

Article

On Schur Forms for Matrices with Simple Eigenvalues

Mihail Mihaylov Konstantinov ¹  and Petko Hristov Petkov ^{2,*} 
¹ Department of Mathematics, University of Architecture, Civil Engineering and Geodesy, 1064 Sofia, Bulgaria; mmk_fte@uacg.bg

² Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

* Correspondence: php@tu-sofia.bg

Abstract: In this paper, we consider various aspects of the Schur problem for a square complex matrix A , namely the similarity unitary transformation of A into upper triangular form containing the eigenvalues of A on its diagonal. Since the profound work of I. Schur published in 1909, this has become a fundamental issue in the theory and applications of matrices. Nevertheless, certain details concerning the Schur problem need further clarification, especially in connection with the perturbation analysis of the Schur decomposition relative to perturbations in the matrix A . We consider both canonical and condensed Schur forms. Special attention is paid to matrices with simple eigenvalues. Some new concepts, such as quasi-Schur forms and diagonally spectral matrices, are also introduced and studied.

Keywords: Schur canonical form; Schur condensed form; diagonally spectral matrix; quasi-Schur form; perturbations of Schur form

MSC: 15A21; 65G30



Citation: Konstantinov, M.M.; Petkov, P.H. On Schur Forms for Matrices with Simple Eigenvalues. *Axioms* **2024**, *13*, 839. <https://doi.org/10.3390/axioms13120839>

Academic Editor: Hari Mohan Srivastava

Received: 10 October 2024

Revised: 24 November 2024

Accepted: 27 November 2024

Published: 28 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction and Notation

The Schur decomposition of a general square matrix [1] and its generalizations are major tools both in the theory and applications of matrix analysis; see, e.g., refs. [2–5]. In this paper, we consider the main definitions and properties of the Schur decomposition of a square matrix (see the definitions given later), which are important from the point of view of the perturbation analysis of the Schur problem; see [6–10]. Various aspects of Schur and Schur-like decompositions are considered in [11–14]. The monographs [15,16] also deal with these problems. Schur-like forms are also proposed for pairs of matrices (or matrix pencils) [3], for control systems [17–19], etc. A perturbation analysis of these constructions is presented in [18,20,21] and the references therein.

In this paper, we introduce new concepts in this field such as quasi-canonical Schur forms and diagonally spectral matrices. A number of examples are given for illustration of the results presented. This is a specific issue, and we shall need a large number of notations. For the convenience of the reader, the general notations are gathered below in this section, while some specific notations appear further in the text. Some of the matrix notations are inspired by the language of the program system MATLAB[®] Version 9.9 (R2020b) [22].

If $\mathcal{K}$ is a finite set, then $\text{card}(\mathcal{K})$ is the number of its elements. Let $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ be the set of integers and $m, n \in \mathbb{Z}$, where $m \leq n$. We denote by $\mathbb{Z}[m, n] = \{k \in \mathbb{Z} : m \leq k \leq n\}$ the set of $n - m + 1$ integers $m, m + 1, \dots, n$. We write $\mathbb{Z}[m, m] = m$ and $\mathbb{Z}[n, m] = \emptyset$ when $n > m$.

The set of real (resp. complex) numbers is denoted by $\mathbb{R}$ (resp. $\mathbb{C} \simeq \mathbb{R} \times \mathbb{R}$), and $i = \sqrt{-1}$ is the imaginary unit. A quantity $0 \neq z \in \mathbb{C} \setminus \mathbb{R}$ is said to be genuinely complex. An array (vector or matrix) is genuinely complex if at least one of its elements is genuinely complex.

A complex number z is written as $z = x + iy$ with $x, y \in \mathbb{R}$, or $z = |z| \exp(i\varphi)$, where $|z| = \sqrt{x^2 + y^2}$ is the absolute value and $\varphi \in (-\pi, \pi]$ is the angle of z . The complex conjugate of z is denoted as $\bar{z} = x - iy = |z| \exp(-i\varphi)$.

The sign function $\text{sign} : \mathbb{R} \rightarrow \mathbb{Z}[-1, 1]$ of the scalar argument is defined as $\text{sign}(x) = -1$, $\text{sign}(x) = 0$ and $\text{sign}(x) = 1$ for $x < 0$, $x = 0$ and $x > 0$, resp. The sign function for real n -tuples $x = (x_1, x_2, \dots, x_n)$ is defined by the expression

$$\text{sign}(x) = \sum_{k=1}^n 2^{1-k} \text{sign}(x_k).$$

The lexicographical order $\prec$ for n -tuples is defined as $x \prec \tilde{x}$, $x = \tilde{x}$ and $x \succ \tilde{x}$ if $\text{sign}(x - \tilde{x}) < 0$, $\text{sign}(x - \tilde{x}) = 0$ and $\text{sign}(x - \tilde{x}) > 0$, resp. Otherwise, $x \prec \tilde{x}$ when either $x_1 < \tilde{x}_1$, or there exists $m \in \mathbb{Z}[2, n]$ such that $x_k = \tilde{x}_k$ for $k \in \mathbb{Z}[1, m-1]$ and $x_m < \tilde{x}_m$.

We use this lexicographical order for complex numbers $z = x + iy$ written as real pairs (x, y) and for pairs (m, n) of integers $m, n \in \mathbb{Z}$. For example, for the fourth roots $\{\pm 1, \pm i\}$ of 1, we have $-1 \prec -i \prec i \prec 1$.

We denote by $\mathbb{C}(m, n)$ (resp. $\mathbb{R}(m, n)$) the space of $m \times n$ complex (resp. real) matrices $A = [A(i, j)]$ with elements $A(i, j)$ and we set $\mathbb{C}(n) = \mathbb{C}(n, n)$, $\mathbb{R}(n) = \mathbb{R}(n, n)$. The i -th row and the j -th column of A are denoted as $A(i, :) \in \mathbb{C}(1, n)$ and $A(:, j) \in \mathbb{C}(m, 1)$, respectively.

The column m -vector $b \in \mathbb{C}(m, 1)$ with elements $b_i = b(i)$ is written as $b = [b_1; b_2; \dots; b_m]$, while the row n -vector $c \in \mathbb{C}(1, n)$ with elements $c_j = c(j)$ is denoted as $c = [c_1, c_2, \dots, c_n]$.

The Kronecker delta symbol is $d(i, j) = 1 - \text{sign}^2(i - j)$. The identity $n \times n$ matrix is denoted as $I_n \in \mathbb{R}(n)$ and has elements $I_n(i, j) = d(i, j)$. The anti-identity matrix $P_n \in \mathbb{R}(n)$ has elements $P_n(i, j) = d(i, n + 1 - j)$.

The zero $m \times n$ matrix is denoted as $O_{m,n} \in \mathbb{R}(m, n)$ with $O_n = O_{n,n}$, or simply as O . We denote by $L_n \in \mathbb{R}(n)$ the strictly lower triangular matrix with ones below its main diagonal and zeros otherwise, i.e., $L_n(i, j) = 1$ if $i > j$ and $L_n(i, j) = 0$ if $i \leq j$. The elementary matrix with element 1 in position (p, q) and zero otherwise is denoted as $E_{p,q} \in \mathbb{R}(m, n)$, i.e., $E_{p,q}(i, j) = d(i, p)d(j, q)$.

The absolute value of the matrix $A \in \mathbb{C}(m, n)$ is the matrix $|A| \in \mathbb{R}(m, n)$ with elements $|A|(i, j) = |A(i, j)|$. The transpose of the matrix $A \in \mathbb{C}(m, n)$ is denoted $A^\top \in \mathbb{C}(n, m)$ and has elements $A^\top(i, j) = A(j, i)$. The complex conjugate transpose of $A \in \mathbb{C}(m, n)$ is denoted $A^H \in \mathbb{C}(n, m)$ and has elements $A^H(i, j) = \overline{A(j, i)}$.

The inverse of the non-singular matrix A is denoted as A^{-1} . The flipped matrix $A^F \in \mathbb{C}(m, n)$ of the matrix $A \in \mathbb{C}(m, n)$ is $A^F = P_n A P_n$ and has elements $A^F(i, j) = A(m + 1 - i, n + 1 - j)$.

For $A, B \in \mathbb{C}(m, n)$, we denote by $A \circ B \in \mathbb{C}(m, n)$ the element-wise product of A and B , i.e., $(A \circ B)(i, j) = A(i, j)B(i, j)$. The spectral and the Frobenius norms of the matrix $A \in \mathbb{C}(m, n)$ are denoted as $\|A\|$ and $\|A\|_F$, respectively.

The spectrum $\text{spect}(A)$ of the matrix $A \in \mathbb{C}(n)$ is the collection, or the multiset, of the eigenvalues $\lambda_k(A) \in \mathbb{C}$ of A , $k \in \mathbb{Z}[1, n]$, counted according to their algebraic multiplicities. With certain abuse of notation, we write $\text{spect}(A) \subset \mathbb{C}$ in the general case, and $\text{spect}(A) \subset \mathbb{R}$ in the case when all eigenvalues of A are real.

The multiplicative group of non-singular matrices $U \in \mathbb{C}(n)$ is denoted as $\mathbf{GL}(n)$, while the group of unitary matrices $U \in \mathbb{C}(n)$ such that $U^H U = I_n$ is denoted by $\mathbf{U}(n)$. The group of orthonormal matrices $U \in \mathbb{R}(n)$ such that $U^\top U = I_n$ is denoted as $\mathbf{O}(n)$.

For $A \in \mathbb{C}(n)$, we denote by $\text{Low}(A) = A \circ L_n$ and $\text{Diag}(A) = A \circ I_n$ the strictly lower triangular and the diagonal parts of A , respectively. If x is an n -vector with elements $x(i) \in \mathbb{C}$ then $\text{diag}(x) \in \mathbb{C}(n)$ is the matrix with elements $\text{diag}(x)(i, j) = x(i)d(i, j)$.

The set of upper triangular matrices $A = A - \text{Low}(A)$ is denoted as $\mathbf{T}(n) \subset \mathbb{C}(n)$, while the set of diagonal matrices $A = \text{Diag}(A)$ is denoted as $\mathbf{D}(n) \subset \mathbf{T}(n)$. For $n \geq 2$, the group of diagonal matrices of the form

$$\text{diag}(1, \exp(i\varphi_2), \dots, \exp(i\varphi_n)),$$

where $\varphi_k \in \mathbb{R}$, is denoted as $\mathbf{D}^*(n) \subset \mathbf{U}(n)$.

The set of $(n-1)$ -tuples of pairs $\{(i_1, j_1), (i_2, j_2), \dots, (i_{n-1}, j_{n-1})\}$ of integers $i_k \in \mathbb{Z}[1, n-1]$, $j_k \in \mathbb{Z}[2, n]$, where $i_k < j_k$, is denoted as $\mathcal{K}(n)$.

We set $v_n = n(n-1)/2$ and

$$\mu_n = \text{card}(\mathcal{K}(n)) = \frac{v_n!}{(n-1)!(v_n - n + 1)!}$$

In particular, we have $v_4 = 6$ and $\mu_4 = 6!/3!/3! = 20$.

Unspecified matrix blocks are denoted by star.

We use the following abbreviations: CSF—canonical Schur form; ConSF—condensed Schur form; DPS—diagonal preserving solution; GJF—generalized Jordan form; PSP—perturbed Schur problem; SP—Schur problem.

2. Condensed Schur Forms

Let a nonzero matrix $A \in \mathbb{C}(n)$, $n \geq 2$, be given. A condensed form UAV of A relative to a set of invertible matrices $U, V \in \mathbb{C}(n)$ is a matrix with preliminary fixed elements (usually zeros and/or ones) at certain positions. Most often, the condensed form is a triangular matrix with at least v_n zero elements. According to the famous Schur result [1], there exists a factorization $A = UTU^H$ of the matrix A , where $U \in \mathbf{U}(n)$ and $T \in \mathbf{T}(n)$. Two things here deserve mentioning: the proof of the Schur result is elementary, and it appeared too late in the history of mathematics.

