Article

Not peer-reviewed version

On Schur Forms for Matrices with Simple Eigenvalues

Mihail Mihailov Konstantinov and Petko Hristov Petkov *

Posted Date: 10 October 2024

doi: 10.20944/preprints202410.0785.v1

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

Article

On Schur Forms for Matrices with Simple Eigenvalues

Mihail Konstantinov [1] and **Petko Petkov** [2,*]

[1] University of Architecture, Civil Engineering and Geodesy, Sofia 1064, Bulgaria
[2] Bulgarian Academy of Sciences, Sofia 1040, Bulgaria
* Correspondence: php@tu-sofia.bg

Abstract: In this paper we consider the standard Schur problem for a square matrix A, namely the similarity unitary transformation of A into upper Schur form containing the eigenvalues of A on its diagonal. Since the profound work of Issai Schur (1909), this is 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 A. In particular, the concept of regular solution to the perturbed Schur form is introduced and illustrated by several examples. We also introduce the concepts of diagonally spectral matrices and of quasi-Schur condensed forms of a matrix A, and show that they may be much less sensitive to perturbations in A.

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

MSC: 15A21; 65G30

---

1. Introduction and Notation

The Schur decomposition of general square matrix and its generalizations are major tools both in the theory and applications of matrix analysis [5]. In this paper we consider the main definitions and properties of the Schur decomposition of a square matrix which are important from the point of view of the perturbation analysis. We also introduce new concepts in this field. A number of examples is given for illustration of the results presented. This is a specific issue and we shall need a large number of notations. For 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 [10].

Let $\mathbb{Z} = \{0, \pm1, \pm2, \dots\}$ be the set of integers and $m, n \in \mathbb{Z}$, where $m \leq n$. We denote by

$$\mathbb{Z}[m,n] = \{k \in \mathbb{Z} : m \leq k \leq n\}$$

(or by $m:n$) the set of $n - m + 1$ integers $m, m+1, \dots, n$. We write $Z[m,m] = m$ and $\mathbb{Z}[n,m] = \varnothing$ 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 $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. A complex number z is written as $z = x + \mathrm{i}y$ with $x, y \in \mathbb{R}$, or $z = |z|\exp(\mathrm{i}\varphi)$, 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 $\bar{z} = x - \mathrm{i}y = |z|\exp(-\mathrm{i}\varphi)$.

The sign function sign: $\mathbb{R} \to \mathbb{Z}[-1,1]$ for scalar arguments is defined as $\mathrm{sign}(x) = -1$, $\mathrm{sign}(x) = 0$ and $\mathrm{sign}(x) = 1$ 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

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

The lexicographical order $\prec$ for n-tuples is defined as $x \preceq \tilde{x}$, $x = \tilde{x}$ and $x \succ \tilde{x}$ if $\mathrm{sign}(x - \tilde{x}) < 0$, $\mathrm{sign}(x - \tilde{x}) = 0$ and $\mathrm{sign}(x - \tilde{x}) > 0$, resp. Otherwise speaking, $x \preceq \tilde{x}$ 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 write $x \preceq \tilde{x}$ if either $x \preceq \tilde{x}$ or $x = \tilde{x}$. We use this lexicographical order for complex numbers $z = x + \mathrm{i}y$ written as real pairs (x, y)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

and for pairs (m,n) of integers $m,n \in \mathbb{Z}$. For example, for the fourth roots $\{\pm 1, \pm \mathrm{i}\}$ of 1, we have $-1 \preceq -\mathrm{i} \preceq \mathrm{i} \preceq 1$.

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 column m-vector $b \in \mathbb{C}(m,1)$ with elements $b_i = b(i)$ is written as $b = [b_1; b_2; \ldots; b_m]$, while the row n-vector $c \in \mathbb{C}(1,n)$ with elements $c_j = c(j)$ is denoted as $c = [c_1, c_2, \ldots, c_n]$. A quantity $z \in \mathbb{C}\backslash\mathbb{R}$ is said to be genuinely complex. A vector or a matrix is genuinely complex if at least one of its elements is complex.

The identity $n \times n$ matrix is denoted as $I_n \in \mathbb{R}(n)$. The element $I_n(i,j)$ of I_n is the Kronecker delta symbol $d(i,j) = 1 - \mathrm{sign}^2(i-j)$. 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 $i \le j$. 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 by $A^{\mathrm{H}} \in \mathbb{C}(n,m)$ and has elements $A^{\mathrm{H}}(i,j) = \overline{A(j,i)}$. 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. 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 $\|A\|_{\mathrm{F}}$, resp.

