New SLICOT Routines for Standard and Descriptor Systems

Vasile Sima[1] and Andreas Varga[2]
[1]*National Institute for Research & Development in Informatics, 8–10 Bd. Mareşal Averescu, Bucharest, Romania*
[2]*DLR-Oberpfaffenhofen (Retired), Gilching, Germany*

Keywords: Descriptor Systems, Generalized Eigenvalues, Numerical Methods, Software, Stability.

Abstract: New routines included into the SLICOT Library for standard and descriptor systems are presented. The underlying numerical methods, the functionality and the main features of the added software are described. The topics covered include stable/unstable and finite/infinite spectrum separation, additive spectral decomposition, and removing the non-dynamic modes. The main implementation issues are also addressed. Numerical results obtained using the software on a large set of examples of various complexity and difficulty are summarized. The results reported highlight the performance and capabilities of this SLICOT Library extension.

1 INTRODUCTION

Consider a linear time-invariant system in a descriptor state-space representation of the form

$$E\lambda x(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t), \quad (1)$$

with $x(t)$ the n-dimensional state vector, $u(t)$ and $y(t)$ the m-dimensional and p-dimensional input and output vectors, respectively, and E, A, B, C, and D the descriptor, state, input, output, and feedthrough matrices of suitable sizes, with E and A square and E possibly singular. According to the system type, $\lambda x(t) := \dot{x}(t)$ or $\lambda x(t) := x(t+1)$, for a continuous- or discrete-time system, respectively. The system (1) is alternatively denoted as $(A-\lambda E, B, C, D)$, where D can be omitted if $D = 0$. The case $E = I_n$, the identity matrix of order n, corresponds to a standard state-space realization and any descriptor realization with nonsingular E can be reduced to an equivalent standard system $(E^{-1}A - \lambda I_n, E^{-1}B, C, D)$. However, this reduction must be generally avoided if E is ill-conditioned with respect to inversion.

If the matrix pencil $A-\lambda E$ is regular (i.e., $\det(A-\lambda E) \not\equiv 0$), then (1) can also be considered as an order n descriptor system realization of the rational $p \times m$ transfer-function matrix,

$$G(\lambda) = C(\lambda E - A)^{-1}B + D, \quad (2)$$

where λ is here a complex variable corresponding to the Laplace transform or Z-transform. Any $p \times m$ rational matrix $G(\lambda)$ has a descriptor realization of the form (1) which satisfies (2). Also, there exist minimal order realizations, with least possible order n. This connection between descriptor systems and rational matrices is the basis for the development of numerically reliable algorithms for manipulating rational matrices via their equivalent descriptor realizations.

Continuous-time descriptor systems are often used for modelling certain mechanical systems, e.g., with contact phenomena, or interconnected systems with algebraic loops, while discrete-time descriptor systems are often encountered for modelling economic processes. Descriptor representations are essential for appropriate numerical operations with rational (transfer-function) matrices, as highlighted in (Varga, 2017) for solving fault diagnosis problems.

Two linear matrix pencils play an important role in defining and characterizing the properties of transfer-function matrices (2) via their equivalent descriptor realizations (1). The regular pole pencil $P(\lambda) = A - \lambda E$ is useful to characterize the pole structure of (2) via its Weierstrass canonical form, while the system matrix pencil

$$S(\lambda) = \left[\begin{array}{cc} A-\lambda E & B \\ C & D \end{array} \right],$$

defines the zero and singularity structures of (1) via its Kronecker canonical form. The knowledge of the pole and zero structures allows to state simple conditions to characterize important properties of the transfer-function matrix (2), such as properness, stability, minimum-phase, or of the descriptor system (1), such as controllability, observability, finite or infinite stabilizability or detectability, or ir-

Sima, V. and Varga, A.
New SLICOT Routines for Standard and Descriptor Systems.
DOI: 10.5220/0006396605350542
In *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2017) - Volume 1*, pages 535-542
ISBN: 978-989-758-263-9

reducibility. The numerical computation of Weierstrass and Kronecker canonical forms involves possibly ill-conditioned (non-orthogonal) transformation matrices. Numerically reliable algorithms and procedures for descriptor systems, as discussed in the following sections, resort to alternative forms, like generalized Schur form and Kronecker-like forms, which provide all the above structural information and are obtainable using orthogonal transformations.

Descriptor systems are investigated in several books, e.g., (Duan, 2010), and many technical papers, e.g., (Demmel and Kågström, 1993; Kågström and Van Dooren, 1990; Misra et al., 1994; Van Dooren, 1981; Varga, 1990; Varga, 1996). Related numerical issues are addressed, for instance, in (Anderson et al., 1999; Demmel and Kågström, 1993; Golub and Van Loan, 2012; Van Dooren, 1981; Varga, 1996; Varga, 2004; Varga, 2017).

This paper presents the main new routines included by the authors into the SLICOT Library[1] for standard and descriptor systems, the underlying numerical techniques, as well as a summary of the results obtained using this software on a large set of examples of various complexity and difficulty from the COMPl$_e$ib collection (Leibfritz and Lipinski, 2003). The organization of the paper is as follows. Section 2 summarises the main numerical techniques for stable/unstable and finite/infinite spectrum separation, additive spectral decomposition, and non-dynamic modes removal. Section 3 describes the functionality of the main newly added SLICOT routines for standard and descriptor systems, as well as the essential implementation issues. Section 4 summarizes part of the numerical results obtained, highlighting the performance and capabilities of this SLICOT Library extension. Section 5 contains some conclusions.