Definition 1. *The pair*

$$(U, T) \in \mathbf{U}(n) \times \mathbf{T}(n), \quad T = U^H A U, \quad (1)$$

is said to be a Schur decomposition, or an upper triangular unitary decomposition of the matrix A . The matrix T is referred to as a condensed Schur form, or ConSF of A . The columns of the unitary transformation matrix U form a Schur basis for the space $\mathbb{C}(n, 1)$ relative to the matrix A .

Thus, the definition of the ConSF T of A is not unique, and hence, it is not canonical in the sense of Definitions 7 and 8 below. If $A \in \mathbb{R}(n)$ and $\text{spect}(A) \subset \mathbb{R}$ is real, then the matrix U may be chosen as $U \in \mathbf{O}(n)$, and we have $T = U^T A U \in \mathbb{R}(n)$. If $A \in \mathbb{R}(n)$ has at least one pair of complex eigenvalues $\alpha \pm i\beta$, $\beta \neq 0$, then a real block-diagonal Schur form with diagonal blocks $\lambda \in \mathbb{R}$ and $\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \in \mathbb{R}(2)$ may be constructed by orthogonal similarity transformations; see e.g., refs. [23,24].

Next, we define two sets of matrices depending on the matrix A that play an important role in our analysis. Denote

$$\mathcal{U}(A) = \{U \in \mathbf{U}(n) : U^H A U \in \mathbf{T}(n)\} \subset \mathbf{U}(n)$$

and

$$\mathcal{T}(A) = \{U^H A U : U \in \mathcal{U}(A)\} \subset \mathbf{T}(n).$$

Thus, $\mathcal{U}(A) \subset \mathbf{U}(n)$ is the set of unitary matrices transforming the matrix A into ConSF, and $\mathcal{T}(A)$ is the set of ConSF of A . For matrices $A \in \mathbb{R}(n)$ with real spectra, we denote

$$\mathcal{O}(A) = \{U \in \mathbf{O}(n) : U^T A U \in \mathbf{T}(n)\} \subset \mathbf{O}(n).$$

In general, the set $\mathcal{U}(A)$ is not a group and not even a groupoid, i.e., $U_1, U_2 \in \mathcal{U}(A)$ does not imply $U_1 U_2 \in \mathcal{U}(A)$.

The most important property of the ConSF T of A is that its diagonal elements are the eigenvalues of A . Another application of ConSF is the evaluation of functions of matrices defined by a matrix power series or by Padé approximations [25].

Because of the only condition $\text{Low}(T) = 0$ imposed on T , the matrix T is a condensed form (rather than a canonical form) of A relative to the similarity action $\mathbf{U}(n) \times \mathbb{C}(n) \rightarrow \mathbf{T}(n)$, defined by $(U, A) \mapsto U^H A U$, of the group $\mathbf{U}(n)$ on the set $\mathbb{C}(n)$.

Definition 2. The problem of finding the ConSF (1) is referred to as the Schur problem (SP) for the matrix $A \in \mathbb{C}(n)$. The general solution of the SP is the set

$$\mathbf{S}(A) = \left\{ (U, U^H A U) : U \in \mathbf{U}(n), U^H A U \in \mathbf{T}(n) \right\} \subset \mathcal{U}(A) \times \mathcal{T}(A)$$

of all ConSF of A . A pair $(U, T) \in \mathbf{S}(A)$ is a particular solution of the SP for the matrix A .

Sometimes the matrices U and T in a particular solution (U, T) of SP for A are written as $U(A)$ and $T(A)$ to emphasize their dependence on A . This dependence, however, is not a functional one. Indeed, the matrix U is always not unique. For example, $(U, T) \in \mathbf{S}(A)$ and $|V| = I_n$ imply $(UV, V^H T V) \in \mathbf{S}(A)$ and, in particular, $(-U, T) \in \mathbf{S}(A)$. With the exception of the case $A = \lambda I_n$, $\lambda \in \mathbb{C}$, when it is fulfilled $T(A) = A$, the matrix $T(A)$ is also not unique. For $A = \lambda I_n$, we have $\mathbf{S}(A) = \mathbf{U}(n) \times \{A\}$.

All upper triangular unitary equivalent forms of a given matrix are unitary similar. In particular, the next proposition is a direct corollary of the definitions; see, e.g., ref. [24].

Proposition 1. Let $\Pi_1 = (U_1, T_1)$ and $\Pi_2 = (U_2, T_2)$ be two solutions of the SP for A . Then, $T_2 = U_2^H U_1 T_1 U_1^H U_2$.

Proof. It suffices to observe that $A = U_1 T_1 U_1^H = U_2 T_2 U_2^H$. $\square$

Definition 3. The solutions Π_1 and Π_2 are said to be diagonally equal if $\text{Diag}(T_1) = \text{Diag}(T_2)$, and diagonally different if $\text{Diag}(T_1) \neq \text{Diag}(T_2)$.

The next proposition has generally been known since 1933 and is attributed to H. Röseler; see Theorem 2.3 of [24]. It gives sufficient and almost necessary conditions for diagonal equality of the solutions of SP. The formulation and proof of the results below are slightly different from the known ones.

Proposition 2. The following assertions hold true.

1. If $V = U_1^H U_2 \in \mathbf{D}^*(n)$, then the solutions Π_1 and Π_2 are diagonally equal.
2. If the matrix A has pair-wise distinct eigenvalues and the solutions Π_1 and Π_2 are diagonally equal then $V \in \mathbf{D}^*(n)$.

Proof. To prove Assertion 1 note that the condition $V \in \mathbf{D}^*(n)$ is equivalent to the existence of a matrix $D \in \mathbf{D}^*(n)$ such that $U_2 = U_1 D$. In this case

$$T_1(i, j) = D(i, i) \overline{D(j, j)} T_2(i, j), \quad T_1(i, i) = T_2(i, i)$$

and $\text{Diag}(T_1) = \text{Diag}(T_2)$.

To prove 2, we use the fact that $T_1 V = V T_2$. Partition the matrices in this equality as

$$T_1 = \begin{bmatrix} \lambda & * \\ 0 & \Lambda \end{bmatrix}, \quad T_2 = \begin{bmatrix} \lambda & * \\ 0 & * \end{bmatrix}, \quad V = \begin{bmatrix} \mu & u \\ v & W \end{bmatrix},$$

where $\lambda \in \text{spect}(A)$, $\Lambda \in \mathbb{C}(n-1)$, $\mu \in \mathbb{C}$, $W \in \mathbb{C}(n-1)$ and $*$ is a matrix block of corresponding size. We have

$$T_1 V = \begin{bmatrix} * & * \\ \Lambda v & \Lambda W \end{bmatrix}, \quad V T_2 = \begin{bmatrix} \lambda \mu & * \\ \lambda v & * \end{bmatrix}$$

and comparing the (2,1)-blocks of these matrices, we obtain $\Lambda v = \lambda v$. Since $\lambda \notin \text{spect}(\Lambda)$, we obtain $v = 0$. Hence $|\mu| = 1$, $u = 0$ and $V = \text{diag}(\mu, W)$. Now, the proof is completed by induction. $\square$

The MATLAB[®] command $[U, T] = \text{schur}(A)$ computes a particular solution (U, T) of SP for a matrix $A \in \mathbb{C}(n)$. One of the aims of computing a ConSF T of a general matrix $A \in \mathbb{C}(n)$ is to determine the eigenvalues $\lambda_k(A)$ of A as the diagonal elements $T(k, k)$ of $T \in \mathbf{T}(n)$. Another aim is to evaluate matrix functions

$$f(A) = f(UTU^H) = Uf(T)U^H,$$

where $f(A)$ is a power matrix series [25] in A .

A particular challenge is to define SP for matrices $A \in \mathbf{T}(n)$, which are already in ConSF. Here, the problem is not to transform A into ConSF (it already is), but rather to find the matrices $U \in \mathbf{U}(n)$ that keep $A \in \mathbf{T}(n)$ in ConSF, i.e., $U^H A U \in \mathbf{T}(n)$.

For matrices $A \in \mathbf{T}(n)$, the MATLAB[®] command $[U, T] = \text{schur}(A)$ returns the solution $U = I_n$, $T = A$ of the SP. Next, for $1 \leq m < n$ and a matrix

$$A = \begin{bmatrix} A_1 & * \\ 0 & A_2 \end{bmatrix}, \quad A_1 \notin \mathbf{T}(n-m), \quad A_2 \in \mathbf{T}(m)$$

the computed solution $[U, T]$ is of the form

$$U = \begin{bmatrix} U_1 & 0 \\ 0 & I_m \end{bmatrix}, \quad T = \begin{bmatrix} T_1 & * \\ 0 & A_2 \end{bmatrix}, \quad T_1 = U_1^H A_1 U_1$$

where $U_1 \in \mathbf{U}(n-m)$. In particular, for $A = I_n$, the command $[U, T] = \text{schur}(\text{eye}(n))$ returns $U = I_n$, $T = I_n$.

The MATLAB[®] (Maple) command $[V, J] = \text{jordan}(A)$ for computing an invertible matrix V and a Jordan canonical form $J = V^{-1}AV$ of A , in case $A = I_n$, also gives $V = I_n$, $J = I_n$.

It is interesting to reveal the action of existing computer codes on matrices A close to I_n in double-precision floating-point arithmetic.

Example 1. Let $n = 2$ and $A(\varepsilon) = I_2 + \varepsilon E_{2,1}$, where ε is small. We have $T = I_2 \pm \varepsilon E_{1,2}$. For $\varepsilon = 2^{-51}$, the solution of SP for $A(\varepsilon)$ is computed wrongly by the MATLAB[®] code `schur` as $U = I_2$, $T = I_2$. Choosing ε as the next machine number, which is $2^{-51} + 2^{-103}$, the code `schur` computes exactly one of the solutions of SP for $A(\varepsilon)$, namely $U = E_{2,1} - E_{1,2}$, $T = I_2 - \varepsilon E_{1,2}$.

The code `jordan` exactly computes the Jordan canonical form (V, J) of A as $V = E_{1,2} + \varepsilon E_{2,1}$, $J = I_2 + E_{1,2}$ for values of $\varepsilon > 0$ as small as 2^{-1074} . For $\varepsilon = 2^{-1075}$, the matrix $A(\varepsilon)$ is rounded to I_2 , and the computed Jordan form is $V = I_2$, $J = I_2$.

Note that the code `jordan` works with variable precision arithmetic.

The SP for $A \in \mathbf{T}(n)$ has infinitely many solutions (U, T) for both U and T with one exception for T . Namely, the matrix T is uniquely determined only when $A = \lambda I_n$, where $\lambda \in \mathbb{C}$. In this case, $T = A$ and $U \in \mathbf{U}(n)$ is arbitrary, i.e., the general solution of SP is $\mathbf{U}(n) \times \{A\}$. It is also interesting to see whether the order of the diagonal elements of $A \in \mathbf{T}(n)$ is the same as the order of the diagonal elements of $T \in \mathbf{T}(n)$. These considerations lead to the following definition.

Definition 4. Let $A \in \mathbf{T}(n)$. Then, the pair (U, T) , where $U \in \mathbf{U}(n)$ and $T = U^H A U \in \mathbf{T}(n)$, is said to be a diagonal preserving solution (DPS) of the SP for A if $\text{Diag}(T) = \text{Diag}(A)$.

Otherwise speaking, the pair (U, T) is the DPS of SP when the upper triangular matrices A and T are diagonally equal, see Definition 3. The DPS thus defined as a

particular solution of SP. The general DPS of SP for $A \in \mathbf{T}(n)$ is the set of all particular DPS solutions. In practice, we are interested in finding a particular DPS rather than determining the general DPS as a subset of $\mathbf{S}(A)$.

