_{ii}/\hat{b}_{ii}$ are called the *generalized eigenvalues* of the pencil $\lambda B - A$. The set $\{\lambda_1, \cdots, \lambda_n\}$ is called the *spectrum* of $\lambda B - A$ and is denoted by $\Lambda(B, A)$; it may contain repeated elements. Notice that λ_i may be infinite (when $\hat{b}_{ii} = 0$) but it is never undetermined (i.e., $\lambda_i = 0/0$), since $\hat{a}_{ii} = \hat{b}_{ii} = 0$ implies $\det(\lambda \hat{B} - \hat{A}) \equiv 0$ and hence $\det(\lambda B - A) \equiv 0$. As a consequence the matrix $\hat{a}_{ii}B - \hat{b}_{ii}A$ is singular. The vectors x_i satisfying

$$(4) \qquad (\hat{a}_{ii}B - \hat{b}_{ii}A)x_i = 0$$

are called *generalized eigenvectors* of $\lambda B - A$ corresponding to λ_i. If the eigenvalue $\lambda_i = \hat{a}_{ii}/\hat{b}_{ii}$ has a larger multiplicity than the number of independent solutions x_i of (4), then one can define *generalized principal vectors* $\tilde{x}_i$ of $\lambda B - A$ corresponding to λ_i. Since we do not need this concept in the sequel, we do not go into further details about it.

In the real case the decomposition (3) also exists but involves complex matrices Q, Z, $\hat{A}$ and $\hat{B}$ when $\Lambda(B, A)$ contains complex elements. Under orthogonal transformations Q and Z, $\lambda B - A$ can be transformed to the quasi upper triangular form

$$(5) \qquad Q'(\lambda B - A)Z = \lambda \hat{B} - \hat{A} = \lambda \begin{bmatrix} \hat{B}_{11} & \cdots & * \\ & \ddots & \vdots \\ 0 & & \hat{B}_{kk} \end{bmatrix} - \begin{bmatrix} \hat{A}_{11} & \cdots & * \\ & \ddots & \vdots \\ 0 & & \hat{A}_{kk} \end{bmatrix},$$

where the diagonal pencils $\lambda \hat{B}_{ii} - \hat{A}_{ii}$ have sizes $d_i = 1$ or 2 and the $\hat{B}_{ii}$ are upper triangular. If $d_i = 1$ then $\Lambda(\hat{B}_{ii}, \hat{A}_{ii})$ is real (possibly infinite). If $d_i = 2$ then $\Lambda(\hat{B}_{ii}, \hat{A}_{ii})$ contains two (finite) complex conjugate numbers. The spectrum of $\lambda B - A$ is the union of the sets $\Lambda(\hat{B}_{ii}, \hat{A}_{ii})$, as can be seen from an additional (unitary) reduction of (5) to (3). An algorithm has been derived recently to obtain decompositions of the type (3) and (5) in a numerically stable way [13]. When $B = I$, (1) boils down to the *standard eigenvalue problem* (SEP):

$$(6) \qquad Ax = \lambda x.$$

It is readily verified that the decompositions (5) and (3) then reduce to the classical Schur decompositions of the real or complex matrix A, respectively. We therefore call (3) and (5) *generalized Schur decompositions* of the regular pencil $\lambda B - A$. In the sequel we drop the term "generalized" when no confusion is possible from the context. The notion of eigenvector in the GEP can be extended to the notion of *deflating subspace* $\mathcal{X}$ of a regular pencil $\lambda B - A$, satisfying

$$(7) \qquad \dim(B\mathcal{X} + A\mathcal{X}) = \dim \mathcal{X},$$

where $\dim \mathcal{S}$ denotes the dimension of a subspace $\mathcal{S}$. Let $\mathcal{X}$ have dimension l, and suppose that the l first columns of the unitary matrices Q and Z, partitioned as

$$(8) \qquad Z = [\underbrace{Z_1}_{l} | \underbrace{Z_2}_{n-l}], \qquad Q = [\underbrace{Q_1}_{l} | \underbrace{Q_2}_{n-l}],$$

span the spaces $\mathcal{X}$ and $A\mathcal{X} + B\mathcal{X}$, respectively. Then it follows from (7) that $Q_2^* A Z_1 =$

$Q_2^* B Z_1 = 0$, or

$$Q^*(\lambda B - A)Z = \lambda \left[\begin{array}{c|c} \hat{B}_{11} & \hat{B}_{12} \\ \hline 0 & \hat{B}_{22} \end{array}\right] - \left[\begin{array}{c|c} \hat{A}_{11} & \hat{A}_{12} \\ \hline 0 & \hat{A}_{22} \end{array}\right] \begin{matrix} \} l \\ \} n-l \end{matrix}. \tag{9}$$

Conversely, if (8), (9) hold then the columns of Z_1 span a deflating subspace $\mathscr{X}$ according to (7). For $l = 1$, $\mathscr{X}$ is an eigenvector of $\lambda B - A$ corresponding to the eigenvalue $\Lambda(\hat{B}_{11}, \hat{A}_{11})$. For any l, $\Lambda(\hat{B}_{11}, \hat{A}_{11})$ is a subset of $\Lambda(B, A)$ and is denoted as $\Lambda(B, A)|_{\mathscr{X}}$ (the spectrum of $\lambda B - A$ restricted to $\mathscr{X}$). The deflating subspace $\mathscr{X}$ is uniquely determined by $\Lambda(B, A)|_{\mathscr{X}}$ when this subset is disjoint from the rest of $\Lambda(B, A)$ ($\mathscr{X}$ is then spanned by the eigenvectors and principal vectors corresponding to the spectrum $\Lambda(B, A)|_{\mathscr{X}}$). All this also holds for the real case. For the case $B = I$, the definition (7) of a deflating subspace reduces to the definition of an *invariant subspace* $\mathscr{X}$ of A, since $\dim(\mathscr{X} + A\mathscr{X}) = \dim \mathscr{X}$ is equivalent to $A\mathscr{X} \subset \mathscr{X}$. Notice also that in the SEP Q is equal to Z in (8), (9).