2 NUMERICAL TECHNIQUES FOR DESCRIPTOR SYSTEMS

Many procedures for descriptor systems analysis and control design involve the orthogonal reduction of a matrix pencil $A - \lambda E$, defined by a pair of square matrices, (A, E), to the *generalized real Schur form* (GRSF), using the standard QZ algorithm, see e.g., (Golub and Van Loan, 2012). The transformed pair is $(\widetilde{A}, \widetilde{E})$, with $\widetilde{A} = Q^T AZ$, $\widetilde{E} = Q^T EZ$, where $Q^T Q = QQ^T = I$, $Z^T Z = ZZ^T = I$, $\widetilde{E}$ is upper triangular, and $\widetilde{A}$ is block upper triangular, with diagonal blocks of order 1 and 2, corresponding to the real and complex conjugate eigenvalues of the pencil, respectively. (If the 2×2 pairs of diagonal blocks may have real eigenvalues, the pair is called *upper quasi-triangular*.) The GRSF can be reordered, using also orthogonal transformation matrices, so that the eigenvalues appear in any desired order along the diagonal blocks of the transformed matrix pair. The cost of reordering is small compared to the initial reduction cost. The real Schur form (RSF) of a matrix, $\widetilde{A} = Z^T AZ$, and an orthogonal transformation matrix, Z, are similarly used for standard systems. These computations can be performed using state-of-the-art linear algebra software, such as LAPACK (Anderson et al., 1999) or MATLAB®. The main advantage of using orthogonal transformations is that the problem conditioning is preserved, so that the errors in the data are essentially not magnified during the calculations.

Several key problems for descriptor systems are summarized below.

2.1 Stable/Unstable Separation

The reordering of GRSF can be used, for instance, to separate the stable and unstable parts of a system. Let $(A - \lambda E, B, C)$ be a descriptor system (1), and let α define the desired boundary of the stability domain, with $\alpha \leq 0$ for the real parts of the eigenvalues, and $\alpha \leq 1$ for the moduli of the eigenvalues, for continuous- and discrete-time systems, respectively. (Choosing $\alpha = -\tau$ or $\alpha = 1 - \tau$ in the continuous- or discrete-time case, respectively, where $\tau > 0$, imposes a certain stability degree τ.) For convenience, assume here that E is nonsingular. Denote $\Lambda(A, E)$ the *spectrum* of the pencil $A - \lambda E$, i.e., the set of its eigenvalues, $\{\lambda_i, \, i = 1 : n\}$ (multiplicities counted). Let Q and Z be orthogonal matrices so that

$$\begin{aligned} Q^T AZ \quad &=: \quad \widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ 0 & \widetilde{A}_{22} \end{bmatrix}, \\ Q^T EZ \quad &=: \quad \widetilde{E} = \begin{bmatrix} \widetilde{E}_{11} & \widetilde{E}_{12} \\ 0 & \widetilde{E}_{22} \end{bmatrix}, \end{aligned} \qquad (3)$$

with $\widetilde{A} - \lambda\widetilde{E}$ in GSRF, $\widetilde{A}_{jj}$, $\widetilde{E}_{jj}$ of order n_j, $j = 1, 2$, and

$$\begin{aligned} \Lambda(\widetilde{A}_{11}, \widetilde{E}_{11}) \quad &= \quad \{\lambda_i \,|\, \Re(\lambda_i) < \alpha\}, \\ \Lambda(\widetilde{A}_{22}, \widetilde{E}_{22}) \quad &= \quad \{\lambda_i \,|\, \Re(\lambda_i) \geq \alpha\}, \end{aligned}$$

for a continuos-time system, and similarly, for a discrete-time system, when $\Re(\lambda_i)$, the real part of λ_i above, is replaced by $|\lambda_i|$, the modulus of λ_i. Therefore, the system $(\widetilde{A} - \lambda\widetilde{E}, \widetilde{B}, \widetilde{C})$, with $\widetilde{B} := Q^T B$, and $\widetilde{C} := CZ$, is an equivalent system having all stable eigenvalues in the leading positions. Then, the first n_1 columns of the matrices Q and Z form orthogonal

[1] www.slicot.org

bases for the left and right stable deflating subspaces of (A,E). When E is singular, the infinite eigenvalues can be deflated, and the remaining system can be handled in a similar way. The procedure is, however, more involved.

2.2 Block Diagonalization

Some applications, as, for example, the computation of additive spectral decompositions of transfer-function matrices, require that the separated parts in (3) be decoupled (i.e., $\widetilde{A}_{12} = 0$ and $\widetilde{E}_{12} = 0$). This can be achieved by block-diagonalization with two diagonal blocks. However, the procedure involves non-orthogonal transformations. Here, it is not necessary to make any stability related assumption in (3), but only assuming that the spectra of $(\widetilde{A}_{11},\widetilde{E}_{11})$ and $(\widetilde{A}_{22},\widetilde{E}_{22})$ are disjoint, i.e., $\Lambda(\widetilde{A}_{11},\widetilde{E}_{11})\bigcap\Lambda(\widetilde{A}_{22},\widetilde{E}_{22}) = \emptyset$. (A GRSF is needed to proceed with efficient computations.) It is possible to simultaneously annihilate $\widetilde{A}_{12}$ and $\widetilde{E}_{12}$ by solving for X and Y the *generalized Sylvester equation* (Kågström and Van Dooren, 1990):