If the matrix $A \in \mathbf{T}(n)$ has simple eigenvalues, the matrix U in any particular DPS (U, T) is diagonal, i.e., $|U| = I_n$. To illustrate this, let $n = 2$ and

$$A = \text{diag}(\lambda, 0) + aE_{1,2}, \quad T = \text{diag}(\lambda, 0) + tE_{1,2}, \quad |t| = |a|,$$

where $\lambda \neq 0$. We have $(AU)(2, 1) = 0$ and $(UT)(2, 1) = \lambda U(2, 1)$, which yields $U(2, 1) = 0$ and $|U| = I_2$.

If the matrix $A \in \mathbf{T}(n)$ has multiple eigenvalues but its upper triangular part is generically nonzero, i.e., $A(p, q) \neq 0$ for $p > q$, then the matrix U in any particular DPS (U, T) is again diagonal. To illustrate this, let $n = 2$ and $A = aE_{1,2}$, $T = tE_{1,2}$, where $|t| = |a| > 0$. We have $(AU)(1, 1) = aU(2, 1)$ and $(UT)(1, 1) = 0$, which yields $U(2, 1) = 0$ and $|U| = I_2$.

For all $A \in \mathbf{T}(n)$, a simple choice for U in any particular DPS is $U = I_n$, which allows treating all cases in a unified manner. Choosing any other matrix U with $|U| = I_n$ is of course also possible. In the real case, orthogonal diagonal matrices U have diagonal elements ± 1 .

The case when the matrix A has multiple eigenvalues and its upper triangular part is not generically nonzero, e.g., $A = \lambda I_n$ or $A = [\lambda I_m, *, 0, A_2]$, where $1 \leq m < n$ and $A_2 \notin \mathbf{T}(n - m)$, needs special consideration. Here, the matrix U in a particular DPS may not be diagonal. For example, in the most non-generic case $A = \lambda I_n$, any pair (U, A) with $U \in \mathbf{U}(n)$ is a particular solution of SP, and hence, the general solution of SP is $\mathbf{U}(n) \times \{A\}$. Anyway, for definiteness, we choose the particular solution with $U = I_n$ and $T = A$ of SP each time when $A \in \mathbf{T}$.

Any particular choice of U in a DPS for matrices A with multiple eigenvalues and a non-generic upper triangular part may lead to jump effects, as described below.

Suppose that $A = A_1(\varepsilon)$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ is the matrix function of a scalar argument, where $\varepsilon_0 > 0$ is a small parameter. If the matrix $A_1(0) \in \mathbf{T}(n)$ has simple eigenvalues (or multiple eigenvalues but generic upper triangular part), then the matrix $U = U_1(\varepsilon)$ in a DPS for $A_1(\varepsilon)$ may be chosen so that $U_1(0) = I_n$. Here, the function $U = U_1(\varepsilon)$ will be continuous in the interval $(-\varepsilon_0, \varepsilon_0)$. If, however, $A_1(0)$ has multiple eigenvalues and a non-generic upper triangular part, then the choice $U_1(0) = I_n$ may lead to jump of the function $U = U_1(\varepsilon)$ at the point $\varepsilon = 0$.

We may use another function $U = U_2(\varepsilon)$ with $U_2(0) \neq I_n$ in the DPS of the SP for $A_1(0) \in \mathbf{T}(n)$, such that U_2 is continuous at $\varepsilon = 0$ for the above choice of $A = A_1(\varepsilon)$. However, for another choice of $A = A_2(\varepsilon)$, the matrix function U_2 will have a jump at $\varepsilon = 0$. Hence, this type of discontinuity is inherent for the statement of the SP with T being only upper triangular. The same is true if T is always lower triangular.

The next example illustrates these slightly complicated considerations.

Example 2. A possible DPS (U, I_n) of the SP for $A_0 = I_2$ is $U_1(0) = I_2$. Let $A_1(\varepsilon) = A_0 + \varepsilon E_{2,1}$, where $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ and $\varepsilon_0 > 0$ is fixed. Then, for $\varepsilon \neq 0$, we have $|U_1(\varepsilon)| = P_2$, $P_2 = E_{1,2} + E_{2,1}$, and the function $U = U_1(\varepsilon)$ has a jump at $\varepsilon = 0$ with $\|U_1(\varepsilon) - U_1(0)\| = \sqrt{2}$ for $\varepsilon \neq 0$. If another DPS is chosen with $U_2(0) = P_2$, then the matrix function U_2 satisfies $U_2(\varepsilon) = P_2$ for $\varepsilon \neq 0$ and is constant (and hence continuous) at $\varepsilon = 0$. If, however, we take $A_2(\varepsilon) = A_0 + \varepsilon E_{1,2}$ and then $U_2(\varepsilon) = I_2$ for $\varepsilon \neq 0$, and the function $U = U_2(\varepsilon)$ is discontinuous with the same jump at the point $\varepsilon = 0$.

Example 2 shows that the discontinuity of $U = U(A)$ is inherent to this statement of the SP in cases like $A = \lambda I_2$. This is due to the definition of T as an upper triangular matrix (or as a lower triangular matrix). If we allow the condensed form T to be any triangular unitary equivalent form of A , then this type of artificial discontinuity may disappear due

to the definitions. This is the idea of the quasi-Schur condensed forms introduced later on. The latter forms generalize triangular Schur forms.

Without additional assumptions, the matrix $T \in \mathbf{T}(n)$ is only a condensed form rather than a canonical form of A relative to the similarity action of the group $\mathbf{U}(n)$. The only (albeit most important) invariants for this action, which are revealed by the matrix T , are the eigenvalues $\lambda_k(A) = T(k, k)$ of the matrix A .

The definition of complete invariants and canonical forms for the similarity action of $\mathbf{U}(n)$ on $\mathbb{C}(n)$ is mathematically interesting (see [24]), but it is much more complicated and is not considered in full detail here. Further on, we consider, among others, only a partial formulation of Schur canonical forms for generic matrices A , see also [26–28]. From the point of view of applications, the condensed forms provide the same advantages as the canonical forms. Moreover, strict unitary canonical forms of non-generic matrices A are rarely, if ever, used in practice, since they involve complicated conditions and procedures (which are hard to be checked) and are more sensitive to perturbations in A compared with the condensed forms.

Let $U \in \mathcal{U}(A)$ and $|D| = I_n$. Then, $UD \in \mathcal{U}(A)$ as well. In particular, $cU \in \mathcal{U}(A)$ for $|c| = 1$ and $-U \in \mathcal{U}(A)$. This fact has an important implication. The diameter of the set $\mathcal{U}(A)$, i.e., the maximum of $\|U - V\|$ for $U, V \in \mathcal{U}(A)$, is equal to 2 and is achieved for $U = -V \in \mathcal{U}(A)$.

Given a matrix $A \in \mathbb{C}(n)$, neither the ConSF $T \in \mathbf{T}(n)$ of A nor the transformation matrix $U \in \mathbf{U}(n)$ are unique in general. In fact, the matrix T is unique if and only if $A = \lambda I_n$, $\lambda \in \mathbb{C}$, while U is always not unique. In this case $T = A$ and $U \in \mathbf{U}(n)$ is an arbitrary unitary matrix, or, equivalently, $\mathcal{T}(A) = \{A\}$ and $\mathcal{U}(A) = \mathbf{U}(n)$.

If A has at least two different eigenvalues, then we have a set $\mathcal{T}(A)$ of ConSF T with different ordering of the eigenvalues of A on the diagonal of T . The ConSF T also differ in their strictly upper triangular parts.

Suppose that $\text{spect}(A)$ consists of $m \leq n$ pair-wise disjoint elements $\lambda_1, \lambda_2, \dots, \lambda_m$ with multiplicities $n_1, n_2, \dots, n_m$, where $n_1 + n_2 + \dots + n_m = n$. Then, there are

$$N = N(n_1, n_2, \dots, n_m) = \frac{n!}{n_1! n_2! \dots n_m!}$$

different orderings of the elements $T(k, k)$ on the diagonal $\text{Diag}(T)$ of the ConSF T , or N diagonally different solutions of the SP for A .

Here, one of the ConSF of A is the block matrix $T = [T_{k,l}]$ with $T_{k,l} \in \mathbb{C}(n_k, n_l)$ and $T(k, k) \in \mathbf{T}(n_k)$, where $\text{Diag}(T_{k,k}) = \lambda_k I_{n_k}$. In the generic case, $m = n$, we have $N(1, 1, \dots, 1) = n!$ diagonally different ConSF, while in the most non-generic case $m = 1$, we have $N(n) = 1$ and all ConSF are diagonally equal.

3. Canonical Schur Forms for Generic Matrices

In this section, we summarize and reformulate some of the results concerning Schur canonical forms for the unitary similarity action of $\mathbf{U}(n)$ on the set $\mathbb{C}(n)$. The canonical Schur form $T \in \mathbf{T}(n)$ of the matrix $A \in \mathbb{C}(n)$ is a ConSF with additional conditions imposed on its elements; see [24] and the references therein. We consider only generic matrices A with pair-wise disjoint eigenvalues for which the solution (U, T) of the Schur problem is continuous as a function of the matrix A . At the same time, the Schur basis U for condensed forms (and hence for canonical forms as well) of a matrix A with multiple eigenvalues may be discontinuous as a function of A .

Definition 5. For a given matrix $A \in \mathbb{C}(n)$, the set

$$\text{Orb}(A) = \{U^H A U : U \in \mathbf{U}(n)\} \subset \mathbb{C}(n)$$

is called equivalence class, or orbit, of A relative to the similarity action of the unitary group $\mathbf{U}(n)$.

Obviously, $B \in \text{Orb}(A)$ implies $A \in \text{Orb}(B)$ and vice versa. Let $\mathfrak{A} \subset \mathbb{C}(n)$ and $\mathfrak{C} \subset \mathbb{T}(n)$ be certain sets.

Definition 6. The matrices $A, B \in \mathbb{C}(n)$ are said to be unitary equivalent (denoted as $A \sim B$) if $B \in \text{Orb}(A)$.

Definition 7. The function $\gamma : \mathfrak{A} \rightarrow \mathfrak{C}$ is said to be canonical mapping for the similarity action of the group $\mathbf{U}(n)$ on the set $\mathfrak{A}$ when the equality $\gamma(A) = \gamma(B)$ holds if and only if $A \sim B$.

Hence, the canonical mapping $\gamma : \mathfrak{A} \rightarrow \mathfrak{C}$ is a complete invariant [29] for the similarity action of the group $\mathbf{U}(n)$ on the set $\mathfrak{A}$, but the opposite, of course, is not true. Next, the canonical form of A is defined as the image $\gamma(A)$ of A under γ .

Definition 8. The image $\gamma(A) \in \mathbb{T}(n)$ of the matrix A under the canonical mapping γ is said to be unitary canonical form, or Schur canonical form, of A .

Definition 9. The subset $\mathfrak{A}$ of $\mathbb{C}(n)$ is said to be closed in the Zariski topology if it is the union of the zeros of a system of polynomials in $z \in \mathbb{C}(n)$. The subset $\mathfrak{A} \subset \mathbb{C}(n)$ is said to be open in the Zariski topology if its complement $\mathbb{C}(n) \setminus \mathfrak{A} \subset \mathbb{C}(n)$ is closed in this topology.

Definition 10. A property $\mathcal{P}$ of a matrix $A \in \mathbb{C}(n)$ is said to be generic if it is fulfilled on a subset $\mathfrak{A} \subset \mathbb{C}(n)$, which is open in the Zariski topology.

Informally, the matrix A is said to be generic relative to a given property if this property is generic.

Proposition 3. The following properties of a matrix $A \in \mathbb{C}(n)$ are generic.