It follows now immediately from (9) that the Z-matrix in the Schur decomposition (3) yields orthonormal bases for deflating subspaces of dimension 1 to $n-1$, since the right-hand side of (3) has a block partitioning of the type (9) for $l = 1, \cdots, n-1$. This also holds for the "real" Schur decomposition (5) for those l that are conformable with the block partitioning in (5), namely

$$l = \sum_{j=1}^{i} d_j \quad \text{for } i = 1, \cdots, k-1. \tag{10}$$

In this paper we consider the computation of a deflating subspace $\mathscr{X}$ with prescribed spectrum $\Lambda(B, A)|_{\mathscr{X}} = \{\mu_1, \cdots, \mu_l\}$. From the above it follows that the l first columns of Z in (3) form an orthonormal basis for such a space $\mathscr{X}$ if and only if the sets $\{\lambda_i = \hat{a}_{ii}/\hat{b}_{ii} | i = 1, \cdots, l\}$ and $\{\mu_i | i = 1, \cdots, l\}$ are equal except for the ordering of their elements. In the real case, this also holds for the matrix Z in (5) when l satisfies (10). The complex elements in $\{\mu_i | i = 1, \cdots, l\}$ must therefore appear in conjugate pairs.

The problem thus reduces to obtaining decompositions of the type (3) and (5) but with prescribed ordering of the eigenvalues occurring on the diagonal. In the next section we show how to solve this problem by deriving a method to interchange the order of the eigenvalues in the decompositions (3) and (5), which were previously obtained by the QZ-algorithm. The method is proved to be numerically stable. In § 3 we apply this new tool to derive new methods for solving Riccati equations arising in linear system theory. In these methods, deflating subspaces with specified spectrum (namely, all the eigenvalues inside the unit circle or all the eigenvalues in the left half-plane) have to be computed. In § 4 we give some numerical examples.

2. Reordering. It is clear that the 1×1 and 2×2 diagonal blocks in the decompositions (3) and (5) can be reordered in an arbitrary way by using a method to interchange two consecutive blocks only. This idea was used, e.g., in the SEP to obtain standard Schur forms with an arbitrary ordering of the eigenvalues [8], [17], [21]. The method described hereafter can be viewed as a stable generalization of it to the GEP. (An unstable generalization was attempted in [15].) We thus want to find unitary transformations Q and Z such that

$$Q^* A Z = Q^* \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right] Z = \hat{A} = \left[\begin{array}{c|c} \hat{A}_{11} & \hat{A}_{12} \\ \hline 0 & \hat{A}_{22} \end{array}\right], \tag{11a}$$

$$Q^* B Z = Q^* \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right] Z = \hat{B} = \left[\begin{array}{c|c} \hat{B}_{11} & \hat{B}_{12} \\ \hline 0 & \hat{B}_{22} \end{array}\right], \tag{11b}$$

where $\Lambda(B_{11}, A_{11}) = \Lambda(\hat{B}_{22}, \hat{A}_{22})$ and $\Lambda(B_{22}, A_{22}) = \Lambda(\hat{B}_{11}, \hat{A}_{11})$, and where the dimensions d_1 and d_2 are either 1 or 2.

Moreover, we want the transformations Q and Z to be numerically stable. In order to prove this, we use a standard error analysis [23] of (possibly complex) transformations of the type

$$G^* y = G^* \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ 0 \end{bmatrix}, \tag{12a}$$

where G is the (possibly complex) Givens transformation

$$G = \begin{bmatrix} c & -\bar{s} \\ s & \bar{c} \end{bmatrix}, \qquad c\bar{c} + s\bar{s} = 1, \tag{12b}$$

constructed to annihilate y_2. Let $\tilde{c}, \tilde{s}$ (defining $\tilde{G}$) and $\tilde{y}_1$ be the computed versions of c, s and $\hat{y}_1$, respectively, and let ε be the machine precision of the computer; then a backward error analysis yields (for a standard construction of such transformations)

$$\tilde{G}^*(y + e_y) = \begin{bmatrix} \tilde{y}_1 \\ 0 \end{bmatrix}, \qquad \|e_y\|_2 \leqq 6 \cdot \varepsilon \|y\|_2. \tag{13}$$

Here we assume that the 0 element is not computed but put equal to zero. When performing the transformation $G^* z = \hat{z}$ for an arbitrary vector z, we have, similarly,

$$\tilde{G}^*(z + e_z) = \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix}, \qquad \|e_z\|_2 \leqq 6 \cdot \varepsilon \|z\|_2. \tag{14}$$

In the sequel $\mathcal{G}_{ij}$ denotes the class of matrices representing Givens transformations between columns or rows i and j. We prove that, by using transformations in this class for the reduction (11), the backward error can be bounded with respect to

$$\Delta = \max \{\|A\|_2, \|B\|_2\}. \tag{15}$$

Case I. $d_1 = d_2 = 1$. This may occur in both decompositions (3) and (5). We thus assume that the matrices can be complex. We have the following configuration:

$$Q^* A Z = Q^* \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} Z = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} \\ 0 & \hat{a}_{22} \end{bmatrix} = \hat{A}, \tag{16a}$$

$$Q^* B Z = Q^* \begin{bmatrix} b_{11} & b_{12} \\ 0 & b_{22} \end{bmatrix} Z = \begin{bmatrix} \hat{b}_{11} & \hat{b}_{12} \\ 0 & \hat{b}_{22} \end{bmatrix} = \hat{B}. \tag{16b}$$

We can assume without loss of generality that $|b_{22}| \geqq |a_{22}|$ (if this is not the case the role of A and B should be interchanged). A construction of Q and Z such that the order of the eigenvalues is interchanged follows then immediately from (8), (9). Indeed, we have $\Lambda(b_{22}, a_{22}) = \Lambda(\hat{b}_{11}, \hat{a}_{11})$ if the first column z_1 of Z is an eigenvector of $\lambda B - A$ corresponding to $\Lambda(b_{22}, a_{22})$ or

$$(a_{22}B - b_{22}A)Z = \left[\begin{array}{c|c} 0 & \\ 0 & * \end{array}\right]. \tag{17}$$

Notice that the last row of $H = (a_{22}B - b_{22}A)$ is zero:

$$H = \begin{bmatrix} x_1 & x \\ 0 & 0 \end{bmatrix}. \tag{18}$$

In order to obtain (17), we thus can choose a $Z \in \mathcal{G}_{12}$ annihilating x_1 in (18). It follows from (17) that Bz_1 and Az_1 are parallel, and (16) is then obtained by choosing a $Q \in \mathcal{G}_{12}$ annihilating x_2 in Bz_1:

$$Q^*BZ = Q^*\begin{bmatrix} x & x \\ x_2 & x \end{bmatrix} = \begin{bmatrix} x & x \\ 0 & x \end{bmatrix}. \tag{19}$$

The assumption $|b_{22}|/|a_{22}| = |\hat{b}_{11}|/|\hat{a}_{11}| \geqq 1$ implies that $\hat{b}_{11} \neq 0$, and Q^*Az_1 can then only be parallel to Q^*Bz_1 if Q^*AZ is indeed upper triangular.