$$\widetilde{A}_{11}Y - X\widetilde{A}_{22} = \sigma\widetilde{A}_{12},\ \widetilde{E}_{11}Y - X\widetilde{E}_{22} = \sigma\widetilde{E}_{12}, \quad (4)$$

where $0 \le \sigma \le 1$ is a scaling factor, set to avoid overflow in X and Y. The condition on disjoint sets of eigenvalues theoretically guarantees the existence of a solution to (4) with $\sigma = 1$. The transformation matrices for decoupling are

$$Q = \begin{bmatrix} I_{n_1} & X \\ 0 & I_{n_2} \end{bmatrix},\ Z = \begin{bmatrix} I_{n_1} & -Y \\ 0 & I_{n_2} \end{bmatrix},$$

so that the two diagonal pairs of submatrices of $(Q\widetilde{A}Z, Q\widetilde{E}Z)$ coincide to those of $(\widetilde{A},\widetilde{E})$, while the off-diagonal blocks, $\widetilde{A}_{12}$ and $\widetilde{E}_{12}$, are annihilated. The transformed descriptor system with $(\widetilde{A},\widetilde{E})$ in block diagonal form has $\widetilde{B} = QB$, and $\widetilde{C} = CZ$. The matrix D is unchanged. With a partition of $\widetilde{B} = \begin{bmatrix}\widetilde{B}_1^T\ \widetilde{B}_2^T\end{bmatrix}^T$ and $\widetilde{C} = \begin{bmatrix}\widetilde{C}_1\ \widetilde{C}_2\end{bmatrix}$, then $(\widetilde{A}_1 - \lambda\widetilde{E}_1, \widetilde{B}_1, \widetilde{C}_1, D)$ and $(\widetilde{A}_2 - \lambda\widetilde{E}_2, \widetilde{B}_2, \widetilde{C}_2)$ represent an additive spectral decomposition of the system transfer-function matrix.

When $E = I_n$, the similarity transformation to a system with state matrix in block-diagonal form is

$$\widetilde{A} = U^{-1}AU,\ \widetilde{B} = U^{-1}B,\ \widetilde{C} = CU,\ U = \begin{bmatrix} I_{n_1} & -X \\ 0 & I_{n_2} \end{bmatrix},$$

where X is the solution of the *standard Sylvester equation*, $\widetilde{A}_{11}X - X\widetilde{A}_{22} = \sigma\widetilde{A}_{12}$. The inverse of U is obtained by replacing $-X$ by X.

2.3 Finite/Infinite Separation

When E is singular, some applications require decompositions of the form in (3) with the (1,2) blocks nonzero or zero, but with a separation between finite and infinite eigenvalues. Either finite or infinite eigenvalues can appear in the resulting leading diagonal block pair. Let assume that the regular pole pencil $A - \lambda E$ of the descriptor system $(A - \lambda E, B, C)$ has been transformed to one of the forms (5) or (6)

$$Q^TAZ = \begin{bmatrix} A_f & \star \\ 0 & A_i \end{bmatrix},\ Q^TEZ = \begin{bmatrix} E_f & \star \\ 0 & E_i \end{bmatrix}, \quad (5)$$

$$Q^TAZ = \begin{bmatrix} A_i & \star \\ 0 & A_f \end{bmatrix},\ Q^TEZ = \begin{bmatrix} E_i & \star \\ 0 & E_f \end{bmatrix}, \quad (6)$$

where the subpencil $A_f - \lambda E_f$ contains the finite eigenvalues, the subpencil $A_i - \lambda E_i$ contains the infinite eigenvalues, and $\star$ denotes sumatrices whose values are not important in this context. Clearly, this transformation requires that E_f and A_i be nonsingular; in implementations, both E_f and A_i are upper triangular. Other procedures than those in Subsections 2.1 and 2.2 are needed in this case. Specifically, the reduction algorithm in (Misra et al., 1994) can be used. For (5), the first step of the algorithm reduces the system to a *SVD-like coordinate form,*

$$Q^TAZ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad Q^TEZ = \begin{bmatrix} E_r & 0 \\ 0 & 0 \end{bmatrix},$$

where E_r is an upper triangular nonsingular matrix of order $r = \text{rank}(E) \le n$. This can be obtained using either singular value decomposition (SVD), or, more efficiently, a (truncated) QR decomposition with column pivoting, followed by a special RQ decomposition (Golub and Van Loan, 2012). Then, the A_{22} matrix is reduced similarly to the *QR-form*

$$A_{22} = \begin{bmatrix} A_r & X \\ 0 & 0 \end{bmatrix},$$

with A_r an upper triangular nonsingular matrix of order $r_2 = \text{rank}(A_{22})$, and X a full matrix. The second step performs the finite-infinite separation. The computations are done on a block column permutation of the matrices obtained in the first step, rewritten as

$$A := \begin{bmatrix} B_1 & A_1 \\ D_1 & C_1 \end{bmatrix},\ E := \begin{bmatrix} 0 & E_1 \\ 0 & 0 \end{bmatrix}, \quad (7)$$