1. The matrix A is totally different from any fixed matrix $A_0 \in \mathbb{C}(n)$, i.e., $A(k, l) \neq A_0(k, l)$ for $k, l \in \mathbb{Z}[1, n]$; in particular, $A(k, l) \neq 0$ for any given pair (k, l) .
2. The matrix A is not normal, i.e., $A^H A \neq A A^H$; in particular, the matrix A is not unitary.
3. The singular values of the matrix A are positive and pair-wise different; in particular, $\text{rank}(A) = n$.
4. The eigenvalues λ_k of the matrix A satisfy the inequalities $\text{Re}(\lambda_k) \neq \text{Re}(\lambda_l)$ and $\text{Im}(\lambda_k) \neq \text{Im}(\lambda_l)$ for $k \neq l$; in particular, $\lambda_k \neq \lambda_l$ for $k \neq l$ and the Jordan canonical form of A is diagonal.
5. Any ConSF T of the matrix A has nonzero and pair-wise different elements on and above its diagonal, i.e., $T(k, l) \neq 0$ and $T(i, j) \neq T(k, l)$ for $i \leq j, k \leq l$ and $(i, j) \neq (k, l)$.

4. Geometry of Schur Canonical Sets

Let $\omega = (\omega_1, \omega_2, \dots, \omega_n) : \mathbb{Z}[1, n] \rightarrow \mathbb{Z}[1, n]$ be a permutation of the integers $1, 2, \dots, n$ and recall that $\mathbb{Z}[1, n-1] = \{1, 2, \dots, n-1\}$ and $\mathbb{Z}[2, n] = \{2, 3, \dots, n\}$. Set $\mathbb{Z}_n = \mathbb{Z}[1, n-1] \times \mathbb{Z}[2, n]$.

Below, we describe a possible set of canonical forms for the similarity action of the group $\mathbf{U}(n)$ on the subset $\mathbb{C}(n)$ of matrices with simple eigenvalues. Let $\mathcal{K}(n) \subset \mathbb{Z}_n^{n-1}$ be the set of $(n-1)$ -tuples

$$\{(i_1, j_1), (i_2, j_2), \dots, (i_{n-1}, j_{n-1})\}$$

of integer pairs $p_k = (i_k, j_k), k \in \mathbb{Z}[1, n-1]$, where $i_k < j_k$. There are μ_n such $(n-1)$ -tuples; see Table 1. Later on, we shall define three important types of such sets.

Table 1. Number of generic canonical Schur forms.

n	2	3	4	5	6	7	8	9	10
μ_n	1	3	20	210	3003	54,264	1,184,040	30,260,340	886,163,135

Definition 11. The conjugate pair of the pair $p = (i, j) \in \mathbb{Z}_n$ is

$$p^\tau = (i, j)^\tau = (n + 1 - j, n + 1 - i) \in \mathbb{Z}_n$$

The pair p is self-conjugate if $p^\tau = p$.

Obviously, the pair (i, j) is self-conjugate if and only if $i + j = n + 1$.

Definition 12. The conjugate $(n - 1)$ -tuple of the $(n - 1)$ -tuple $\theta = (p_1, p_2, \dots, p_{n-1})$, where $p_k = (i_k, j_k)$, $k \in \mathbb{Z}[1, n - 1]$, is

$$\theta^\tau = (p_1^\tau, p_2^\tau, \dots, p_{n-1}^\tau).$$

The $(n - 1)$ -tuple θ is self-conjugate if $\theta^\tau = \theta$.

The conjugation for pairs p and $(n - 1)$ -tuples θ is an involution, i.e., $((i, j)^\tau)^\tau = (i, j)$ and $(\theta^\tau)^\tau = \theta$. It corresponds to reflection relative to the anti-diagonal $(1, n), (2, n - 1), \dots, (n, 1)$ of $n \times n$ arrays.

Definition 13. The set $\mathcal{K}(n)$ has the following important subsets.

1. The set $\mathcal{K}_1(n) \subset \mathcal{K}(n)$ is of type 1 if its elements are of the form

$$\{(i_1, 1), (i_2, 2), \dots, (i_{n-1}, n - 1)\}$$

2. The set $\mathcal{K}_2(n) \subset \mathcal{K}(n)$ is of type 2 if its elements are of the form

$$\{(1, j_1), (2, j_2), \dots, (n - 1, j_{n-1})\}$$

3. The set $\mathcal{K}_3(n) = \mathcal{K}(n) \setminus (\mathcal{K}_1(n) \cup \mathcal{K}_2(n)) \subset \mathcal{K}(n)$ is of type 3 if it is neither of type 1 nor of type 2.

Note that the elements of the set $\mathcal{K}_1(n)$ are conjugate to the elements of the set $\mathcal{K}_2(n)$.

Proposition 4. The intersection $\mathcal{K}_1(n) \cap \mathcal{K}_2(n)$ has a single element

$$\theta^\circledast = \{(1, 2), (2, 3), \dots, (n - 1, n)\}$$

which is a self-conjugate $(n - 1)$ -tuple.

Definition 14. The elements of the set $\mathcal{K}_1(n) \cup \mathcal{K}_2(n)$ are said to be proper. The elements of the set $\mathcal{K}_3(n)$ are said to be improper.

There are $(n - 1)!$ elements in each of the sets $\mathcal{K}_1(n)$ and $\mathcal{K}_2(n)$ and one joint element of $\mathcal{K}_1(n)$ and $\mathcal{K}_2(n)$. Thus, we have

$$\begin{aligned} \text{card}(\mathcal{K}_1(n) \cup \mathcal{K}_2(n)) &= (n - 1)! + (n - 1)! - 1 = 2(n - 1)! - 1 \\ \text{card}(\mathcal{K}_3(n)) &= \mu_n - 2(n - 1)! + 1 \end{aligned}$$

Example 3. For $n = 2$, there is $\mu_2 = 1$ pair of indexes $(1, 2)$ and it is proper. For $n = 3$, there are $\mu_3 = 3$ sets of pairs of indexes

$$\{(1, 2), (1, 3)\}, \{(1, 2), (2, 3)\}, \{(1, 3), (2, 3)\}$$

and they are all proper. For $n = 4$, there are $\mu_4 = 20$ triples of pairs of indexes of which 11 are proper, namely

$$\begin{aligned} &\{(1, 2), (2, 3), (3, 4)\}, \{(1, 2), (1, 3), (1, 4)\}, \{(1, 4), (2, 4), (3, 4)\} \\ &\{(1, 2), (1, 3), (2, 4)\}, \{(1, 3), (2, 4), (3, 4)\}, \{(1, 2), (1, 3), (3, 4)\} \\ &\{(1, 3), (2, 3), (3, 4)\}, \{(1, 2), (2, 3), (2, 4)\}, \{(1, 4), (2, 3), (3, 4)\} \\ &\{(1, 4), (2, 3), (3, 4)\}, \{(1, 2), (2, 3), (1, 4)\} \end{aligned}$$

and 9 are improper, namely

$$\begin{aligned} &\{(1, 3), (2, 3), (3, 4)\}, \{(1, 3), (1, 4), (2, 4)\}, \{(1, 2), (1, 4), (3, 4)\} \\ &\{(1, 2), (1, 3), (2, 4)\}, \{(1, 3), (2, 4), (3, 4)\}, \{(1, 2), (1, 3), (3, 4)\} \\ &\{(1, 2), (1, 3), (2, 3)\}, \{(1, 3), (1, 4), (3, 4)\}, \{(1, 2), (1, 4), (2, 4)\} \end{aligned}$$

Proposition 5. The minimal and maximal elements relative to the order relation $\prec$ on the set $\mathcal{K}_2(n)$ are

$$\theta_1 = \{(1, 2), (1, 3), \dots, (1, n)\}$$

and

$$\theta_{(n-1)!} = \theta^@ = \{(1, 2), (2, 3), \dots, (n-1, n)\},$$

respectively. The minimal and maximal elements of the set $\mathcal{K}_1(n)$ are $\theta^@$ and

$$\theta_{2(n-1)!-1} = \{(1, n), (2, n), \dots, (n-1, n)\}$$

respectively.

Now, we are in position to define possible sets of canonical Schur forms (CSF) $S \in \mathbf{T}(n)$ for generic matrices $A \in \mathbb{C}(n)$. There are $n!(2(n-1)! - 1)$ such sets. The multiplier $n!$ comes from the different orders of the (simple) eigenvalues of A on the diagonal of S . The multiplier $2(n-1)! - 1$ corresponds to different choices of proper $(n-1)$ -tuples

$$\{(i_1, j_1), (i_2, j_2), \dots, (i_{n-1}, j_{n-1})\}$$

such that the elements $S(i_k, j_k)$, $k \in \mathbb{Z}[1, n-1]$, of S are positive.

If the eigenvalues of S are ordered as $\lambda_1 \prec \lambda_2 \prec \dots \prec \lambda_n$, then there remain $2(n-1)! - 1$ sets of CSF. Note that any fixed order of the (simple) eigenvalues of A on the diagonal of S is preserved only by unitary similarity transformations with diagonal matrices U , i.e., $U(j, k) = d(j, k) \exp(i\varphi_k)$.

If, in particular, we choose a given $(n-1)$ -tuple, say

$$\{(1, 2), (2, 3), \dots, (n-1, n)\} \in \mathcal{K}_1(n) \cap \mathcal{K}_2(n)$$

then the set of CSF is uniquely fixed. In this case, the CSF have the form

$$\begin{bmatrix} \lambda_1 & \oplus & * & \cdots & * \\ 0 & \lambda_2 & \oplus & \cdots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \oplus \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

where $\oplus$ denotes a positive element. The two other CSFs are

$$\begin{bmatrix} \lambda_1 & \oplus & \oplus & \cdots & \oplus \\ 0 & \lambda_2 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & * \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \begin{bmatrix} \lambda_1 & * & * & \cdots & \oplus \\ 0 & \lambda_2 & * & \cdots & \oplus \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & \oplus \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Note that there is a similar problem with canonical Jordan forms of matrices $A \in \mathbb{C}$ relative to general similarity transformations. Usually, it is assumed that different orders of the Jordan blocks do not produce different Jordan forms. Formally, this means that the canonical Jordan form of A is not a single block-diagonal matrix $J \in \mathbb{C}(n)$ but a class of block-diagonal matrices, which are permutationally equivalent to J .

Definition 15. A set $\mathfrak{C} \subset \mathbf{T}(n)$ of CSF $T \in \mathbf{T}(n)$ for generic matrices $A \in \mathbb{C}(n)$ and a fixed $(n-1)$ -tuple

$$\{(i_1, j_1), (i_2, j_2), \dots, (i_{n-1}, j_{n-1})\} \in \mathcal{K}_1 \cup \mathcal{K}_2$$

is characterized as follows.

1. The n diagonal elements of the matrix T are ordered as

$$T(1, 1) \prec T(2, 2) \prec \cdots \prec T(n, n).$$

2. The $n-1$ elements $T(i_1, j_1), T(i_2, j_2), \dots, T(i_{n-1}, j_{n-1})$ of T over the diagonal are real and positive.

Of course, we may choose the elements $T(i_k, j_k)$ to be real and negative as well, or to have angles equal to a fixed value $\varphi_0 \in (-\pi, \pi]$, etc.

A matrix $A \in \mathbb{C}$ with eigenvalues $\lambda_1 \prec \lambda_2 \prec \cdots \prec \lambda_n$ may be transformed into CSF $C \in \mathbf{T}(n)$ by the next three steps.

1. The matrix A is transformed into any ConSF $T_1 = U_1^H A U_1 \in \mathbf{T}(n)$ by a matrix $U_1 \in \mathbf{U}(n)$. Numerically this is performed by the QR algorithm [3]. For this purpose, the code `schur` from MATLAB[®] may be used [22].
2. A ConSF $T_2 = U_2^H T_1 U_2 \in \mathbf{T}(n)$ is constructed so that $T_2(k, k) = \lambda_k, k \in \mathbb{Z}[1, n]$. This may be performed by complex plane rotations, which interchange the positions of two diagonal elements $T_1(i, i)$ and $T_1(j, j)$ of T_1 such that $i < j$ but $T(i, i) > T(j, j)$; see, e.g., ref. [3].
3. A diagonal matrix $U_3 \in \mathbf{U}(n)$ with elements $U_3(1, 1) = 1, U_3(k, k) = \exp(i\varphi_k), \varphi_k \in \mathbb{R}, k \in \mathbb{Z}[2, n]$ is chosen so that the matrix $T = U_3^H T_2 U_3$ has positive elements in positions (i_k, j_k) .