We now prove the numerical stability of the method. As in (13), (14), computed elements are denoted by an upper tilde $\tilde{(\cdot)}$. Using the analysis (13) (14) above, it is easy to prove that all the ε_i, $i = 1, \cdots, 9$ below are of the order of the machine accuracy ε of the computer.

An error analysis of (17), (18) yields

$$(a_{22}B - b_{22}A + F)\tilde{Z} = \begin{bmatrix} 0 & x \\ 0 & 0 \end{bmatrix} \quad \text{for } \|F\|_2 = \varepsilon_1 \|a_{22}B - b_{22}A\|_2, \tag{20}$$

and of (19) yields

$$\tilde{Q}^*(B + E_b)\tilde{Z} = \begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} \\ 0 & \tilde{b}_{22} \end{bmatrix} \quad \text{for } \|E_b\|_2 = \varepsilon_2 \Delta. \tag{21}$$

We prove that there also exists a backward error E_a such that

$$\tilde{Q}^*(A + E_a)\tilde{Z} = \begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} \\ 0 & \tilde{a}_{22} \end{bmatrix} \quad \text{for } \|E_a\|_2 = \varepsilon_3 \Delta. \tag{22}$$

An error analysis of Q^*AZ using (14) yields

$$\tilde{Q}^*(A + E_c)\tilde{Z} = \begin{bmatrix} \tilde{a}_{11} & \tilde{a}_{12} \\ \tilde{a}_{21} & \tilde{a}_{22} \end{bmatrix} \quad \text{for } \|E_c\|_2 = \varepsilon_4 \Delta. \tag{23}$$

We only have to prove that $\tilde{a}_{21} \simeq \varepsilon \Delta$ in order to obtain (22) by putting $\tilde{a}_{21}$ equal to zero.

Let us therefore denote the (2, 1) elements of $\tilde{Q}^*(a_{22}B - b_{22}A)\tilde{Z}$, $\tilde{Q}^*B\tilde{Z}$ and $\tilde{Q}^*A\tilde{Z}$ by η_1, η_2 and η_3, respectively. They clearly satisfy the relation

$$a_{22} \cdot \eta_2 - b_{22} \cdot \eta_3 = \eta_1. \tag{24}$$

From (20), (21) and (23) it follows that

$$|\eta_1| \leqq \varepsilon_5\{|a_{22}|\|B\|_2 + |b_{22}|\|A\|_2\}, \tag{25a}$$

$$|\eta_2| \leqq \varepsilon_6 \Delta, \tag{25b}$$

$$|\tilde{a}_{21}| \leqq |\eta_3| + \varepsilon_7 \Delta. \tag{25c}$$

Using (25) and the assumption $|b_{22}| \geqq |a_{22}|$ in (24) we obtain

$$\begin{aligned} |\eta_3| &\leqq |\eta_2||a_{22}|/|b_{22}| + |\eta_1|/|b_{22}| \\ &\leqq \varepsilon_6 \Delta + \varepsilon_5\{\Delta + \Delta\} = \varepsilon_8 \Delta, \end{aligned} \tag{26a}$$

$$|\tilde{a}_{21}| \leqq (\varepsilon_8 + \varepsilon_7)\Delta = \varepsilon_9 \Delta. \tag{26b}$$

This shows the importance of the assumption $|b_{22}| \geqq |a_{22}|$ in order to guarantee the stability of the algorithm. In case $|b_{22}| < |a_{22}|$, Q is constructed to reduce A to triangular form instead of B, and a similar analysis is then possible.

Case II. $d_1 = 2$, $d_2 = 1$. We now have the following configuration (all matrices are real):

$$Q'AZ = Q'\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} Z = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \hat{a}_{13} \\ 0 & \hat{a}_{22} & \hat{a}_{23} \\ 0 & \hat{a}_{32} & \hat{a}_{33} \end{bmatrix} = \hat{A}, \tag{27a}$$

$$Q'BZ = Q'\begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{bmatrix} Z = \begin{bmatrix} \hat{b}_{11} & \hat{b}_{12} & \hat{b}_{13} \\ 0 & \hat{b}_{22} & \hat{b}_{23} \\ 0 & 0 & \hat{b}_{33} \end{bmatrix} = \hat{B}. \tag{27b}$$

We assume that $|b_{33}| \geq |a_{33}|$. If this is not the case, we can always interchange the role of A and B by transforming the first two columns of A and the last two columns of $\hat{A}$ in order to annihilate a_{21} and $\hat{a}_{32}$ and to create b_{21} and $\hat{b}_{32}$.

It follows again from (8), (9) that $\Lambda(b_{33}, a_{33}) = \Lambda(\hat{b}_{11}, \hat{a}_{11})$ if the first column z_1 of Z is an eigenvector of $\lambda B - A$ corresponding to $\Lambda(b_{33}, a_{33})$. Therefore we have (with R any invertible row transformation):

$$R'(a_{33}B - b_{33}A)Z = \left[\begin{array}{c|cc} 0 & & \\ 0 & * & \\ 0 & & \end{array}\right]. \tag{28}$$

Notice that the last row of $H = (a_{33}B - b_{33}A)$ is zero and that we can choose $R \in \mathcal{G}_{12}$ to annihilate the (2, 1) element of H. We then have

$$R'H = \begin{bmatrix} x_2 & x & x \\ 0 & x_1 & x \\ 0 & 0 & 0 \end{bmatrix}. \tag{29}$$

In order to obtain (28) we thus can choose $Z = Z_1 \cdot Z_2$, with $Z_1 \in \mathcal{G}_{23}$ and $Z_2 \in \mathcal{G}_{12}$ annihilating x_1 and x_2, respectively. Q is then constructed to have $\hat{B} = Q'BZ$ in upper triangular form. We therefore take $Q = Q_1 \cdot Q_2$, where $Q_1 \in \mathcal{G}_{23}$ is chosen to annihilate the (3, 2) element created by Z_1 (i.e., $Q_1'BZ_1$ is upper triangular) and where $Q_2 \in \mathcal{G}_{12}$ is chosen to annihilate the (2, 1) element created by Z_2 (i.e., $Q_2'Q_1'BZ_1Z_2$ is upper triangular). $\hat{B}$ now satisfies (27b). Since $|b_{33}|/|a_{33}| = |\hat{b}_{11}|/|\hat{a}_{11}| \geq 1$, we have $\hat{b}_{11} \neq 0$ and because of (28), $Q'Az_1$ and $Q'Bz_1$ are parallel. This ensures that $\hat{A} = Q'AZ$ also satisfies (27a).