where, initially, $E_1 = E_r$, $A_1 = A_{11}$, $C_1 = A_{21}$, $B_1 = A_{12}$, and $D_1 = A_{22}$. The same block column permutation acts on the transformed output matrix C. A finite sequence of iterations is then applied to the matrices in (7), exploiting the current structure. Initially, the order of the matrix D_1 is $n - r$, with the last $\rho = n - r - r_2$ rows zero, and the leading $r_2 \times r_2$ submatrix, A_r, upper triangular. At each subsequent iteration, the current matrix D_1 is row compressed using standard and rank-revealing QR factorizations with column pivoting, and the transformations are applied to B and C_1 from the left. Then, the bottom part of C_1, corresponding to the zeroed part in D_1, is column compressed via RQ factorization, while keeping E_1 upper triangular. The transformations are also applied to A_1, B_1, B, from the left, and to A_1 and C, from the right. The matrices D_1 and C_1 are redefined after omitting the zeroed parts. The process ends when D_1 has maximal row rank. Another block column permutation returns

$$A = \begin{bmatrix} A_1 & B_1 & \star \\ C_1 & D_1 & \star \\ 0 & 0 & A_i \end{bmatrix}, \; E = \begin{bmatrix} E_1 & 0 & \star \\ 0 & 0 & \star \\ 0 & 0 & E_i \end{bmatrix},$$

and C_1 is annihilated by rotations from the right (with pivots in the diagonal elements of D_1). Finally, the matrices A_i and E_i have the staircase forms

$$A_i = \begin{bmatrix} A_{0,0} & A_{0,k} & \cdots & A_{0,1} \\ 0 & A_{k,k} & \cdots & A_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{1,1} \end{bmatrix}, \; E_i = \begin{bmatrix} 0 & E_{0,k} & \cdots & E_{0,1} \\ 0 & 0 & \cdots & E_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{8}$$

where $A_{j,j}$, $j = 0, 1, \ldots, k$, are nonsingular upper triangular submatrices.

The same procedure is used for (6) by working on the dual system (and interchanging m with p, and Q with Z), and transforming the results by *pertransposition*,

$$A := PA^TP,\; E := PE^TP,\; B := PC^T,\; C := B^TP,$$

to preserve the shapes of A and E, where P is a matrix with 1 on the secondary diagonal, and with 0 in the other entries. (The transformation matrices are also updated as $Q := QP$, $Z := ZP$.) Finally, the matrices A_i and E_i have the staircase forms

$$A_i = \begin{bmatrix} A_{1,1} & \cdots & A_{1,k} & A_{1,0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & A_{k,k} & A_{k,0} \\ 0 & \cdots & 0 & A_{0,0} \end{bmatrix}, \; E_i = \begin{bmatrix} 0 & E_{1,k} & \cdots & E_{1,0} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E_{k,0} \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{9}$$

where $A_{j,j}$, $j = 0, 1, \ldots, k$, are nonsingular upper triangular submatrices. In both cases, the pair $(A_{0,0}, 0)$ contains the system non-dynamic infinite modes.

2.4 Removing the Non-dynamic Modes

The reduction to the SVD-like coordinate form can be used to elliminate the non-dynamic modes of a descriptor system, using the following procedure:

1. Reduce the system to the SVD-like coordinate form $(Q^TAZ - \lambda Q^TEZ, Q^TB, CZ)$, where
$$\begin{aligned} Q^TAZ &= \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & 0 \\ A_{31} & 0 & 0 \end{bmatrix}, \\ Q^TEZ &= \begin{bmatrix} E_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; Q^TB = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}, \\ CZ &= \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix}, \end{aligned} \tag{10}$$
where E_{11} and A_{22} are upper triangular invertible matrices.
2. Compute the reduced descriptor system without non-dynamic modes as $(A_r - \lambda E_r, B_r, C_r, D_r)$, where
$$\begin{aligned} A_r &= \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & A_{13} \\ A_{31} & 0 \end{bmatrix}, \\ E_r &= \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \; B_r = \begin{bmatrix} B_1 - A_{12}A_{22}^{-1}B_2 \\ B_3 \end{bmatrix}, \\ C_r &= \begin{bmatrix} C_1 - C_2A_{22}^{-1}A_{21} & C_3 \end{bmatrix}, \\ D_r &= D - C_2A_{22}^{-1}B_2 . \end{aligned} \tag{11}$$
3. Optionally, reduce the descriptor system to the normalized form with $A_r := \mathrm{diag}(E_{11}^{-1}, I)A_r$, $B_r := \mathrm{diag}(E_{11}^{-1}, I)B_r$, $E_r = \mathrm{diag}(I, 0)$.

The reduced system order is $n - r_2$, $r_2 := \mathrm{rank}(A_{22})$.

3 IMPLEMENTATION ISSUES

The functionality of several new main routines for standard and descriptor systems, added to the SLICOT Library, is briefly described.

`MB03QG` reorders the diagonal blocks of a selected principal subpencil of an upper quasi-triangular matrix pencil $A - \lambda E$, together with their generalized eigenvalues, using orthogonal equivalence transformations Q and Z as in (3). After reordering, the leading block of the subpencil has generalized eigenvalues in a suitably defined domain of interest, usually related to stability/instability in a continuous- or discrete-time sense. The system type, stability option, and a bound α for the real parts or moduli of the generalized eigenvalues have to be specified on input.

`TB01PX` finds a reduced (controllable, observable, or minimal) state-space representation (A_r, B_r, C_r) for an original state-space representation (A, B, C) using the following procedure:

1. If a minimal or controllable realization is desired, the pair (A, B) is reduced by orthogonal similarity transformations to the controllability staircase

form (Van Dooren, 1981), and a controllable realization (A_c, B_c, C_c) is extracted, where A_c results in an upper block Hessenberg form.