In connection with step 3, we note that unitary similarity transformations that introduce positive elements in certain positions of the transformed matrices are considered in [30].

We recall that to introduce CSF in the set $\mathbb{C}(n)$ relative to the similarity action of $\mathbf{U}(n)$, we use the lexicographical order $\prec$ on $\mathbb{C} \simeq \mathbb{R} \times \mathbb{R}$. For $z_k = x_k + iy_k \in \mathbb{C}$, where $x_k, y_k \in \mathbb{R}$, $k = 1, 2$, we write $z_1 \prec z_2$ if either $x_1 < x_2$, or $x_1 = x_2$ and $y_1 < y_2$.

There are

$$\mu_n = \binom{v_n}{n-1}, \quad v_n = n(n-1)/2,$$

sets of generic canonical forms $\mathfrak{C}_k \subset \mathbf{T}(n), k \in \mathbb{Z}[1, \mu_n]$, for $A \in \mathbb{C}(n)$. The values of μ_n for small values of n are given in Table 1.

There are μ_n different pairs of $p_k = (i_k, j_k)$, $1 \leq i < j \leq n$. They are ordered lexicographically according to the rule $(i_1, j_1) \prec (i_2, j_2)$ if either $i_1 < i_2$, or $i_1 = i_2$ and $j_1 < j_2$. We may order the pairs p_k as $p_1 \prec p_2 \prec \dots \prec p_{\mu_n}$, where

$$k = k_n(i, j) = j + n(i - 1) - \frac{i(i + 1)}{2}. \quad (2)$$

Thus, we have the chain of inequalities

$$(1, 2) \prec (1, 3) \prec \dots \prec (1, n) \prec (2, 3) \prec \dots \prec (2, n) \prec \dots \prec (n - 2, n - 1) \prec (n - 2, n) \prec (n - 1, n)$$

For any $p = (i, j)$, denote the conjugate pair $p^\tau = (n + 1 - j, n + 1 - i)$, symmetric to p relative to the anti-diagonal of elements in positions $(i, n + 1 - i)$, $i \in \mathbb{Z}[1, n]$.

It follows from (2) that for n fixed and any $k \in \mathbb{Z}[1, \nu_n]$ there exists a unique integer $i = i_n(k)$ such that $a_n(i) \leq k \leq b_n(i)$, where $a_n(i) = 1 + (i - 1)(2n - i)/2$ and $b_n(i) = i(2n - i - 1)/2$.

The integer $i_n(k)$ may be defined from

$$a_n(i_n(k)) = \max_i \{a_n(i) \leq k\}$$

or

$$b_n(i_n(k)) = \min_i \{b_n(i) \geq k\}.$$

Finally, set

$$j_n(k) = k - n(i_n(k) - 1) + \frac{i_n(k)(i_n(k) + 1)}{2}.$$

Thus, we have defined a bijection

$$(i, j) \mapsto k = k_n(i, j), \quad k \mapsto (i, j) = (i_n(k), j_n(k)),$$

between the ordered sets of integers $\mathbb{Z}[1, \nu_n]$ and integer pairs (i, j) , where $1 \leq i < j \leq n$.

Proposition 6. The triple of pairs of indexes $((i, k), (i, j), (l, j))$, where $i < k, i < j, l < j$ and $k < j, i < l$, is said to be improper. The triple of pairs of indexes $((k, j), (i, j), (i, l))$, where $k < j, i < j, i < l$ and $k < i, j < l$, is said to be improper.

Theorem 1. Each set $\theta \in \mathcal{K}_1(n) \cup \mathcal{K}_2(n)$ of proper integer $(n - 1)$ -tuples defines a class $\mathfrak{C}(\theta) \subset \mathbf{T}(n)$ of canonical forms for the unitary similarity action of the group $\mathbf{U}(n)$ on generic matrices $A \in \mathbb{C}(n)$. These forms are upper triangular matrices S with $S(k, k) \prec S(k + 1, k + 1)$ for $k \in \mathbb{Z}[1, n - 1]$ and $T(i, j) \in \mathbb{R}$, $T(i, j) > 0$ for $(i, j) \in \theta$.

If the matrix $A \in \mathbb{C}(n)$ with eigenvalues $\lambda_1 \prec \lambda_2 \prec \dots \prec \lambda_n$ is already transformed into ConSF, i.e., $A \in \mathbf{T}(n)$, it is then easily put into CSF as follows. First, a matrix $U \in \mathbf{U}(n)$ is chosen so as

$$\text{Diag}(S) = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), \quad S = U^H A U.$$

Then, a diagonal unitary matrix D with $D(1, 1) = 1$ is found so that $T = D^H S D \in \mathfrak{C}(\theta)$. Denoting $S(i, j) = |S(i, j)| \exp(i\alpha(i, j))$ and $D(i, i) = \exp(i\varphi(i))$, $\varphi(1) = 0$, where $\alpha(i, j), \varphi(i) \in (-\pi, \pi]$, the conditions $S(i, j) > 0$ give the system of $n - 1$ linear equations

$$\varphi(i) - \varphi(j) = \alpha(i, j), \quad (i, j) \in \theta, \quad (3)$$

for $\varphi(2), \varphi(3), \dots, \varphi(n)$. If it happens that $\varphi(k) \notin (-\pi, \pi]$ for some k then $\varphi(k)$ is replaced by

$$\tilde{\varphi}(k) = \varphi(k) \bmod (2\pi) \in (-\pi, \pi]. \quad (4)$$

Three special sets of Schur canonical forms for generic matrices $A \in \mathbb{C}(n)$ deserve attention. For these sets, the system (3) for $\varphi(i)$, $i \in \mathbb{Z}[2, n]$, is solved explicitly as follows.

1. The first set corresponds to pairs of indexes $(1, j)$, $j \in \mathbb{Z}[2, n]$, and here $\varphi(i) = -\alpha(1, i)$.
2. The second set corresponds to pairs of index pairs (i, n) , $i \in \mathbb{Z}[1, n - 1]$, and here $\varphi(i) = \alpha(i, n) - \alpha(1, n)$.
3. The third set corresponds to pairs of indexes $(i, i + 1)$, $i \in \mathbb{Z}[1, n - 1]$, and here $\varphi(i) = -\alpha(1, 2) - \alpha(2, 3) - \dots - \alpha(i - 1, i)$.

In all these cases, the fulfilment of the convention (4) is presupposed.

The restrictions assumed in this section, and in particular, the condition that the eigenvalues of A are simple, seem serious, but in fact, their violation can make the perturbation analysis of this statement of SP meaningless. If, for example, A has two or more equal eigenvalues, then the Schur basis $U(A + \delta A)$ of the perturbed SP may be discontinuous as a function of the perturbation δA in A ; see, e.g., refs. [6].

5. Real Schur Canonical Forms

The considerations above are valid for real or genuinely complex matrices with spectra that may in turn be real or genuinely complex. In particular, we have the following four possibilities.

1. The matrix A is real and has a real spectrum.
2. The matrix A is real and has a genuinely complex spectrum (i.e., there is at least one complex conjugate pair $\alpha \pm i\beta$ of eigenvalues, where $0 < \beta \in \mathbb{R}$).
3. The matrix A is genuinely complex and has a real spectrum.
4. The matrix A is genuinely complex and has a genuinely complex spectrum.

When $A \in \mathbb{R}(n)$ (cases 1 and 2), we may use orthogonal transformation matrices $U \in \mathbf{O}(n)$ instead of unitary ones to obtain the real Schur canonical form and the real Schur condensed forms of A . In case 1, the transformation matrix is taken as $U \in \mathbf{O}(n)$ and both the SCF and the ConSF of A are real upper triangular matrices T with the eigenvalues of A on their main diagonals.

Case 2 is slightly more subtle. Here, the transformation matrix U may be chosen as orthogonal [23,24], while the canonical form and the condensed forms of A are upper block-triangular matrices with 1×1 or 2×2 blocks ($\lambda_k \in \mathbb{R}$ or $\Lambda_k \in \mathbb{R}(2)$) on the main diagonal. In this case, there is at least one 2×2 block

$$\Lambda_k = \begin{bmatrix} \alpha_k & \beta_k \\ -\beta_k & \alpha_k \end{bmatrix} \in \mathbb{R}(2)$$

corresponding to the eigenvalues $\alpha_k \pm i\beta_k$ of A , where $\alpha_k, \beta_k \in \mathbb{R}$ and $\beta_k > 0$.

Let $n > 2$ and suppose that the spectrum $\text{spect}(A)$ contains m real elements $\lambda_1, \lambda_2, \dots, \lambda_m$ and $n - m$ genuinely complex elements $\alpha_k + i\beta_k, \alpha_k - i\beta_k$, $k = 1, 2, \dots, (n - m)/2$, where the number $n - m$ is even. Set $q = (n - m)/2$. Then, the orthogonal canonical form of A has the structure

$$S = U^T A U = \begin{bmatrix} S_{1,1} & S_{1,2} \\ O_{n-m,m} & S_{2,2} \end{bmatrix}$$

Here $S_{1,1} \in \mathbb{R}(m)$, $S_{1,2} \in \mathbb{R}(m, n - m)$, $S_{2,2} \in \mathbb{R}(n - m)$,

$$S_{1,1} = \begin{bmatrix} \lambda_1 & s_{1,2} & \cdots & s_{1,m} \\ 0 & \lambda_2 & \cdots & s_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}, \quad S_{2,2} = \begin{bmatrix} \Lambda_1 & S_{1,2} & \cdots & S_{1,q} \\ O_2 & \Lambda_2 & \cdots & S_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ O_2 & O_2 & \cdots & \Lambda_q \end{bmatrix}$$

and $s_{i,j} \in \mathbb{R}$, $S_{i,j} \in \mathbb{R}(2)$. The diagonal blocks are ordered as $\lambda_1 < \lambda_2 < \dots < \lambda_m$ and $\alpha_t + i\beta_t < \alpha_{t+1} + i\beta_{t+1}$, $t = 1, 2, \dots, q$.

6. Perturbations of the Schur Problem

Let $(U, T) \in \mathbf{U}(n) \times \mathbf{T}(n)$ be a particular DPS of SP for a matrix $A \in \mathbb{C}(n)$, i.e., $T = U^H A U$ and $\text{Diag}(T) = \text{Diag}(A)$. If A is already in ConSF, i.e., $A \in \mathbf{T}(n)$, there is nothing to transform. Here, we may choose $U = I_n$ for definiteness among all unitary matrices U that keep the diagonal of A under similarity transformations $(U, A) \rightarrow U^H A U$. In addition, if the matrix $A \in \mathbf{T}(n)$ has a simple spectrum, and/or has multiple eigenvalues but its upper triangular form is generically nonzero, then any unitary matrix U in the DPS is necessarily diagonal, i.e., $|U| = I_n$.

The choice of U as the simplest diagonal matrix I_n for $A \in \mathbf{T}(n)$ is justified by a number of additional arguments as follows.

1. It works for arbitrary matrices A with simple eigenvalues as well as with multiple eigenvalues.
2. When using ConSF for the evaluation of matrix functions, the computational algorithm has to work with a value for U , which is applicable to all input matrices A . For this reason, the QR algorithm [3] for computing T first checks whether $A \in \mathbf{T}(n)$. If yes, it assumes $U = I_n$ and $T = A$.
3. Algorithms for computing a Jordan form $J = V^{-1} A V$ of A act similarly. Indeed, for A already in Jordan form, computational algorithms for finding (V, J) assume $V = I_n$ and $J = A$.