We now prove the numerical stability of the method. Using (13), (14) it can be checked that all the ε_i, $i = 1, \cdots, 4$ below are of the order of the machine accuracy ε. An error analysis of (28), (29) yields

$$\tilde{R}'(a_{33}B - b_{33}A + F)\tilde{Z}_1\tilde{Z}_2 = \begin{bmatrix} 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \end{bmatrix} \quad \text{for } \|F\|_2 = \varepsilon_1 \|a_{33}B - b_{33}A\|_2, \tag{30}$$

and of the constructed product $Q_2'Q_1'BZ_1Z_2$ yields

$$\tilde{Q}_2'\tilde{Q}_1'(B + E_b)\tilde{Z}_1\tilde{Z}_2 = \begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} & \tilde{b}_{13} \\ 0 & \tilde{b}_{22} & \tilde{b}_{23} \\ 0 & 0 & \tilde{b}_{33} \end{bmatrix} \quad \text{for } \|E_b\|_2 = \varepsilon_2 \Delta. \tag{31}$$

We prove that there also exists a backward error E_a such that

$$\tilde{Q}_2'\tilde{Q}_1'(A+E_a)\tilde{Z}_1\tilde{Z}_2=\begin{bmatrix}\tilde{a}_{11} & \tilde{a}_{12} & \tilde{a}_{13}\\ 0 & \tilde{a}_{22} & \tilde{a}_{23}\\ 0 & \tilde{a}_{32} & \tilde{a}_{33}\end{bmatrix} \quad \text{for } \|E_a\|_2=\varepsilon_3\Delta. \tag{32}$$

An error analysis of $Q_2'Q_1'AZ_1Z_2$ yields

$$\tilde{Q}_2'\tilde{Q}_1'(A+E_c)\tilde{Z}_1\tilde{Z}_2=\begin{bmatrix}\tilde{a}_{11} & \tilde{a}_{12} & \tilde{a}_{13}\\ \tilde{a}_{21} & \tilde{a}_{22} & \tilde{a}_{23}\\ \tilde{a}_{31} & \tilde{a}_{32} & \tilde{a}_{33}\end{bmatrix} \quad \text{for } \|E_c\|=\varepsilon_4\Delta. \tag{33}$$

We only have to prove that the elements $\tilde{a}_{i1}$, $i=2, 3$ are ε-small, in order to obtain (32) by putting $\tilde{a}_{i1}=0$, $i=2, 3$. This is easily proved using similar reasoning to (24)–(26). Here again the assumption $|b_{33}|\geqq|a_{33}|$ is crucial in the proof of backward stability. Therefore, in the case $|b_{33}|<|a_{33}|$, the roles of B and A have to be interchanged.

Case III. $d_1=1$, $d_2=2$. This case is dual to the previous case and can be reduced to it by pertransposition (transposition over the antidiagonal).

Case IV. $d_1=d_2=2$. A detailed configuration of (11) is then

$$Q'AZ=Q'\begin{bmatrix}a_{11} & a_{12} & a_{13} & a_{14}\\ a_{21} & a_{22} & a_{23} & a_{24}\\ 0 & 0 & a_{33} & a_{34}\\ 0 & 0 & a_{43} & a_{44}\end{bmatrix}Z=\begin{bmatrix}\hat{a}_{11} & \hat{a}_{12} & \hat{a}_{13} & \hat{a}_{14}\\ \hat{a}_{21} & \hat{a}_{22} & \hat{a}_{23} & \hat{a}_{24}\\ 0 & 0 & \hat{a}_{33} & \hat{a}_{34}\\ 0 & 0 & \hat{a}_{43} & \hat{a}_{44}\end{bmatrix}=\hat{A}, \tag{34a}$$

$$Q'BZ=Q'\begin{bmatrix}b_{11} & b_{12} & b_{13} & b_{14}\\ 0 & b_{22} & b_{23} & b_{24}\\ 0 & 0 & b_{33} & b_{34}\\ 0 & 0 & 0 & b_{44}\end{bmatrix}Z=\begin{bmatrix}\hat{b}_{11} & \hat{b}_{12} & \hat{b}_{13} & \hat{b}_{14}\\ 0 & \hat{b}_{22} & \hat{b}_{23} & \hat{b}_{24}\\ 0 & 0 & \hat{b}_{33} & \hat{b}_{34}\\ 0 & 0 & 0 & \hat{b}_{44}\end{bmatrix}=\hat{B}, \tag{34b}$$

where all the elements are real and B and $\hat{B}$ are invertible. In order to have $\Lambda(B_{22}, A_{22})=\Lambda(\hat{B}_{11}, \hat{A}_{11})$ the first two columns of Z must span the deflating subspace of $\lambda B-A$ corresponding to $\Lambda(B_{22}, A_{22})$ or, equivalently, the two (complex) eigenvectors corresponding to the eigenvalues λ_2 and $\bar{\lambda}_2$ of $\Lambda(B_{22}, A_{22})$. Such a Z also satisfies

$$(\lambda_2 I-B^{-1}A)(\bar{\lambda}_2 I-B^{-1}A)Z=\left[\begin{array}{cc|c}0 & 0 & \\ 0 & 0 & \\ 0 & 0 & *\\ 0 & 0 & \end{array}\right], \tag{35}$$

and could be constructed through (35). Unfortunately, this approach is not recommended from a numerical point of view because of the occurrence of B^{-1} and of the product $(\lambda_2 I-B^{-1}A)(\bar{\lambda}_2 I-B^{-1}A)$. An error analysis of (35) would yield a negligible relative error for this product but not for A and B individually.