2. If a minimal or observable realization is desired, the same algorithm is applied to the dual of the system (A_c, B_c, C_c) or (A, B, C), respectively, to extract an observable realization (A_r, B_r, C_r). In the first case, the resulting realization is also controllable, and thus minimal. The state matrix A_r is in an upper block Hessenberg staircase form.

`TB01UY` finds a controllable realization for the linear time-invariant multi-input system

$$\lambda x = Ax + B_1 u_1 + B_2 u_2, \quad y = Cx,$$

where A, B_1, B_2, and C are $n \times n$, $n \times m_1$, $n \times m_2$, and $p \times n$ matrices, respectively, and A and $\begin{bmatrix} B_1 \; B_2 \end{bmatrix}$ are reduced to an orthogonal canonical form using (and optionally accumulating) orthogonal similarity transformations, which are also applied to C. Specifically, the system $(A, \begin{bmatrix} B_1 \; B_2 \end{bmatrix}, C)$ is reduced to the triplet $(\widetilde{A}, \begin{bmatrix} \widetilde{B}_1 \; \widetilde{B}_2 \end{bmatrix}, \widetilde{C})$, where $\widetilde{A} = U^T A U$, $\begin{bmatrix} \widetilde{B}_1 \; \widetilde{B}_2 \end{bmatrix} = U^T \begin{bmatrix} B_1 \; B_2 \end{bmatrix}$, $\widetilde{C} = CU$, with

$$\widetilde{A} = \begin{bmatrix} A_c & \star \\ 0 & A_{nc} \end{bmatrix}, \quad \begin{bmatrix} \widetilde{B}_1 & \widetilde{B}_2 \end{bmatrix} = \begin{bmatrix} B_{c_1} & B_{c_2} \\ 0 & 0 \end{bmatrix},$$

and A_c and $\begin{bmatrix} B_{c_1} \; B_{c_2} \end{bmatrix}$ given by

$$A_c = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1,q-2} & A_{1,q-1} & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2,q-2} & A_{2,q-1} & A_{2q} \\ A_{31} & A_{32} & \cdots & A_{3,q-2} & A_{3,q-1} & A_{3q} \\ 0 & A_{42} & \cdots & A_{4,q-2} & A_{4,q-1} & A_{4q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & A_{q,q-2} & A_{q,q-1} & A_{qq} \end{bmatrix},$$

$$\begin{bmatrix} B_{c_1} & B_{c_2} \end{bmatrix} = \begin{bmatrix} A_{1,-1} & A_{1,0} \\ 0 & A_{2,0} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \qquad (12)$$

where the block $A_{i,i-2}$ is a $\tau_i \times \tau_{i-2}$ full row rank matrix (with $\tau_{-1} = m_1$, $\tau_0 = m_2$), $i = 1, \ldots, q$, and $q/2$ is the controllability index of the pair $(A, \begin{bmatrix} B_1 \; B_2 \end{bmatrix})$. The size of the block A_{nc} is equal to the dimension of the uncontrollable subspace of the pair $(A, \begin{bmatrix} B_1 \; B_2 \end{bmatrix})$. The implemented algorithm (Varga, 2003) represents a specialization of the controllability staircase algorithm of (Varga, 1981) to the special structure of B.

`TG01HU` is primarily intended for a descriptor system $(A - \lambda E, B, C)$ with nonsingular E, where $B = \begin{bmatrix} B_1 \; B_2 \end{bmatrix}$ is an $n \times (m_1 + m_2)$ matrix, with B_i an $n \times m_i$ submatrix, $i = 1, 2$, and it reduces the pair $(A - \lambda E, [B_1 \; B_2])$ to the form

$$Q^T \begin{bmatrix} B_1 & B_2 & A - \lambda E \end{bmatrix} \mathrm{diag}(I_{m_1+m_2}, Z) = \begin{bmatrix} B_{c_1} & B_{c_2} & A_c - \lambda E_c & \star \\ 0 & 0 & 0 & A_{nc} - \lambda E_{nc} \end{bmatrix}, \quad (13)$$

where Q and Z are orthogonal, and

1) the pencil $\begin{bmatrix} B_{c_1} \; B_{c_2} \; A_c - \lambda E_c \end{bmatrix}$ has full row rank n_r for all finite $\lambda \in \mathbf{C}$, and it is defined by (12) and

$$E_c = \begin{bmatrix} E_{11} & E_{12} & \cdots & E_{1q} \\ 0 & E_{22} & \cdots & E_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E_{qq} \end{bmatrix},$$

where E_{ii} is a $\tau_i \times \tau_i$ upper triangular matrix.

2) the pencil $A_{nc} - \lambda E_{nc}$ is regular, of order $n - n_r$, with E_{nc} upper triangular, and it contains the uncontrollable finite eigenvalues of the pencil $A - \lambda E$ and possibly some of the uncontrollable infinite eigenvalues.

The right transformations are also applied to the matrix C. The implemented algorithm (Varga, 2004) is a specialization of the controllability staircase algorithm in (Varga, 1990) to the special structure of B. The implementation is more general: A and E are upper block triangular, and E may have a given number of nonzero subdiagonals.

`TG01GD` finds a reduced descriptor representation (11), $(A_r - \lambda E_r, B_r, C_r, D_r)$, without nondynamic modes, for a descriptor representation $(A - \lambda E, B, C, D)$. Optionally, the reduced descriptor system can be put into a standard form with the leading diagonal block of E_r identity.