As shown later on, the choice $U(0) = I_n$ in the solution of SP for $A_0 = \lambda I_n$ may lead to discontinuity of the function $U = U(\varepsilon)$ at the point $\varepsilon = 0$, in particular solutions $(U(\varepsilon), T(\varepsilon))$ of SP for a perturbed matrix $A + \varepsilon E$ with a given E . If we choose $U(0) \neq I_n$, this discontinuity disappears. However, for another perturbed matrix $A + \varepsilon F$, the discontinuity occurs as well. This inevitable discontinuity is due to the fact that ConSF T is upper triangular and for a lower triangular perturbation εE , the perturbed transformation matrix $U(\varepsilon)$ is a flip matrix such as P_n .

Let $\delta A \in \mathbb{C}(n)$ be a perturbation in A . Usually (but not always), we suppose that the matrix δA is small relative to A , e.g., $\|\delta A\| \leq \rho \|A\|$, where $\rho = 2^{-53} \simeq 1.1102 \times 10^{-16}$ is the rounding unit of FPA used in the computations [31].

We often assume that the perturbation δA is a 1-parameter family $\delta A(\varepsilon) = \varepsilon E$, where $\varepsilon > 0$ is a small parameter, and $E \in \mathbb{C}(n)$ is a fixed matrix with $\|E\| = 1$, i.e., $\|\delta A\| = \varepsilon$. The technique of the so-called fictitious small parameter can also be used in the perturbation analysis of matrix problems. Assuming that $\|\delta A\|$ is small relative to $\|A\|$, we use the identity $\delta A = \varepsilon E$, where $E = \delta A / \|\delta A\|$ and ε is finally set to $\|\delta A\|$.

The formulation of the perturbed Schur problem (PSP), i.e., the SP for a perturbed matrix $A + \delta A$, is not trivial. First, we mention two facts.

1. If PSP for a perturbed matrix $A + \delta A$ has a perturbed solution $(U + \delta U, T + \delta T)$ with $\|\delta U\|, \|\delta T\| \rightarrow 0$ as $\|\delta A\| \rightarrow 0$, then it also has a perturbed solution $(U + \Delta U, T + \delta T)$ with $\Delta U = -2U - \delta U$ and $\|\Delta U\| \rightarrow 2$ as $\delta \rightarrow 0$. The reason is that $U + \Delta U = -(U + \delta U)$.
2. The solution of PSP for a matrix $A + \delta A$ may have the form $(U, T + \delta T)$ with $\delta U = 0$. This will happen when $U^H \delta A U \in \mathbf{T}(n)$, and in this case, $\|\delta T\| = \|\delta A\|$.

Let $(U_0, T_0) \in \mathbf{U}(n) \times \mathbf{T}(n)$ be a particular solution of SP for a matrix $A_0 \in \mathbb{C}(n)$. For definiteness and based on the above considerations, we assume that, if $A_0 \in \mathbf{T}(n)$, then $U_0 = I_n$ and $T_0 = A_0$.

Consider the perturbation $\delta A = \varepsilon E$, where $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\varepsilon_0 > 0$ is a small parameter and $E \in \mathbb{C}(n)$ is a fixed matrix with $\|E\| = 1$. Let $(U(\varepsilon), T(\varepsilon))$ be a particular solution to PSP for the matrix $A(\varepsilon) = A_0 + \varepsilon E$, i.e.,

$$T(\varepsilon) = U^H(\varepsilon) A(\varepsilon) U(\varepsilon) \in \mathbf{T}(n), \quad U(\varepsilon) \in \mathbf{U}(n). \quad (5)$$

Since the solution of PSP always exists, we have defined functions $U: (-\varepsilon_0, \varepsilon_0) \rightarrow \mathbf{U}(n)$ and $T: (-\varepsilon_0, \varepsilon_0) \rightarrow \mathbf{T}(n)$ through relations (5). The problem is that there are many such

functions and not all of them are suitable for perturbation analysis. The aim of the next definition is to clarify the concepts in this area.

Definition 16. The pair $(U(\varepsilon), T(\varepsilon))$ is said to be a regular solution of the PSP for the matrix $A(\varepsilon) = A_0 + \varepsilon E$ if the functions U and T are continuous on the interval $(-\varepsilon_0, \varepsilon_0)$ and, if $A_0 \in \mathbf{T}(n)$, then $(U(0), T(0))$ is the DPS of SP for A_0 .

Example 4. Let $A = \lambda I_n$, $\lambda \in \mathbb{C}$. Then, the general solution of SP for A is $\mathbf{U}(n) \times \{A\}$. The opposite statement is also true in the form of the next two assertions.

1. If $\mathcal{U}(A) = \mathbf{U}(n)$, then $A = \lambda I_n$ and $\mathcal{T}(A) = \{A\}$.
2. If $\mathcal{T}(A) = \{A\}$, then $A = \lambda I_n$ and $\mathcal{U}(A) = \mathbf{U}(n)$.

Example 5. Let $A = \lambda I_n + J_n$, where $\lambda \in \mathbb{C}$ and $J_n = [0, I_{n-1}; 0, 0] \in \mathbb{R}(n)$ is a Jordan block with a zero eigenvalue. Then,

$$\mathcal{U}(A) = \{U \in \mathbf{U}(n) : |U| = I_n\}, \quad \mathcal{T}(A) = \{\lambda I_n + X : |X| = J_n\}.$$

A number of examples of ConSF for $A \in \mathbb{R}(2)$ presented in the next section illustrate the structure of these forms and the behavior of their perturbations; see also [7].

7. Examples of Real 2×2 Matrices

In this section, we consider several examples illustrating the concepts introduced so far. The examples are for SP and PSP for matrices $A \in \mathbb{R}(2)$ with $\text{spect}(A) \subset \mathbb{R}$ for which the transformation group is $\mathbf{O}(2)$. This is the simplest non-trivial case. However, the effects observed are valid for matrices $A \in \mathbb{C}(n)$ of the form

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ O_{n-2,2} & A_{2,2} \end{bmatrix}$$

where $n \geq 3$, $A_{1,1} \in \mathbb{R}(2)$, $A_{1,2} \in \mathbb{C}(2, n-2)$ and $A_{2,2} \in \mathbb{C}(n-2)$.

Matrices $A \in \mathbb{R}(2)$ correspond to linear operators $\mathbb{R}(2, 1) \rightarrow \mathbb{R}(2, 1)$ and have the simplest nontrivial albeit rich structure. A surprisingly large number of facts about general linear operators is revealed by such matrices; see, e.g., ref. [32] and the examples below.

Example 6. Let the matrix $A \in \mathbb{R}(2)$ have eigenvalues λ_1, λ_2 and set

$$r = \sqrt{\|A\|_{\mathbb{F}}^2 - |\lambda_1|^2 - |\lambda_2|^2}.$$

Then, the following four cases are possible in which the statements are reversible.

- If $\lambda_1 = \lambda_2 = \lambda$ and $r = 0$, then there exists a unique ConSF λI_2 of the matrix A .
- If $\lambda_1 = \lambda_2 = \lambda$ and $r > 0$, then there exist two ConSF $\lambda I_2 \pm r E_{1,2}$ of the matrix A .
- If $\lambda_1 \neq \lambda_2$ and $r = 0$, then there exist two ConSF $\text{diag}(\lambda_1, \lambda_2)$ and $\text{diag}(\lambda_2, \lambda_1)$ of the matrix A .
- If $\lambda_1 \neq \lambda_2$ and $r > 0$, then there exist four ConSF $\text{diag}(\lambda_1, \lambda_2) \pm r E_{1,2}$, $\text{diag}(\lambda_2, \lambda_1) \pm r E_{1,2}$ of the matrix A .

Example 7. Let $A_0 = \lambda I_2$, $0 \neq \lambda \in \mathbb{R}$. We have $\mathcal{U}(A_0) = \mathbf{O}(2)$ and $\mathcal{T}(A_0) = \{A_0\}$. Since A_0 is in ConSF, we consider the simplest particular DPS, which is (I_2, A_0) . Let the matrix A_0 be perturbed to $A(\varepsilon) = A_0 + \varepsilon E_{2,1}$, where ε is a small parameter. Then, a particular solution $(U(\varepsilon), T(\varepsilon))$ of PSP may be written as $U(\varepsilon) = I_2 + \delta U(\varepsilon)$, $T(\varepsilon) = A_0 + \delta T(\varepsilon)$, where $\delta T(\varepsilon) = \varepsilon U^H(\varepsilon) E_{2,1} U(\varepsilon)$. Here, the perturbation $\delta T(\varepsilon)$ is small for order ε , but the perturbation $\delta U(\varepsilon)$ is not small even though ε is small.

For $\varepsilon \neq 0$, the set of transformation matrices $\mathcal{U}(\varepsilon)$ consists of the matrices $U_1 = E_{1,2} + E_{2,1}$, $U_2 = E_{1,2} - E_{2,1}$ and their negations. In view of the equalities $U_k = I_2 + \delta U_k$, we have $\|\delta U_1\| = \|I_2 \pm U_1\| = 2$ and $\|\delta U_2\| = \|I_2 \pm U_2\| = \sqrt{2}$. At the same time, for $\varepsilon \neq 0$, the set of Schur

forms $\mathcal{T}(A(\varepsilon))$ consists of two matrices $A_0 \pm \varepsilon E_{1,2}$. Thus, the transformation matrix $U(\varepsilon)$ is discontinuous at the point $\varepsilon = 0$.

Consider the multivalued function $\Psi : \mathbb{R} \rightarrow 2^{\mathbf{O}(2)}$, where $2^{\mathbf{O}(2)}$ is the set of subsets of $\mathbf{O}(2)$, defined by $\varepsilon \mapsto \Psi(\varepsilon) := \mathcal{U}(\varepsilon)$. We have $\Psi(0) = \mathbf{O}(2)$ and $\Psi(\varepsilon) = \{\pm U_1, \pm U_2\}$ for $\varepsilon \neq 0$. Hence, the function Ψ , i.e., the Schur basis for $\mathbb{R}(2, 1)$ relative to the matrix $A(\varepsilon)$, is discontinuous at the point $\varepsilon = 0$, while the Schur forms $T = T(\varepsilon)$ of $A(\varepsilon)$ are continuous in ε .

Example 8. Let $A_0 = \lambda I_2 + E_{1,2} \in \mathbb{R}(2)$ be a Jordan block with eigenvalue $\lambda \in \mathbb{R}$. The set $\mathcal{T}(A_0)$ contains two matrices $T_{0,1} = \lambda I_2 + E_{1,2}$ and $T_{0,2} = \lambda I_2 - E_{1,2}$, while the set $\mathcal{U}(A_0)$ contains the matrices $I_2, E_{1,1} - E_{2,2}$ and their negations.