A different approach is therefore recommended here, namely the double shift QZ-step. Implicitly, this is a double shift QR-step working on the matrix AB^{-1}, but the actual implementation avoids the construction of AB^{-1} and works instead directly on B and A [13]. For our 4×4 pencil (34) the scheme can be implemented economically with Givens rotations:

- Construct $Q_1\in\mathcal{G}_{23}$ and $Q_2\in\mathcal{G}_{12}$ according to the "double shift technique", and construct $Z_1\in\mathcal{G}_{23}$ and $Z_2\in\mathcal{G}_{12}$ such that $Q_2'Q_1'BZ_1Z_2$ is upper triangular.

$Q_2'Q_1'AZ_1Z_2$ and $Q_2'Q_1'BZ_1Z_2$ then look like

$$\begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x_4 & x & x & x \\ x_3 & x_5 & x & x \end{bmatrix} \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix}. \tag{36}$$

• Construct $Q_3 \in \mathcal{G}_{34}$, $Q_4 \in \mathcal{G}_{23}$ and $Q_5 \in \mathcal{G}_{34}$, annihilating x_3, x_4 and x_5, respectively in (36). Construct $Z_3 \in \mathcal{G}_{34}$, $Z_4 \in \mathcal{G}_{23}$ and $Z_5 \in \mathcal{G}_{34}$ such that $Q'BZ$, with $Q = Q_1Q_2Q_3Q_4Q_5$ and $Z = Z_1Z_2Z_3Z_4Z_5$, is upper triangular. $Q'AZ$ is now upper Hessenberg and $Q'BZ$ upper triangular. This form is clearly maintained by a QZ-step. In order to obtain (34) we want moreover that $\hat{a}_{32} = 0$ and $\Lambda(\hat{B}_{22}, \hat{A}_{22}) = \Lambda(B_{11}, A_{11})$. According to the properties of the double shift method [13], this will be the case when $\{\lambda_1, \bar{\lambda}_1\} = \Lambda(B_{11}, A_{11})$ is chosen to determine the double shift (i.e., Q_1 and Q_2), and if in addition $a_{32} \neq 0$. Since in (34) the latter is not satisfied, we first perform a QZ-step with random shift such that $a_{32} \neq 0$, and we then perform a second QZ-step with double shift based on $\{\lambda_1, \bar{\lambda}_1\}$.

The numerical properties of the QZ-step are discussed in [13]. The algorithm is backward stable, but under the presence of rounding errors the element $\hat{a}_{32}$ may not be negligible. Several QZ-steps with double shift $\{\lambda_1, \bar{\lambda}_1\}$ are then performed, and $\hat{a}_{32}$ is shown to converge very fast to zero [13]. Only in pathological cases is more than one step required to obtain $|\hat{a}_{32}| \leqq \varepsilon\Delta$.

Operation count. The combination of a pair of left and right Givens transformations Q_i, Z_i, requires approximately $12n$ operations (1 operation = 1 addition + 1 multiplication). The number of operations for the different cases is then (for Case IV we assume only 2 QZ-steps are needed):

Case I: $12n$.

Case II and III: $32n$ (average).

Case IV: $120n$.

Since Cases II and III correspond to two interchanges of eigenvalues and Case IV to four interchanges, we finally have an average of $20n$ operations for interchanging two adjacent eigenvalues.

When a deflating subspace with a specified spectrum $\{\mu_1, \cdots, \mu_l\}$ has to be computed and a QZ-decomposition is already available, then at most $l \cdot (n-l) \leqq n^2/4$ such interchanges are required (namely when all μ_i, $i = 1, \cdots, l$ are in the bottom right corner). A reasonable estimate is thus $5n^3$ operations for computing a specific deflating subspace from a QZ-decomposition, while the latter requires approximately $25n^3$ operations. In order to obtain all possible orderings of eigenvalues in the QZ decomposition, and thus all possible deflating subspaces (if no eigenvalues are repeated), $n!$ such interchanges are required [9]. That is to be expected since it is a combinatorial problem.

3. Riccati equations. In this section we apply the above ideas to the solution of certain Riccati equations arising in linear system theory. We first briefly restate the four problems we will focus on, and we refer to the literature for a more complete discussion. We will everywhere assume that the matrices involved are real, since this is usually the case in practice. Extensions to the complex case are trivial.

Problem I. *Optimal control: continuous time case* [11][12][24]. Given the stabilizable system

$$\dot{x}(t) = A_{nn}x(t) + B_{nm}u(t), \tag{37}$$

find the control $u(t) = -Kx(t)$ minimizing the functional

$$J = \int_0^\infty [x'(t)Q_{nn}x(t) + u'(t)R_{mm}u(t)]\,dt, \tag{38}$$

where (A, Q) is detectable, $Q \geqq 0$ and $R \geqq 0$. When R is invertible this problem reduces to the computation of the unique nonnegative definite solution P of the algebraic Riccati equation

$$Q + A'P + PA - PBR^{-1}B'P = 0. \tag{39}$$

K is then equal to $R^{-1}B'P$. Equivalently [12], one can compute the invariant subspace $\mathscr{X}_s$ of the matrix

$$H = \begin{bmatrix} A & -BR^{-1}B' \\ -Q & -A' \end{bmatrix}, \tag{40}$$

where $\Lambda(H)|_{\mathscr{X}_s}$ contains all the stable eigenvalues (i.e., $\mathrm{Re}\,(\lambda) < 0$) of H. If $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is a basis for this subspace, then $P = X_2X_1^{-1}$.

Problem II. *Optimal control problem: discrete time case* [5][7][15]. Given the stabilizable system

$$x_{i+1} = F_{nn}x_i + G_{nm}u_i, \tag{41}$$

find the control $u_i = -Kx_i$ minimizing the functional

$$J = \sum_{i=0}^{\infty} [x_i'Q_{nn}x_i + u_i'R_{mm}u_i], \tag{42}$$

where (F, Q) is detectable, $Q \geqq 0$ and $R \geqq 0$.