`TG01LD` computes orthogonal transformation matrices Q and Z which reduce the regular pole pencil $A - \lambda E$ of the descriptor system $(A - \lambda E, B, C)$ to the form (5) or to the form (6), where the subpencil $A_f - \lambda E_f$, with E_f nonsingular and upper triangular, contains the finite eigenvalues, and the subpencil $A_i - \lambda E_i$, with A_i nonsingular and upper triangular, contains the infinite eigenvalues. The subpencil $A_i - \lambda E_i$ is in a staircase form (8) or (9), respectively. Optionally, the submatrix A_f is further reduced to an upper Hessenberg form.

`TG01MD` computes orthogonal transformation matrices Q and Z which reduce the regular pole pencil $A - \lambda E$ of the descriptor system $(A - \lambda E, B, C)$ to the form (5) or (6), where the pair (A_f, E_f) is in a GRSF, with E_f nonsingular and upper triangular, and A_f in RSF. The subpencil $A_f - \lambda E_f$ contains the finite eigenvalues. The pair (A_i, E_i) is in a GRSF with both A_i and E_i upper triangular. The subpencil $A_i - \lambda E_i$, with A_i nonsingular and E_i nilpotent, contains the infinite eigenvalues, and it is in a block staircase form (8) or (9), respectively.

TG01ND computes equivalence transformation matrices Q and Z which reduce the regular pole pencil $A-\lambda E$ of the descriptor system $(A-\lambda E,B,C)$ to one of the forms

$$QAZ=\left[\begin{array}{cc} A_f & 0 \\ 0 & A_i \end{array}\right],\quad QEZ=\left[\begin{array}{cc} E_f & 0 \\ 0 & E_i \end{array}\right],$$

$$QAZ=\left[\begin{array}{cc} A_i & 0 \\ 0 & A_f \end{array}\right],\quad QEZ=\left[\begin{array}{cc} E_i & 0 \\ 0 & E_f \end{array}\right],$$

where the pair (A_f,E_f) is in a GRSF, with E_f nonsingular and upper triangular, and A_f in RSF. The subpencil $A_f-\lambda E_f$ contains the finite eigenvalues. The pair (A_i,E_i) is in a GRSF with both A_i and E_i upper triangular. The subpencil $A_i-\lambda E_i$, with A_i nonsingular and E_i nilpotent, contains the infinite eigenvalues, and it is in a block staircase form (8) or (9), respectively. This decomposition corresponds to an additive decomposition of the transfer-function matrix of the descriptor system as the sum of a proper term and a polynomial term.

TG01PD computes orthogonal transformation matrices Q and Z which reduce the regular pole pencil $A-\lambda E$ of the descriptor system $(A-\lambda E,B,C)$ to a GRSF with ordered generalized eigenvalues. The pair (A,E) is reduced to the form

$$Q^TAZ=\left[\begin{array}{cccc} \star & \star & \star & \star \\ 0 & A_1 & \star & \star \\ 0 & 0 & A_2 & \star \\ 0 & 0 & 0 & \star \end{array}\right],\; Q^TEZ=\left[\begin{array}{cccc} \star & \star & \star & \star \\ 0 & E_1 & \star & \star \\ 0 & 0 & E_2 & \star \\ 0 & 0 & 0 & \star \end{array}\right],$$

where the subpencil $A_1-\lambda E_1$ contains the eigenvalues which belong to a suitably defined domain of interest, and the subpencil $A_2-\lambda E_2$ contains the eigenvalues which are outside of the domain of interest. Optionally, the pair (A,E) is assumed to be already in a GRSF and the reduction is performed only on the subpencil $A_{12}-\lambda E_{12}$ defined by rows and columns from l to u of $A-\lambda E$, with $1\le l\le u\le n$.

TG01QD computes orthogonal transformation matrices Q and Z which reduce the regular pole pencil $A-\lambda E$ of the descriptor system $(A-\lambda E,B,C)$ to a GRSF with ordered generalized eigenvalues. The pair (A,E) is reduced to the form

$$Q^TAZ=\left[\begin{array}{ccc} A_1 & \star & \star \\ 0 & A_2 & \star \\ 0 & 0 & A_3 \end{array}\right],\; Q^TEZ=\left[\begin{array}{ccc} E_1 & \star & \star \\ 0 & E_2 & \star \\ 0 & 0 & E_3 \end{array}\right],$$

where the subpencils $A_k-\lambda E_k$, for $k=1,2,3$, contain the generalized eigenvalues which belong to certain domains of interest.

Most of the algorithms are backward stable. The number of floating point operations is of the order of n^3. State-of-the-art numerical linear algebra algorithms are used, including QR and RQ factorization with pivoting, incremental condition estimation, Householder transformations and Givens rotations, etc. BLAS 3 calls are used whenever possible, to increase efficiency. Optimal workspace size can optionally be precomputed by several subroutines. Options are available to deal individually with the one or two transformation matrices. The options are 'N' (do not compute a transformation matrix), 'I' (compute it with identity initialization), or 'U' (update the matrix given on entry). TG01ND has an option to compute the direct or inverse transformation matrices, and TG01ND and TG01QD have an option to exhibit either the finite or the infinite eigenvalues in the leading positions.