Let the matrix A_0 be perturbed to $A(\varepsilon) = A_0 + \varepsilon E_{2,1}$, where $\varepsilon > 0$. The eigenvalues of $A(\varepsilon)$ are $\lambda_1(\varepsilon) = \lambda - \sqrt{\varepsilon}$, $\lambda_2(\varepsilon) = \lambda + \sqrt{\varepsilon}$. Setting

$$c(\varepsilon) = \frac{1}{\sqrt{1+\varepsilon}}, \quad s(\varepsilon) = -\frac{\sqrt{\varepsilon}}{\sqrt{1+\varepsilon}}$$

we see that there are four ConSF

$$T_1(\varepsilon) = \begin{bmatrix} \lambda_1(\varepsilon) & 1-\varepsilon \\ 0 & \lambda_2(\varepsilon) \end{bmatrix}, \quad T_2(\varepsilon) = \begin{bmatrix} \lambda_1(\varepsilon) & \varepsilon-1 \\ 0 & \lambda_2(\varepsilon) \end{bmatrix},$$

and

$$T_3(\varepsilon) = \begin{bmatrix} \lambda_2(\varepsilon) & 1-\varepsilon \\ 0 & \lambda_1(\varepsilon) \end{bmatrix}, \quad T_4(\varepsilon) = \begin{bmatrix} \lambda_2(\varepsilon) & \varepsilon-1 \\ 0 & \lambda_1(\varepsilon) \end{bmatrix},$$

The orthogonal matrices $U_k(\varepsilon)$ that transform $A(\varepsilon)$ into $T_k(\varepsilon)$ are

$$U_1(\varepsilon) = \begin{bmatrix} c(\varepsilon) & -s(\varepsilon) \\ s(\varepsilon) & c(\varepsilon) \end{bmatrix}, \quad U_2(\varepsilon) = \begin{bmatrix} c(\varepsilon) & s(\varepsilon) \\ s(\varepsilon) & -c(\varepsilon) \end{bmatrix},$$

and

$$U_3(\varepsilon) = \begin{bmatrix} c(\varepsilon) & s(\varepsilon) \\ -s(\varepsilon) & c(\varepsilon) \end{bmatrix}, \quad U_4(\varepsilon) = \begin{bmatrix} c(\varepsilon) & -s(\varepsilon) \\ -s(\varepsilon) & -c(\varepsilon) \end{bmatrix}$$

Hence, there are two regular solutions of this PSP, namely $(U_1(\varepsilon), T_1(\varepsilon))$ and $(U_3(\varepsilon), T_3(\varepsilon))$ corresponding to the unperturbed ConSF $T_{0,1}$ and $T_{0,2}$, respectively.

Example 9. Let $A_0 = \text{diag}(\lambda_1, \lambda_2)$, where $\lambda_1 \neq \lambda_2$. Here, the set $\mathcal{T}(A_0)$ contains two diagonally different ConSF $T_{0,1} = \text{diag}(\lambda_1, \lambda_2)$ and $T_{0,2} = \text{diag}(\lambda_2, \lambda_1)$ of A_0 , while the set $\mathcal{U}(A_0)$ has eight elements, namely $\pm E_{1,1} \pm E_{2,2}$ and $\pm E_{1,2} \pm E_{2,1}$. Let us choose $\delta A(\varepsilon) = \varepsilon E_{2,1}$. For $\varepsilon \neq 0$, the set $\mathcal{T}(A_0 + \varepsilon E_{2,1})$ has four elements $\tilde{T}_1(\varepsilon) = T_{0,1} + \varepsilon E_{1,2}$, $\tilde{T}_2(\varepsilon) = T_{0,1} - \varepsilon E_{1,2}$ and $\tilde{T}_3(\varepsilon) = T_2 + \varepsilon E_{1,2}$ and $\tilde{T}_4(\varepsilon) = T_2 - \varepsilon E_{1,2}$.

The matrices $U_1, -U_1$ from Example 7 transform the perturbed matrix $A + \delta A(\varepsilon)$ into the ConSF $\tilde{T}_3(\varepsilon)$ and the matrices $U_2, -U_2$ transform $A + \delta A(\varepsilon)$ into the ConSF $\tilde{T}_4(\varepsilon)$ since U_1 and U_2 transform $\delta A(\varepsilon)$ in $\pm \varepsilon E_{1,2} \in \mathbf{T}(2)$, respectively.

Consider the transformation of $A + \delta A$ into some of the Schur forms $\tilde{T}_1(\varepsilon)$ or $\tilde{T}_2(\varepsilon)$. Define the orthogonal matrices

$$U_1(\varepsilon) = \begin{bmatrix} c(\varepsilon) & s(\varepsilon) \\ s(\varepsilon) & -c(\varepsilon) \end{bmatrix}, \quad U_2(\varepsilon) = \begin{bmatrix} c(\varepsilon) & -s(\varepsilon) \\ s(\varepsilon) & c(\varepsilon) \end{bmatrix},$$

where

$$c(\varepsilon) = \frac{\lambda_1 - \lambda_2}{\sqrt{\varepsilon^2 + (\lambda_1 - \lambda_2)^2}}, \quad s(\varepsilon) = \frac{\varepsilon}{\sqrt{\varepsilon^2 + (\lambda_1 - \lambda_2)^2}}.$$

We have

$$U_k^\top(\varepsilon)(A + \delta A(\varepsilon))U_k(\varepsilon) = \tilde{T}_k(\varepsilon), \quad k = 1, 2.$$

Furthermore, $U_2(0) = I_2$ and $U_1(0) = \text{diag}(1, -1)$ are fulfilled. Hence, the regular solution of the PSP is $(U_2(\varepsilon), \tilde{T}_2(\varepsilon))$.

Example 10. Let $A_0 = \text{diag}(\lambda_1, \lambda_2) + aE_{1,2} \in \mathbb{R}(2)$, where $\delta = \lambda_2 - \lambda_1 > 0$ and $a \neq 0$. The set $\mathcal{T}(A_0)$ of ConSF of A_0 contains four matrices:

$$T_{1,2} = \text{diag}(\lambda_1, \lambda_2) \pm aE_{1,2}, \quad T_{3,4} = \text{diag}(\lambda_2, \lambda_1) \pm aE_{1,2}.$$

Let the matrix A_0 be perturbed to $A(\varepsilon) = A_0 + \varepsilon E_{2,1}$, where ε is a small parameter such that $\delta^2 + 4a\varepsilon > 0$, i.e., $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, where $\varepsilon_0 = \delta^2 / (4|a|)$. The ConSF of the matrix $A(\varepsilon)$ are

$$\tilde{T}_{1,2} = \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2) \pm \tilde{a}E_{1,2}, \quad \tilde{T}_{3,4} = \text{diag}(\tilde{\lambda}_2, \tilde{\lambda}_1) \pm \tilde{a}E_{1,2},$$

where the quantities $\tilde{\lambda}_1 = \tilde{\lambda}_1(\varepsilon)$, $\tilde{\lambda}_2 = \tilde{\lambda}_2(\varepsilon)$ and $\tilde{a} = \tilde{a}(\varepsilon)$ are analytical functions of ε . In particular,

$$\begin{aligned} \tilde{\lambda}_1 &= \lambda_1 + \frac{a\varepsilon}{\lambda_1 - \lambda_2} + O(\varepsilon^2), \\ \tilde{\lambda}_2 &= \lambda_2 + \frac{a\varepsilon}{\lambda_2 - \lambda_1} + O(\varepsilon^2), \\ \tilde{a} &= a - \varepsilon + O(\varepsilon^2), \quad \varepsilon \rightarrow 0. \end{aligned}$$

Among the four ConSF, only the matrix $\tilde{T}_1 = \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2) + \tilde{a}E_{1,2}$ corresponds to a regular solution.

8. Diagonally Spectral Matrices

Denote by $\Delta(n) \subset \mathbb{C}(n)$ the set of matrices $A \in \mathbb{C}(n)$ such that the multiset of its diagonal elements is equal to the multiset of its eigenvalues, i.e.,

$$\text{spect}(A) = \{A(1,1), A(2,2), \dots, A(n,n)\} \quad (6)$$

Otherwise speaking, $\Delta(n)$ is the set of matrices A such that

$$\det(A - A(k,k)I_n) = 0, \quad k = 1, 2, \dots, n \quad (7)$$

Definition 17. The matrix $A \in \Delta(n)$, which satisfies (6) or (7), is said to be diagonally spectral.

The set $\Delta(n) \subset \mathbb{C}(n)$ is defined by n algebraic equations (7) (some of them may not be independent) in the n^2 elements of the matrix A and is hence a closed algebraic variety [29] of complex dimension of at least $n^2 - n$.

Upper triangular matrices and lower triangular matrices are diagonally spectral. Schur condensed forms in particular are diagonally spectral. More generally, for $P \in \mathbf{O}(n)$ being a permutation matrix, and $A \in \mathbb{C}(n)$ being a diagonally spectral matrix, the matrix PAP is also diagonally spectral.

Example 11. The elements of the matrix $A \in \Delta(2)$ satisfy one independent algebraic equation $A(1,2)A(2,1) = 0$. Hence, the matrix $A \in \Delta(2)$ has the form

$$\begin{bmatrix} * & * \\ 0 & * \end{bmatrix}, \quad \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}$$

where $*$ denotes unspecified matrix elements.

Example 12. The matrices $A_1, A_1^\top, A_2, A_2^\top, A_3, A_3^\top \in \mathbb{C}(3)$, where

$$A_1 = \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix}, \quad A_2 = \begin{bmatrix} * & 0 & * \\ * & * & * \\ 0 & 0 & * \end{bmatrix}, \quad A_3 = \begin{bmatrix} * & * & * \\ 0 & * & 0 \\ 0 & * & * \end{bmatrix},$$

are diagonally spectral.

Matrices from $\Delta(n)$ may not be condensed in the sense that they may have less than $n(n-1)/2$ zero elements. In particular, matrices from $\Delta(n)$ may have all their elements different from zero.

Example 13. Let $z \in \mathbb{C}$ be a parameter. Then, the matrices

$$A(z) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ z & -z & 3 \end{bmatrix}, B(z) = \begin{bmatrix} 1 & 1 & 1 \\ z & 2 & 1 \\ -z-2 & 2 & 3 \end{bmatrix}$$

are diagonally spectral, i.e., $\text{spect}(A(z)) = \text{spect}(B(z)) = \{1, 2, 3\}$. We stress that $A(0) \in \mathbf{T}(3)$ but $B(z) \notin \mathbf{T}(3)$ for all $z \in \mathbb{C}$.

The main advantage of a Schur canonical or condensed form

$$T = U^T A U \in \mathbf{T}(n), U \in \mathbf{U}(n)$$

of a matrix $A \in \mathbb{C}(n)$ is that it reveals the spectrum $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ of the matrix A as the collection of the diagonal elements

$$\{T(1, 1), T(2, 2), \dots, T(n, n)\}$$

of the form T . Thus, the sets of Schur canonical and condensed forms are subsets of the larger set (closed in the Zarisky topology)

$$\Delta(n) = \Delta_1(n) \cap \Delta_2(n) \cap \dots \cap \Delta_n(n) \subset \mathbb{C}(n),$$

where

$$\Delta_k(n) = \{A \in \mathbb{C}(n) : \det(A - A(k, k)I_n) = 0\}, k \in \mathbb{Z}[1, n],$$

of matrices having spectra equal to the collection of their diagonal elements.

Obviously, the set $\mathbf{T}(n)$ as well as the set of lower triangular $n \times n$ matrices are subsets of $\Delta(n)$. More generally, if $P \in \mathbf{O}(n)$ is a permutation matrix (i.e., the columns of P are a permutation of the columns of the identity matrix I_n) and $T \in \mathbf{T}(n)$ then $PTP \in \Delta(n)$. In particular, if $T \in \mathbf{T}(n)$, then the matrix $R = P_n T P_n$ is lower triangular with $R(k, k) = T(n+1-k, n+1-k), k = 1, 2, \dots, n$.

Example 14. Let $A = \lambda I_2, 0 \neq \lambda \in \mathbb{R}$. One of the DPS (U, T) of SP for A is the pair (I_n, A) . If we perturb A to $\tilde{A} = \lambda I_2 + \varepsilon E_{2,1}$, where $\varepsilon > 0$ is small, the pair (U, T) is transformed to $(\tilde{U}, \tilde{T})$, where $\tilde{T} = A + \varepsilon E_{1,2}$ and $\tilde{U}$ is any of the four matrices $\pm E_{1,2} \pm E_{2,1}$. Thus, $\|U - \tilde{U}\| = 2$ and the transformation matrix $U = U(\varepsilon)$ is discontinuous at the point $\varepsilon = 0$.

To avoid such artificial high sensitivity, in the next section, we introduce the concept of condensed quasi-Schur forms.

9. Condensed Quasi-Schur Forms

Let $m \in \mathbb{Z}[1, n]$ be a fixed integer.