When R is invertible [7], this problem can again be converted to the computation of the unique nonnegative definite solution P of the (discrete time) algebraic Riccati equation

$$P = F'PF - F'PG(R + G'PG)^{-1}G'PF + Q. \tag{43}$$

K is then equal to $(R + G'PG)^{-1}G'P$. This is also equivalent to solving for the "stable" deflating subspace $\mathscr{X}_s$ of the pencil [5][15]

$$\lambda\begin{bmatrix} I & GR^{-1}G' \\ 0 & F' \end{bmatrix} - \begin{bmatrix} F & 0 \\ -Q & I \end{bmatrix}, \tag{44}$$

where this time the stable eigenvalues are those inside the unit circle. If $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is a basis for $\mathscr{X}_s$, then $P = X_2X_1^{-1}$.

Problem III. *Spectral factorization: continuous time case* [2]. Given an $m \times m$ "positive real" rational matrix $Z(s)$, i.e.,

$$Z(s) \text{ analytic} \quad \text{and} \quad Z(s) + Z^*(s) \geqq 0 \quad \text{in } \mathrm{Re}\,(s) > 0, \tag{45}$$

find a "spectral factorization"

$$Z(s) + Z'(-s) = R(s) \cdot R'(-s), \tag{46}$$

where $R(s)$ has only stable poles and zeros (i.e., $\mathrm{Re}\,(s) < 0$).

When $Z(s)$ is given by a minimal realization $C(sI_n - A)^{-1}B + D$ and $(D + D')$ is invertible, then this problem reduces to the computation of the unique positive definite

solution of the algebraic Riccati equation [2]

$$B(D+D')^{-1}B' + P[A - B(D+D')^{-1}C]' + [A - B(D+D')^{-1}C]P + PC'(D+D')^{-1}CP = 0. \tag{47}$$

This is again equivalent to the computation of the stable invariant subspace $\mathscr{X}_s$ of the matrix

$$H = \begin{bmatrix} A - B(D+D')^{-1}C & B(D+D')^{-1}B' \\ -C'(D+D')^{-1}C & -[A - B(D+D')^{-1}C]' \end{bmatrix}. \tag{48}$$

Problem IV. *Spectral factorization: discrete time case* [1][4]. Given an $m \times m$ "positive real" discrete time matrix $Z(z)$, i.e.,

$$Z(z) \text{ analytic} \quad \text{and} \quad Z(z) + Z^*(z) \geqq 0 \quad \text{for } |z| > 1, \tag{49}$$

find a spectral factorization

$$Z(z) + Z'(z^{-1}) = R(z) \cdot R'(z^{-1}) \tag{50}$$

where $R(z)$ has only stable poles and zeros (i.e., inside the unit circle). Again, when $Z(z)$ is given by a minimal realization $H(zI_n - F)^{-1}G + J$ and $(J + J')$ is invertible, the problem can be reduced to the computation of the unique positive definite solution P of the (discrete time) Riccati equation [4]

$$P = FPF' + (G - FPH')(J + J' - HPH')^{-1}(G' - HPF'). \tag{51}$$

In analogy to (43), (44), one can prove that this is equivalent to computing the stable deflating subspace $\mathscr{X}_s$ of

$$\lambda \begin{bmatrix} I & -G(J+J')^{-1}G' \\ 0 & F' - H'(J+J')^{-1}G' \end{bmatrix} - \begin{bmatrix} F - G(J+J')^{-1}H & 0 \\ -H'(J+J')^{-1}H & I \end{bmatrix}. \tag{52}$$

This, however, was not found in the literature.

Note that in order to be able to write down the Riccati equations, we need certain matrices to be invertible. This also holds for the equivalent SEP's and GEP's, since they are derived from the Riccati equations. Yet, if the matrices to be inverted happen to be badly conditioned, each of these approaches may encounter serious numerical difficulties when computing these inverses. We now present a way to circumvent this by an embedding technique. If D is invertible in the pencil

$$\begin{bmatrix} \lambda E - A & B \\ \underbrace{\lambda F - C}_{p} & \underbrace{D}_{m} \end{bmatrix} \begin{matrix} \}p \\ \}m \end{matrix} \tag{53}$$

then

$$\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \cdot \begin{bmatrix} \lambda E - A & B \\ \lambda F - C & D \end{bmatrix} = \begin{bmatrix} \lambda(E - BD^{-1}F) - (A - BD^{-1}C) & 0 \\ \lambda F - C & D \end{bmatrix}. \tag{54}$$

Let U be an orthogonal transformation reducing $\begin{bmatrix} B \\ D \end{bmatrix}$ to $\begin{bmatrix} 0 \\ \tilde{D} \end{bmatrix}$ with $\tilde{D}$ $m \times m$ and invertible. Partition U conformably with (53); then we have

$$\begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \cdot \begin{bmatrix} \lambda E - A & B \\ \lambda F - C & D \end{bmatrix} = \begin{bmatrix} \lambda \tilde{E} - \tilde{A} & 0 \\ * & \tilde{D} \end{bmatrix}. \tag{55}$$

Since the rows of $[U_{11}|U_{12}]$ and $[I|-BD^{-1}]$ both are a basis for the left null space of $\begin{bmatrix} B \\ D \end{bmatrix}$, they are related by an invertible row transformation which clearly must be U_{11}:

$$U_{11}[I|-BD^{-1}]=[U_{11}|U_{12}] \tag{56}$$

From (54) and (55) it then follows that

$$U_{11}[\lambda(E-BD^{-1}F)-(A-BD^{-1}C)]=\lambda\tilde{E}-\tilde{A}. \tag{57}$$

Therefore the deflating subspaces of $\lambda\tilde{E}-\tilde{A}$ and of $\lambda(E-BD^{-1}F)-(A-BD^{-1}C)$ are the same. According to (7), deflating subspaces of a regular pencil are indeed not affected by an invertible row transformation on the pencil. This technique was also applied in [22] (with $E=I$ and $F=0$) for developing a stable way to compute the deflating subspaces of $\lambda I-(A-BD^{-1}C)$ or, in other words, the invariant subspaces of $A-BD^{-1}C$. This can now be applied to the above four problems. In each of them the pencil (53) takes the form (we always have $p=2n$)

Problem I:

$$\lambda\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}-\begin{bmatrix} A & 0 & B \\ -Q & -A' & 0 \\ 0 & B' & R \end{bmatrix}. \tag{58}$$