4 NUMERICAL RESULTS

An extensive testing has been performed to evaluate the new routines. Some examples from the literature have been used to verify the correctness of the delivered results. For a performance investigation, examples from the COMPl$_e$ib collection (Leibfritz and Lipinski, 2003) have been used. The collection contains 124 standard continuous-time examples, but with variations, a total of 168 systems can be defined. All 151 systems with order less than 1000, and one larger system, have been tried. Note that some COMPl$_e$ib examples were derived from systems with general, but nonsingular matrix E, by multiplying the matrices in the state equation in (1) by E^{-1} from the left. These examples have been modified for this paper in order to be solved as descriptor (actually, generalized) systems. Therefore, results for both standard as well as generalized systems will be presented in the following two subsections.

The computations have been performed in double precision on an Intel Core i7-3820QM portable computer (2.7 GHz, 16 GB RAM), under Windows 7 Professional (Service Pack 1) operating system (64 bit), with Intel Visual Fortran Composer XE 2015, and MATLAB R2015b. Executable MEX-files have been built using SLICOT routines and MATLAB-provided optimized LAPACK and BLAS routines.

4.1 Results for Standard Systems

The set of examples tried have been obtained calling the MATLAB script examplevector from COMPl$_e$ib, with names selected with the string vector Ex_no_sl. The default tolerance, set by the routines to $n^2\varepsilon_M$, where ε_M is the relative machine precision, has been used for rank computations.

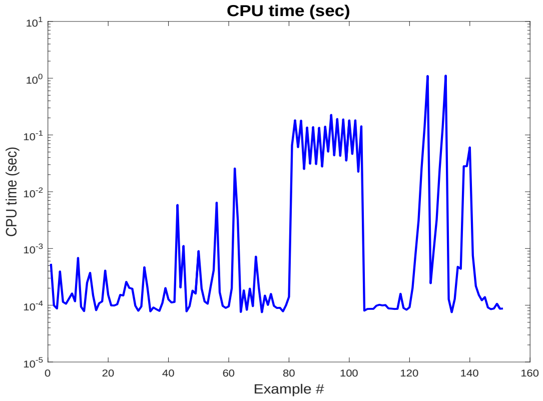

Figure 1: CPU times for computing minimal realizations for 151 standard systems chosen from the COMPl$_e$ib collection.

Figure 1 shows the CPU execution times needed to compute minimal realizations for the 151 standard systems. The maximum CPU time is 1.11 seconds, for the `CM6_IS` example, with $n = 960$, $m = 1$, and $p = 2$. Each of the other examples was solved in less than 0.23 seconds. Moreover, there are 117 examples each solved in less than 0.01 seconds, including 111 solved in less than one millisecond.

There are 32 systems found non-minimal. The reduction took place in step 1 (for 17 examples), and in step 2 (for 21 examples) of the minimal realization procedure in Section 3 (`TB01PX` description). For six examples, the order decreased in both steps.

The separation of the systems spectra into stable and unstable parts has also been investigated. The stability degree τ has been set to -10^{-8}. The CPU times are comparable to those in Fig. 1.

Figure 2 shows the relative errors of the stable and unstable poles after separation, compared to the original ones. Their number is preserved for all problems. There are 44 stable systems. The zero values of the errors have been replaced by a value slightly smaller than the smallest relative error, so that its logarithm can be computed. The largest relative error is for example `CSE2`, and it is due to a pair of real, very close poles, with relative difference of $2.7877 \cdot 10^{-9}$.

4.2 Results for Generalized Systems

As mentioned before, a collection of 34 generalized systems has been derived from the COMPl$_e$ib examples for which the matrix E was available, usually in binary `mat` files (31 examples). Among these, 8 examples (`HF2Di`, with i = 1, 2, ..., 8) have orders greater than or equal to 2025, and only `HF2D1` has been separately tried. Only two examples, `TL` and `FS`, have condition numbers for E larger than 4, namely,

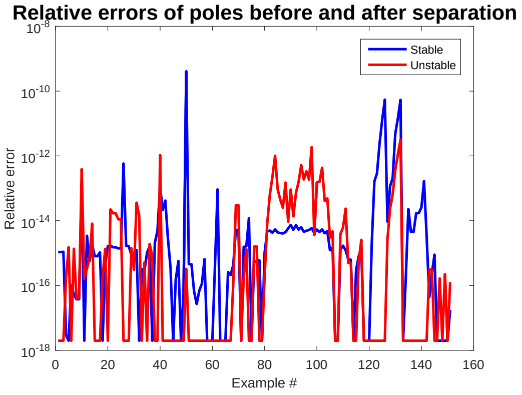

Figure 2: Relative errors of the stable and unstable poles after separation, compared to the original ones, for 151 standard systems chosen from the COMPl$_e$ib collection.

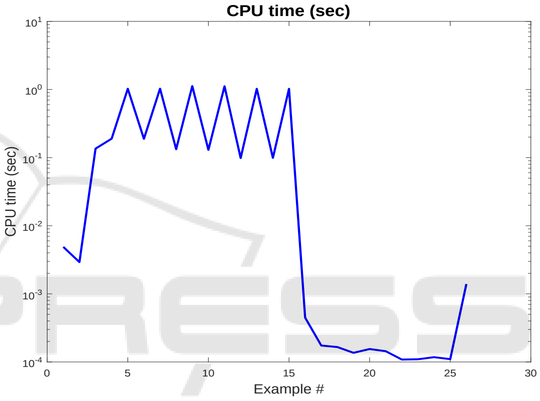

Figure 3: CPU times for computing minimal realizations for 26 generalized systems, with order $n \leq 541$, generated from the COMPl$_e$ib collection.