Definition 18. The m -tuple $v = (n_1, n_2, \dots, n_m)$, where n_k are positive integers, is said to be m -partition of n if $n = n_1 + n_2 + \dots + n_m$.

Next, we define condensed quasi-Schur forms of the matrix $A \in \mathbb{C}(n)$ as block-triangular matrices (upper or lower) such that the blocks on the diagonal of A are, in turn, triangular matrices. In particular, condensed quasi-Schur forms of A are diagonally

spectral and thus reveal the spectrum of A . The idea is that these forms are less sensitive to perturbations in A in comparison to other unitary equivalent forms of A such as ConSF.

Definition 19. A matrix $S = U^H A U \in \mathbb{C}(n)$, where $U \in \mathbf{U}(n)$, is said to be a condensed quasi-Schur form of $A \in \mathbb{C}(n)$ if there exists an m -partition v of n such that S or $S^\top$ is block-upper triangular with diagonal blocks $S_{k,k} \in \mathbb{C}(n_k)$, i.e.,

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,m} \\ O & S_{2,2} & \cdots & S_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & S_{m,m} \end{bmatrix}$$

and either $S_{k,k} \in \mathbf{T}(n_k)$ or $S_{k,k}^\top \in \mathbf{T}(n_k)$, $k = 1, 2, \dots, m$.

Note that the upper triangular matrix $S_{k,k}$ is permutationally equivalent to the lower triangular matrix $S_{k,k}^\top = P_{n_k} S_{k,k} P_{n_k}$. At the same time, a matrix A and its transpose $A^\top = V^{-1} A V$ are similar [33], but we cannot use this result here since the matrix V is not unitary in general. Instead of the transposed matrix $A^\top$, we may use its flipped variant $A^F = P_n A P_n$, which is lower triangular whenever A is upper triangular and P_n is a permutation matrix.

Example 15. For $n = 2$, the condensed quasi-Schur forms are $A_1, A_1^\top$, where $A_1 \in \mathbf{T}(2)$. For $n = 3$, the condensed quasi-Schur are $A_1, A_1^\top, A_2, A_2^\top$ and $A_3, A_3^\top$, where $A_1 \in \mathbf{T}(3)$,

$$A_2 = \begin{bmatrix} * & * & * \\ 0 & * & 0 \\ 0 & * & * \end{bmatrix}, \quad A_3 = \begin{bmatrix} * & 0 & * \\ * & * & * \\ 0 & 0 & * \end{bmatrix}.$$

and the star denotes unspecified elements.

Condensed quasi-Schur forms are diagonally spectral, but the opposite is not true for $n \geq 3$; see Example 13. Obviously, a ConSF is also a condensed quasi-Schur form, but the opposite may not be true (we recall that $n \geq 2$). We stress that the high sensitivity of Schur forms as in Example 14 may not be observed for condensed quasi-Schur forms.

We stress, finally, that the concept of a diagonally spectral matrix as an algebraic object may be of independent interest in matrix analysis and should be independently studied in more detail.

10. Properties of Canonical/Condensed Forms

Canonical and condensed forms of matrices and matrix pencils under the action of matrix transformation groups are widely used in matrix analysis and control theory. These forms have zeros at given positions and, optionally, ones at other positions. Introducing zeros at given positions may lead to discontinuity of the transformation matrix, while introducing ones may cause bad conditioning of this matrix. In particular, the requirement that the condensed form T of A is upper triangular, i.e., $T \in \mathbf{T}(n)$, may lead to extreme sensitivity of the transformation pair (T, U) relative to perturbations in the matrix A . At the same time, this high sensitivity may not be relevant to the problem of computing the spectrum of A . Also, whether a condensed form is upper or lower triangular does not matter from both theoretical and practical points of view. In the concept of a condensed quasi-Schur form of a matrix, we exploit this idea: the condensed form is defined as a set, containing all triangular forms.

This approach may be extended to generalized Jordan forms (GJF) of matrices. We recall that GJF $G = X^{-1} A X$ of the matrix A is a bi-diagonal matrix with the eigenvalues of A on its diagonal and the property that $G(i, i+1) \neq 0$ implies $G(i, i) = G(i+1, i+1)$. The

difference with the standard Jordan form of A is that the nonzero super-diagonal elements $G(i, i + 1)$ of G are not necessarily equal to 1.

11. Conclusions

Novelty

In this paper, we consider condensed Schur forms for a general square matrix A as well as various sets of canonical Schur forms for a square matrix A with distinct eigenvalues. This case is generic, and hence, it is a primary problem [34] in the analysis of Schur decompositions. We also study the most non-generic case of scalar matrices, which belong to a one-dimensional variety in the space of square matrices.

The sensitivity of the Schur forms relative to perturbations in A is also studied. The concepts of a diagonal preserving solution to the Schur problem and of a regular solution to the perturbed Schur problem are introduced and illustrated by many examples.

We also introduce the concepts of diagonally spectral matrices and of quasi-Schur condensed forms of a matrix A . The latter forms are much less sensitive (if ever) to perturbations in the matrix A in comparison with the upper triangular condensed forms of A . The concept of a diagonally spectral matrix is about a new algebraic object that may be studied independently in matrix theory as a closed variety in the Zariski topology.

Author Contributions: Methodology, M.M.K.; Validation, P.H.P.; Formal analysis, M.M.K.; Writing—review & editing, P.H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors of this paper are grateful to the anonymous reviewers for their very useful and detailed comments and suggestions, which helped to improve the text.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Schur, I. Beiträge zur Theorie der Gruppen linearer homogener Substitutionen. *Trans. Am. Math. Soc.* **1909** *10*, 159–175. [\[CrossRef\]](#)
2. Bhatia, R. *Matrix Analysis*; Springer: Berlin/Heidelberg, Germany, 1996; ISBN 978-0387948461.
3. Golub, G.; Loan, C.V. *Matrix Computations*, 4th ed.; The Johns Hopkins University Press: Baltimore, MD, USA, 2013; ISBN 978-1421407944.
4. Horn, R.; Johnson, C. *Matrix Analysis*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2012; ISBN 978-0521839402.
5. Horn, R.; Johnson, C. *Topics in Matrix Analysis*; Cambridge University Press: Cambridge, UK, 1991; ISBN 978-0511840371.
6. Konstantinov, M.; Petkov, P.; Christov, N. Nonlocal perturbation analysis of the Schur system of a matrix. *SIAM J. Matrix Anal. Appl.* **1994**, *15*, 383–392. [\[CrossRef\]](#)
7. Konstantinov, M.; Petkov, P. *Perturbation Methods in Matrix Analysis and Control*; Nova Science Publishers: New York, NY, USA, 2020; ISBN 978-1536174700. [\[CrossRef\]](#)
8. Minenkova, A.; Nitch-Griffin, E.; Olshevsky, V. Gohberg-Kaashoek numbers and forward stability of Schur canonical forms. *arXiv* **2021**, arXiv:2110.15334. [\[CrossRef\]](#)
9. Minenkova, A.; Nitch-Griffin, E.; Olshevsky, V. Backward stability of the Schur decomposition under small perturbations. *Linear Algebra Appl.* **2024**, *in press*. [\[CrossRef\]](#)
10. Zhang, G.; Li, H.; Wel, Y. Componentwise perturbation analysis for the generalized Schur decomposition. *Calcolo* **2022**, *59*, 19. [\[CrossRef\]](#)
11. Konstantinov, M.; Mehrmann, V.; Petkov, P. Perturbation analysis of Hamiltonian Schur and block-Schur forms. *SIAM J. Matrix Anal. Appl.* **2001**, *23*, 387–424. [\[CrossRef\]](#)
12. Chu, E. Pole assignment via the Schur form. *Syst. Control. Lett.* **2007**, *56*, 303–314. [\[CrossRef\]](#)
13. Chu, D.; Liu, X.; Mehrmann, V. A numerical method for computing the Hamiltonian Schur form. *Numer. Math.* **2007**, *105*, 375–412. [\[CrossRef\]](#)
14. Chen, J.; Ma, W.; Miao, Y.; Wei, Y. Perturbations of Tensor-Schur decomposition and its applications to multivariable control systems and facial recognitions. *Neurocomputing* **2023**, *547*, 126446. [\[CrossRef\]](#)

15. Stewart, G.; Sun, J. *Matrix Perturbation Theory*; Academic Press: Cambridge, MA, USA, 1990; ISBN 978-0126702309.
16. Konstantinov, M.; Gu, D.; Mehrmann, V.; Petkov, P. *Perturbation Theory for Matrix Equations*; Science Direct: Amsterdam, The Netherlands, 2003; ISBN 0-444513159.
17. Konstantinov, M.; Petkov, P.; Christov, N. Invariants and canonical forms for linear multivariable systems under the action of orthogonal transformation groups. *Kybernetika* **1981**, *17*, 413–424.
18. Konstantinov, M.; Postlethwhite, I.; Gu, D.; Petkov, P. Perturbation analysis of orthogonal canonical forms. *Linear Algebra Appl.* **1997**, *251*, 267–291. [CrossRef](#)
19. Boley, D.; Datta, B. Numerical Methods for Linear Control Systems. In *Systems and Control in the Twenty-First Century. Systems & Control: Foundations & Applications*; Byrnes, C., Ed.; Birkhäuser: Boston, MA, USA, 1997; Volume 22.
20. Sun, J. Perturbation bounds for the generalized Schur decomposition. *SIAM J. Matrix Anal. Appl.* **1995**, *16*, 1328–1340. [CrossRef](#)
21. Sun, J. Perturbation analysis of system Hessenberg and Hessenberg/triangular forms. *Linear Algebra Appl.* **1996**, *241/243*, 811–849. [CrossRef](#)
22. The MathWorks, Inc. *MATLAB Version 9.9.0.1538559 (R2020b)*; The MathWorks, Inc.: Natick, MA, USA, 2020. Available online: www.mathworks.com (accessed on 27 November 2024).
23. Murnaghan, F.; Wintner, A. A canonical form for real matrices under orthogonal transformations. *Proc. Natl. Acad. Sci. USA* **1931**, *17*, 417–420. [CrossRef](#) [PubMed](#)
24. Shapiro, H. A survey of canonical forms and invariants for unitary similarity. *Linear Algebra Appl.* **1991**, *147*, 101–167. [CrossRef](#)
25. Higham, N. *Functions of Matrices: Theory and Computation*; SIAM: Philadelphia, PA, USA, 2008. [CrossRef](#)
26. Brenner, J. The problem of unitary equivalence, *Acta Math.* **1951**, *86*, 297–308. [CrossRef](#)
27. Littlewood, D. On unitary equivalence. *J. Lond. Math. Soc.* **1953**, *28*, 314–322. [CrossRef](#)
28. Ikramov, K. The canonical Schur form of a matrix with simple eigenvalues. *Dokl. Math.* **2008**, *77*, 359–360. [CrossRef](#)
29. Hartshorne, R. *Algebraic Geometry*; Springer: Berlin/Heidelberg, Germany, 1977; ISBN 978-0387902449.
30. Ikramov, K. On a constructive procedure for verifying whether a matrix can be made real by a unitary similarity transformation. *Comput. Math. Math. Phys.*, **2010**, *50*, 383–386. [CrossRef](#)
31. 754-2019; IEEE Standard for Floating-Point Arithmetic. IEEE Computer Society: Washington, DC, USA, 2019.
32. Glazman, M.; Ljubich, J. *Finite-Dimensional Linear Analysis: A Systematic Presentation in Problem Form (Dover Books in Mathematics)*; Dover Publications: New York, NY, USA, 2006; ISBN 978-0486453323.
33. Tausky, O.; Zassenhaus, H. On the similarity transformation between a matrix and its transpose. *Pac. J. Math.* **1959**, *9*, 893–896. [CrossRef](#)
34. Arnold, V. *Geometric Methods in the Theory of Ordinary Differential Equations*; Springer Science + Business Media: New York, NY, USA, 1988; ISBN 978-1461269946.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.