The spectrum $\mathrm{spect}(A)$ 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 $\mathrm{spect}(A) \subset \mathbb{C}$ in the general case, and $\mathrm{spect}(A) \subset \mathbb{R}$ in the case when all eigenvalues of A are real.

The multiplicative group of unitary matrices $U \in \mathbb{C}(n)$ such that $U^{\mathrm{H}}U = I_n$ 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 $\mathrm{Low}(A) = A \circ L_n$ and $\mathrm{Diag}(A) = A \circ I_n$ 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 $\mathrm{diag}(x) \in \mathbb{C}(n)$ is the matrix with elements $\mathrm{diag}(x)(i,j) = x(i)d(i,j)$.

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

$$\mathrm{diag}(1, \exp(\mathrm{i}\varphi_2), \ldots, \exp(\mathrm{i}\varphi_n)),$$

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

If $\mathcal{K}$ is a finite set then $\mathrm{card}(\mathcal{K})$ is the number of its elements. The set of $(n-1)$-tuples of pairs $\{(i_1, j_1), (i_2, j_2), \ldots, (i_{n-1}, j_{n-1})\}$ 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 $\mathcal{K}(n)$.

Finally we set $\nu_n = n(n-1)/2$ and

$$\mu_n = \mathrm{card}(\mathcal{K}(n)) = (\nu_n)!/(n-1)!$$

In particular $\nu_4 = 6$ and $\mu_4 = 6!/3! = 20$. Unspecified matrix block are denoted by star. The end of definitions, examples and propositions is marked by $\square$.

2. Condensed Schur Forms

Let an arbitrary matrix $A \in \mathbb{C}(n)$, $n \ge 2$, be given. Then according to the famous Schur theorem [13] there exists a factorization $A = UTU^{\mathrm{H}}$ of the matrix A, where $U \in \mathbf{U}(n)$ and $T \in \mathbf{T}(n)$.

Definition 1. The pair

$$(U,T) \in \mathbf{U}(n) \times \mathbf{T}(n),\ T = U^{\mathrm{H}}AU, \tag{1}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

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

Thus defined the condensed Schur form T of A is not unique. Hence the condensed Schur forms are not canonical but rather *quasi-canonical*. If $A \in \mathbb{R}(n)$ and if the spectrum of A is real then the transformation matrix may be chosen as $U \in \mathbf{O}(n)$ and we have $T = U^\top AU \in \mathbb{R}(n)$. If $A \in \mathbb{R}(n)$ and the spectrum of A is genuinely complex then a real block Schur form with 1×1 and 2×2 diagonal blocks may also be constructed.

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

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

and

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

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

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

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 $U_1U_2 \in \mathcal{U}(A)$.

The most important (actually, the only important) property of the ConSF T of the matrix A is that its diagonal elements are the eigenvalues of A, i.e. $T(k,k) = \lambda_k(A)$, $k \in \mathbb{Z}[1,n]$. Because of the only condition $\mathrm{Low}(T) = 0$ the matrix T is only a condensed form (rather than a canonical form) of A relative to the similarity action $\mathbf{U}(n) \times \mathbb{C}(n) \to \mathbf{T}(n)$, defined by $(U,A) \mapsto U^{\mathrm{H}}AU$, 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

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

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

Sometimes the matrices U and T in a ConSF of A are written as $U(A)$ and $T(A)$ to emphasize their dependence on A. This dependence, however, is not functional. Indeed, the transformation matrix $U(A)$ is always not unique, e.g. $(U,T) \in \mathrm{Schur}(A)$ implies $(-U,T) \in \mathrm{Schur}(A)$. With exception of the case $A = \lambda I_n$, $\lambda \in \mathbb{C}$, when it is fulfilled $T(A) = A$, the upper triangular unitary equivalent form $T(A)$ of A is also not unique. For the latter choice of A we have $\mathrm{Schur}(A) = \mathbf{U}(n) \times \{A\}$.

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. [14].

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 $T_2 = U_2^{\mathrm{H}}U_1T_1U_1^{\mathrm{H}}U_2$. □

Proof. It suffices to observe that $A = U_1T_1U_1^{\mathrm{H}} = U_2T_2U_2^{\mathrm{H}}$. □

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

The next proposition is generally known since 1933 and is attributed to H. Röseler, see e.g. [14, Theorem 2.3]. It gives sufficient and "almost necessary" conditions for diagonal equality of the solutions of the Schur problem. The formulation and proof of the results given below are slightly different from the known ones.