$7.7579 \cdot 10^6$ and $3.77 \cdot 10^5$, respectively.

Figure 3 shows the CPU execution times needed to compute minimal realizations for the 26 generalized systems with order $n \leq 541$. The maximum CPU time is 1.12 seconds, for the modified `HF2D1_M541` example, with $n = 541$, $m = 2$, and $p = 3$. Six examples (those with $n \geq 529$) needed slighly more than one second runtime. Each of the other examples was solved in less than 0.25 seconds. Moreover, there are 13 examples each solved in less than 0.01 seconds, including 10 solved in less than one millisecond.

Except for timing, the results reproduced those for the corresponding standard systems. All, but two examples have been found as minimal. The original non-minimal examples are `CSE1` and `CSE2`; for each of them the order has been reduced by 1.

The modified example `HF2D1`, with $n = 3796$, $m = 2$, and $p = 3$, has also been tried, and found as minimal. The CPU time has been 795 seconds. On the other hand, solving the corresponding standard system needed 119 seconds.

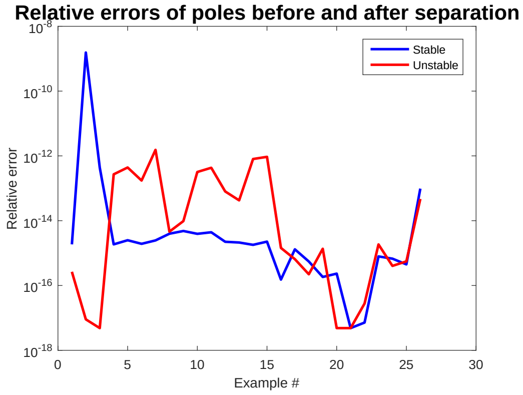

Figure 4: Relative errors of the stable and unstable poles after separation, compared to the original ones, for 26 generalized systems, with order $n \leq 541$.

The separation of the systems spectra into stable and unstable parts has also been investigated. For descriptor systems this includes a preliminary separation into finite and infinite parts. The infinite part is void, since matrix E is nonsingular (but moderately ill-conditioned for TL and FS examples). The stability degree τ has been set to -10^{-8}, and the tolerance has been set to 0, i.e., a default value was used. The CPU times are comparable to those in Fig. 3.

Figure 4 shows the relative errors of the stable and unstable poles after separation, compared to the original ones. Their number is preserved for all problems. There are 3 stable systems: TL, HF2D12, and HF2D13. Example FS has 3 unstable poles, and each of the other 22 examples have one unstable pole. The largest relative error, recorded for the CSE2 example, is due to a pair of complex poles with imaginary parts of about $\pm 6.5854 \cdot 10^{-9}$, while the reordered poles became real. Such a change is entirely motivated theoretically. Omitting these two poles, the relative error for CSE2 example becomes $2.7174 \cdot 10^{-15}$.

5 CONCLUSIONS

Numerical techniques and procedures for computing stable/unstable and finite/infinite spectrum separation, additive spectral decomposition, and removing the non-dynamic modes, have been discussed. These techniques and procedures represent the theoretical and practical foundation for the new routines included by the authors into the SLICOT Library for standard and descriptor systems. The functionality of several main new routines has been briefly described, and their essential features highlighted. Numerical results obtained on a comprehensive set of examples from the COMPl$_e$ib collection have been summarized and illustrate the performance and capabilities of this SLICOT Library extension.

REFERENCES

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide: Third Edition*. SIAM, Philadelphia.

Demmel, J. W. and Kågström, B. (1993). The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part I: Theory and algorithms. Part II: Software and applications. *ACM Trans. Math. Software*, 19:160–174, 175–201.

Duan, G.-R. (2010). *Analysis and Design of Descriptor Linear Systems*, vol. 23 of *Advances in Mechanics and Mathematics*. Springer, New York.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, fourth edition.

Kågström, B. and Van Dooren, P. (1990). Additive decomposition of a transfer function with respect to a specified region. In *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems, MTNS-89*, Amsterdam, The Netherlands, Birkhäuser, Boston.

Leibfritz, F. and Lipinski, W. (2003). Description of the benchmark examples in *COMPl$_e$ib*. Technical report, University of Trier, Germany.

Misra, P., Van Dooren, P., and Varga, A. (1994). Computation of structural invariants of generalized state-space systems. *Automatica*, 30(12):1921–1936.

Van Dooren, P. (1981). The generalized eigenstructure problem in linear system theory. *IEEE Trans. Automat. Contr.*, AC–26:111–129.

Varga, A. (1981). Numerically stable algorithm for standard controllability form determination. *Electronics Letters*, 17:74–75.

Varga, A. (1990). Computation of irreducible generalized state-space realizations. *Kybernetika*, 26(2):89–106.

Varga, A. (1996). Computation of Kronecker-like forms of a system pencil: Applications, algorithms and software. In *Proceedings of the IEEE International Symposium on Computer-Aided Control System Design*, Dearborn, MI, 77–82. IEEE.

Varga, A. (2003). Reliable algorithms for computing minimal dynamic covers. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawai'i USA, 1873–1878. IEEE.

Varga, A. (2004). Reliable algorithms for computing minimal dynamic covers for descriptor systems. In *Proceedings of the Mathematical Theory of Networks and Systems, MTNS 2004*, Leuven, Belgium.

Varga, A. (2017). *Solving Fault Diagnosis Problems: Linear Synthesis Techniques*. Springer, Berlin.