Problem II:

$$\lambda\begin{bmatrix} I & 0 & 0 \\ 0 & F' & 0 \\ 0 & G' & 0 \end{bmatrix}-\begin{bmatrix} F & 0 & -G \\ -Q & I & 0 \\ 0 & 0 & R \end{bmatrix}. \tag{59}$$

Problem III:

$$\lambda\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}-\begin{bmatrix} A & 0 & B \\ 0 & -A' & C' \\ C & -B' & D+D' \end{bmatrix}. \tag{60}$$

Problem IV:

$$\lambda\begin{bmatrix} I & 0 & 0 \\ 0 & F' & 0 \\ 0 & G' & 0 \end{bmatrix}-\begin{bmatrix} F & 0 & -G \\ 0 & I & -H' \\ H & 0 & -(J+J') \end{bmatrix}. \tag{61}$$

For each of these pencils a $2n\times 2n$ pencil $\lambda\tilde{E}-\tilde{A}$ can thus be derived via (55), and its stable deflating subspace $\mathscr{X}_s$ is the one required in the above four problems. This procedure does not involve the inversion of a possibly ill-conditioned matrix. Only orthogonal transformations are used as well in the construction of $\lambda\tilde{E}-\tilde{A}$ as in the computation of the deflating subspace $\mathscr{X}_s$. This guarantees the numerical stability of the method. Unfortunately this is not completely satisfactory yet, since the performed errors do not necessarily respect the structure of the pencils (58)–(61). An (unsuccessful) attempt to restrict the orthogonal transformations to those respecting the structure of the matrices they act upon can be found in the literature for Problems I and III, but in the formulation (40) and (48), respectively [14].

An important remark here is that in the new formulation (58)–(61) no inverses occur any more, and that perhaps this new formulation also gives the correct answer when these inverses do not exist. This would follow from limiting arguments if both the exact solution of the problem and the computed solution from the GEP's (58)–(61) are

continuous. This is true for the eigenvalue problem if the spectrum $\Lambda(\tilde{E}, \tilde{A})|_{\mathcal{X}_s}$ is separated from the rest of the spectrum $\lambda\tilde{E} - \tilde{A}$ [20], and this holds under some weak assumptions in each problem (stabilizability, detectability, positive realness). The continuity of the solution $R(s)$ of Problem III is discussed in [1, p. 243]. It also holds for the more general "minimal factorization problem" [3] for which the above embedding technique was originally derived [22]. It is therefore reasonable to assume that it also holds for the other three problems. This is still under current investigation.

During the elaboration of this research, the author's attention was drawn to the work of A. Emami-Naeini and G. Franklin [6]. Via an independent approach they arrive at the same form (59). No proof is provided, though, that the method also works for singular R. The authors of [6] are presently working on that problem.

4. Numerical examples. In this section we give two examples illustrating the reordering of eigenvalues in order to compute a certain deflating subspace with prescribed spectrum. We use a PDP 11-34 computer with double precision. The machine precision is then $\varepsilon \cong 1.5\ 10^{-17}$. Two routines are used for the reordering of the Schur form [25]. EXCHQZ exchanges two adjacent blocks in a real Schur form and ORDER uses this routine to reorder all the eigenvalues inside the unit circle to the top or bottom of the real Schur form, depending on the value of a parameter IFIRST. This last routine is easily adapted for any region which is symmetric with respect to the real axis. This condition is necessary because the pencils considered are real and complex conjugate eigenvalues need thus to stay together in the real Schur form.

Example I.

$$A - \lambda B = \left[\begin{array}{c|ccc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & .3 & .2 & 4 & 6 & 0 & 0 & 0 \\ 0 & -.2 & .3 & 0 & 0 & 0 & .5 & 0 \\ \hline 0 & 0 & 0 & .5 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 4 & 2.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & -10 & 4 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{array}\right] - \lambda \left[\begin{array}{c|ccc|cc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \tag{62}$$

The first four eigenvalues $\{0, .3 - j.2, .3 + j.2, .5\}$ are outside the unit circle. The last four ones $\{\infty, 4 - j5, 4 + j5, 2\}$ are outside the unit circle. Calling ORDER with IFIRST = 1 interchanges the order of these sets of eigenvalues. The first four columns of the transformation Z required for this then span the unstable subspace $\mathcal{X}_u$ of $A - \lambda B$; see Table 1.

When again calling ORDER but now with IFIRST = −1, we retrieve the ordering of $A - \lambda B$ and the four first columns of the updated Z look like Table 2.

This is ε-close to the real stable deflating subspace $\mathcal{X}_s$ of $A - \lambda B$, which is spanned by $\begin{bmatrix} I_4 \\ 0_4 \end{bmatrix}$. This result is to be expected because of the numerical stability of our method and because the space $\mathcal{X}_s$ of $A - \lambda B$ is well-conditioned. When the gap between the spectra $\Lambda(B, A)|_{\mathcal{X}_s}$ and $\Lambda(B, A)|_{\mathcal{X}_u}$ is large, both spaces $\mathcal{X}_s$ and $\mathcal{X}_u$ are indeed well-conditioned (see [20]).

Example II. Consider Problem II with

$$F = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = [0]. \tag{63}$$

TABLE 1

0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.4472135954999579*d*	00
0.0000000000000000*d*	00	0.1074176896329408*d*	00	0.4536669603618238*d*	−01	0.0000000000000000*d*	00
0.0000000000000000*d*	00	−0.6787906375520325*d*	−01	−0.1235432403982217*d*	00	0.0000000000000000*d*	00
0.0000000000000000*d*	00	0.1029985626780152*d*	00	−0.1992532259759830*d*	00	0.0000000000000000*d*	00
0.1000000000000000*d*	01	0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00
0.0000000000000000*d*	00	−0.1485334378807009*d*	00	−0.9610814937405690*d*	00	0.0000000000000000*d*	00
0.0000000000000000*d*	00	−0.9752861049841944*d*	00	0.1389224422842769*d*	00	0.0000000000000000*d*	00
0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.8944271909999159*d*	00