Proposition 2. The following assertions hold true.

1. If $V = U_1^{\mathrm{H}} U_2 \in \mathbf{D}^*(n)$ 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 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

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

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

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

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

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

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

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

The MATLAB® command `[U,T] = schur(A)` computes a particular solution (U, T) of the SP for the matrix $A \in \mathbb{C}(n)$. The aim of the computation of the ConSF T of a general matrix $A \in \mathbb{C}(n)$ is to determine the eigenvalues of A as the diagonal elements of T. But the Schur problem may be defined also for matrices A which are already in ConSF, i.e. $A \in \mathbf{T}(n)$. For such matrices the above MATLAB® command computes the solution (I_n, A) of the Schur problem for A. At the same time the Schur problem for $A \in \mathbf{T}(n)$ has infinitely many solutions. In order to tie them down we introduce the following definition.

Definition 4. If $A \in \mathbf{T}(n)$ then the pair (I_n, A) is called the *principal solution* of the Schur problem for A. □

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 $\mathbf{U}(n)$. The only (albeit most important) invariants for this action which, revealed by the matrix T, are the eigenvalues $\lambda_k(A) = T(k,k)$, $k \in \mathbb{Z}[1,n]$, of the matrix A which appear on the diagonal of T.

The definition of complete invariants and canonical forms for the similarity action of $\mathbf{U}(n)$ on $\mathbb{C}(n)$, see [14], is more subtle 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 [1,9] and [3]. Note that from point of view of applications the condensed forms provide the same advantages as the canonical forms. Moreover, strict canonical forms of the matrix A are rarely (if ever) used in practice since they are usually defined by complicated conditions and procedures and are more sensitive to perturbations in A.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

Let $U \in \mathcal{U}(A)$ and $D \in \mathbf{D}^*(n)$. Then $UD \in \mathcal{U}(A)$ as well. Thus we have $cU \in \mathcal{U}(A)$, where $c \in \mathbb{C}$, $|c| = 1$, and $-U \in \mathcal{U}(A)$ in particular. 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 $U \in \mathcal{U}(A)$ and $V = -U \in \mathcal{U}(A)$.

Given the 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 ConSF T of $A \in \mathbb{C}(n)$ is unique if and only if $A = \lambda I_n$, where $\lambda \in \mathbb{C}$. 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 also differ in their strictly upper triangular parts.

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

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

different orderings of the elements $T(k,k)$ on the diagonal $\mathrm{Diag}(T)$ 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 $\mathrm{Diag}(T_{k,k}) = \lambda_k I_k$. In the generic case $m = n$ we have $N(1, 1, \ldots, 1) = n!$ 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 [14] 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 $A \in \mathbb{C}(n)$ the set

$$\mathrm{Or}[A, \mathbf{U}(n)] = \{U^{\mathrm{H}} A U : U \in \mathbf{U}(n)\} \subset \mathbb{C}(n)$$

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

Obviously $B \in \mathrm{Or}[A, \mathbf{U}(n)]$ implies $A \in \mathrm{Or}[B, \mathbf{U}(n)]$ and vice versa. Let $\mathfrak{A} \subset \mathbb{C}(n)$ and $\mathfrak{C} \subset \mathbf{T}(n)$ 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 $B \in \mathrm{Or}[A, \mathbf{U}(n)]$. □

Definition 7. The function $\gamma : \mathfrak{A} \to \mathfrak{C}$ is said to be a *canonical form* 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 $A \sim B$. □

Thus the canonical form $\gamma : \mathfrak{A} \to \mathfrak{C}$ is a *complete invariant* [6] for the similarity action of the group $\mathbf{U}(n)$ on the set $\mathfrak{A}$ but the opposite, of course, is not true. The canonical form γ thus defined is a

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

function. Informally, we also say that the image $\gamma(A) \in \mathbf{T}(n)$ of the matrix A under γ is a unitary canonical form, or Schur form, of A.

Definition 8. The subset $\mathfrak{A}$ of $\mathbb{C}(n)$ is said to be *closed* in the Zariski topology [6] 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 $\mathbb{C}(n) \backslash \mathfrak{A} \subset \mathbb{C}(n)$ is closed in this topology. □

Definition 9. 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. $A^{\mathrm{H}}A \neq AA^{\mathrm{H}}$; in particular the matrix A is not unitary.
3. The singular values of the matrix A are positive and pair-wise different; in particular $\mathrm{rank}(A) = n$.
4. The eigenvalues λ_k of the matrix A satisfy the inequalities $\mathrm{Re}(\lambda_k) \neq \mathrm{Re}(\lambda_l)$ and $\mathrm{Im}(\lambda_k) \neq \mathrm{Im}(\lambda_l)$ 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 $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \colon \mathbb{Z}[1,n] \to \mathbb{Z}[1,n]$ be a permutation of the integers $1, 2, \ldots, n$ and recall that $\mathbb{Z}[1,n-1] = \{1,2,\ldots,n-1\}$ and $\mathbb{Z}[2,n] = \{2,3,\ldots,n\}$. 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 pair-wise disjoint eigenvalues. Let $\mathcal{K}(n) \subset \mathbb{Z}_n^{n-1}$ be the set of $(n-1)$-tuples

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

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.

Definition 10. 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 11. The *conjugate* $(n-1)$-tuple of the $(n-1)$-tuple $\theta = (p_1, p_2, \ldots, p_{n-1})$, where $p_k = (i_k, j_k)$, $k \in \mathbb{Z}[1, n-1]$, is

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

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 $(1,n), (2,n-1), \ldots, (n,1)$ of $n \times n$ arrays.

Definition 12. The set $\mathcal{K}(n)$ has the following important subsets.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

1. The set $\mathcal{K}_1(n) \subset \mathcal{K}(n)$ is of *type 1* if its elements are of the form
$$\{(i_1, 1), (i_2, 2), \ldots, (i_{n-1}, n-1)\}$$
2. The set $\mathcal{K}_2(n) \subset \mathcal{K}(n)$ is of *type 2* if its elements are of the form
$$\{(1, j_1), (2, j_2), \ldots, (n-1, j_{n-1})\}$$
3. The set $\mathcal{K}_3(n) = \mathcal{K}(n) \backslash (\mathcal{K}_1(n) \cup \mathcal{K}_2(n)) \subset \mathcal{K}(n)$ 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
$$\theta^{@} = \{(1, 2), (2, 3), \ldots, (n-1, n)\}$$
which is a self-conjugate $(n-1)$-tuple. □

Definition 13. 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 1. 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 all they are 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*
$$\theta_1 = \{(1, 2), (1, 3), \ldots, (1, n)\}$$
and
$$\theta_{(n-1)!} = \theta^{@} = \{(1, 2), (2, 3), \ldots, (n-1, n)\},$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

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

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

resp. □

Now we are in position to define a possible set of Schur canonical forms $S \in \mathbf{T}(n)$ for generic matrices $A \in \mathbb{C}(n)$. There are $n!(2(n-1)!-1)$ such sets. The multiplier $n!$ comes from the different orderings 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

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

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

If we assume that the eigenvalues of S are ordered as $\lambda_1 \prec \lambda_2 \prec \cdots \prec \lambda_n$ then there remain $2(n-1)!-1$ sets of Schur canonical forms. Note that any fixed ordering of the (simple) eigenvalues of A on the diagonal of S is preserved only by unitary similarity transformations with matrices U, such that $U(k,k) = \exp(\mathrm{i}\varphi_k)$, $k \in \mathbb{Z}[1,n]$.

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

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

then the set of Schur canonical forms is uniquely fixed. In this case the Schur canonical forms 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. Two other Schur canonical forms 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 Jordan canonical forms of matrices $A \in \mathbb{C}$ relative to general similarity transformations. Usually it is assumed that different orderings of the Jordan blocks do not produce different Jordan forms. Formally this means that the Jordan canonical 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 14. A set $\mathfrak{C} \subset \mathbf{T}(n)$ of Schur canonical forms $T \in \mathbf{T}(n)$ for generic matrices $A \in \mathbb{C}(n)$ and a fixed $(n-1)$-tuple

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

is characterized as follows.

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

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

2. The $n-1$ elements $T(i_1, j_1), T(i_2, j_2), \ldots, T(i_{n-1}, j_{n-1})$ 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 Schur canonical form $C \in \mathbf{T}(n)$ by the next three steps.