TABLE 2

0.1000000000000000*d*	01	0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00
0.0000000000000000*d*	00	−0.9682688602071642*d*	00	−0.2499108127975237*d*	00	0.3620467925763759*d*	−16
0.0000000000000000*d*	00	−0.2499108127975237*d*	00	0.9682688602071642*d*	00	−0.1119431427711831*d*	−15
0.0000000000000000*d*	00	0.2168404344971009*d*	−17	0.1154675313697062*d*	−15	0.9999999999999999*d*	00
0.0000000000000000*d*	00	−0.3134519117986896*d*	−17	0.1040834085586084*d*	−16	0.4491253335510017*d*	−16
0.0000000000000000*d*	00	0.1040834085586084*d*	−16	0.1543903893619358*d*	−15	−0.5898059818321144*d*	−16
0.0000000000000000*d*	00	−0.3469446951953614*d*	−17	−0.6357761539454998*d*	−15	0.0000000000000000*d*	00
0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00	0.0000000000000000*d*	00

The pencil (59) then looks like

$$\lambda \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{64}$$

An orthogonal row transformation can then be constructed in order to construct a deflated pencil $\lambda\tilde{E} - \tilde{A}$ following (55):

$$\lambda \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{65}$$

The QZ-algorithm permutes the two last columns of (65) to obtain the real Schur form

$$\lambda \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{66}$$

which displays the eigenvalues $\{\infty, \infty, 0, 0\}$. In order to obtain the stable subspace $\mathscr{X}_s$ of $\lambda\tilde{E} - \tilde{A}$ we reorder these eigenvalues and obtain, as a basis for $\mathscr{X}_s$,

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & 0 \\ 0 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & 0 \end{bmatrix} + O(\varepsilon). \tag{67}$$

We then find, up to machine accuracy, the answer $P = I$. One can check that this is the correct answer to Problem II by using another method [7]. This example illustrates that the embedding technique gives a correct result even when R is singular. Moreover, the problem is perfectly well-conditioned as well for the construction of $\lambda\tilde{E} - \tilde{A}$, as for the computation of $\mathscr{X}_s$ and P.

We finally want to draw attention to the fact that the number of operations required for the construction of $\lambda\tilde{E} - \tilde{A}$ from the pencils (58)–(61) is comparable to the amount of work required to construct the pencils (40), (44), (48), (52). From then on, the new approach takes the same amount of computations for Problems II and IV and only slightly more (less than the double) for Problems I and III. The stability of the method and its better conditioning therefore make this new approach particularly attractive.

Acknowledgments. I want to thank A. Emami-Naeini, G. Franklin and A. Laub for drawing my attention to this problem and for several helpful discussions. A. Emami-Naeini also suggested Example II.

REFERENCES

[1] B. ANDERSON AND J. MOORE, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[2] B. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis. A Modern Systems Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[3] H. BART, I. GOHBERG, M. KAASHOEK AND P. VAN DOOREN, *Factorizations of transfer functions*, SIAM J. Control. Optim., 18 (1980), pp. 675–696.

[4] M. DENHAM, *On the factorization of discrete-time rational spectral density matrices*, IEEE Trans. Automat. Contr., AC-20 (1975), pp. 535–537.

[5] A. EMAMI-NAEINI AND G. FRANKLIN, *Design of steady state quadratic loss optimal digital controls for systems with a singular system matrix*, in Proceedings 13th Asilomar Conference on Circ. Syst. & Comp., Nov., 1979, pp. 370–374.

[6] ———, *Comments on "The numerical solution of the discrete time algebraic Riccati equation"*, IEEE Trans. Automat. Contr., AC-25 (1980), pp. 1015–1016.

[7] G. FRANKLIN AND J. POWELL, *Digital Control of Dynamic Systems*, Addison-Wesley, Reading, MA, 1979.

[8] G. GOLUB AND J. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[9] S. JOHNSON, *Generation of permutations by adjacent transposition*, Math. Comp., 17 (1963), pp. 282–285.

[10] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[11] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[12] A. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Contr., AC-24 (1979), pp. 913–921.

[13] C. MOLER AND G. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[14] C. PAIGE AND C. VAN LOAN, *A Hamiltonian-Schur decomposition*, Internal Report, Department of Computer Science, Cornell University, Ithaca, New York, 1979.

[15] T. PAPPAS, A. LAUB AND N. SANDELL JR., *On the numerical solution of the discrete time algebraic Riccati equation*, IEEE Trans. Automat. Contr., AC-25 (1980), pp. 631–641.

[16] G. PETERS AND J. WILKINSON, *$Ax = \lambda Bx$ and the generalized eigenproblem*, SIAM J. Numer. Anal., 7 (1970), pp. 479–492.

[17] A. RUHE, *An algorithm for numerical determination of the structure of a general matrix*, BIT, 10 (1970), pp. 196–216.

[18] G. STEWART, *On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$*, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.

[19] ———, *Introduction to Matrix Computation*, Academic Press, New York, 1973.

[20] ———, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[21] ———, *Algorithm 506: HQR3 and EXCHNG. Fortran subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix*, ACM TOMS, 2 (1976), pp. 275–280.

[22] P. VAN DOOREN, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Contr., AC-26 (1981), to appear.

[23] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[24] M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

[25] P. VAN DOOREN, *A generalized eigenvalue approach for solving Riccati equations*, Internal Report NA-80-02, Department of Computer Science, Stanford University, Stanford, CA, 1980.

SIAM J. SCI. STAT. COMPUT.
Vol. 4, No. 4, December 1983

0196-5204/83/0404-0016 $01.25/0

ERRATUM: A GENERALIZED EIGENVALUE APPROACH FOR SOLVING RICCATI EQUATIONS*

P. VAN DOOREN†

In the above paper, some errors appeared in the treatment of the spectral factorization problems for both the continuous-time and the discrete-time case.

The corrections to be performed are the following. Formulas (46), (47), (50) and (51) have to be replaced by their "dual" forms which appear below:

$$Z(s)+Z'(-s)=R'(-s)\cdot R(s), \tag{46}$$

$$\begin{aligned} &C'(D+D')^{-1}C+P[A-B(D+D')^{-1}C] \\ &\quad +[A-B(D+D')^{-1}C]'P+PB(D+D')^{-1}B'P=0, \end{aligned} \tag{47}$$

$$Z(z)+Z'(z^{-1})=R'(z^{-1})\cdot R(z), \tag{50}$$

$$P=F'PF+(H'-F'PG)(J+J'-G'PG)^{-1}(H-G'PF). \tag{51}$$

Acknowledgment. The author is grateful to Bert van Gent who pointed out these errors.

* This Journal, 2 (1981), pp. 121-135.

‡ Philips Research Laboratory, B-1170 Brussels, Belgium.