1. The matrix A is transformed into any Schur condensed form $T_1 = U_1^{\mathrm{H}} A U_1 \in \mathbf{T}(n)$ by a matrix $U_1 \in \mathbf{U}(n)$. Numerically this is done by the QR algorithm [5]. For this purpose the code `schur` from MATLAB® may be used [10].
2. A condensed Schur form $T_2 = U_2^{\mathrm{H}} T_1 U_2 \in \mathbf{T}(n)$ is constructed so that $T_2(k,k) = \lambda_k$, $k \in \mathbb{Z}[1,n]$. This may be done by several 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. [5].
3. A diagonal matrix $U_3 \in \mathbf{U}(n)$ with elements $U_3(1,1) = 1$, $U_3(k,k) = \exp(\mathrm{i}\varphi_k)$, $\varphi_k \in \mathbb{R}$, $k \in \mathbb{Z}[2,n]$, is chosen so that the matrix $T = U_3^{\mathrm{H}} T_2 U_3$ has positive elements in positions (i_k, j_k).

We recall that to introduce Schur canonical forms 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 $z_k = x_k + \mathrm{i}y_k \in \mathbb{C}$, 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

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

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 at Table 1.

Table 1. Number of generic canonical forms

n	2	3	4	5	6	7	8	9	10
μ_n	1	3	20	210	3003	54264	1184040	30260340	886163135

There are μ_n different pairs of $p_k = (i_k, j_k)$, $1 \le i < j \le n$. 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 $p_1 \prec p_2 \prec \cdots \prec p_{\mu_n}$, where

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

Thus we have the chain of inequalities

$$\begin{aligned}(1,2) &\prec (1,3) \prec \cdots \prec (1,n) \prec (2,3) \prec \cdots \prec (2,n) \prec \cdots \\ &\prec (n-2, n-1) \prec (n-2, n) \prec (n-1, n)\end{aligned}$$

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 $a_n(i) \le k \le b_n(i)$, where

$$a_n(i) = 1 + \frac{(i-1)(2n-i)}{2}, \ b_n(i) = \frac{i(2n-i-1)}{2}.$$

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

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

or

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

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
$$(i,j) \mapsto k = k_n(i,j),\ k \mapsto (i,j) = (i_n(k), j_n(k)),$$
between the ordered sets of integers $\mathbb{Z}[1, \nu_n]$ and integer pairs (i,j), where $1 \le i < j \le n$.

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*. □

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 $\lambda_1 \prec \lambda_2 \prec \cdots \prec \lambda_n$ is already transformed into condensed Schur form, i.e. $A \in \mathbf{T}(n)$, it is then easily put into canonical form as follows. First a matrix $U \in \mathbf{U}(n)$ is chosen so as
$$\mathrm{Diag}(S) = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),\ S = U^{\mathrm{H}} A U.$$
Then a diagonal unitary matrix D with $D(1,1) = 1$ is found so that $T = D^{\mathrm{H}} S D \in \mathfrak{C}(\theta)$. Denoting
$$S(i,j) = |S(i,j)| \exp(\mathrm{i}\alpha(i,j))$$
and $D(i,i) = \exp(\mathrm{i}\varphi(i))$, $\varphi(1) = 0$, where $\alpha(i,j), \varphi(i) \in (-\pi, \pi]$, the conditions $S(i,j) > 0$ give the system of $n-1$ linear equations
$$\varphi(i) - \varphi(j) = \alpha(i,j),\ (i,j) \in \theta, \tag{3}$$
for $\varphi(2), \varphi(3), \ldots, \varphi(n)$. If it happens that $\varphi(k) \notin (-\pi, \pi]$ for some k then $\varphi(k)$ is replaced by
$$\widetilde{\varphi}(k) = \varphi(k) \mod (2\pi) \in (-\pi, \pi]. \tag{4}$$

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 easily solved explicitly. The first set of generic canonical forms corresponds to the pairs of indexes $(1,j)$, $j \in \mathbb{Z}[2, n]$, and here the solution of (3) is
$$\varphi(i) = -\alpha(1,i).$$
The second set corresponds to the index pairs (i,n), $i \in \mathbb{Z}[1, n-1]$, the solution of (3) being
$$\varphi(i) = \alpha(i,n) - \alpha(1,n).$$
The third set corresponds to the indexes $(i, i+1)$, $i \in \mathbb{Z}[1, n-1]$, and here the solution of (3) is
$$\varphi(i) = -\sum_{k=1}^{i-1} \alpha(k, k+1).$$
In all these cases the convention (4) is presupposed.

The restrictions assumed in this section, and in particular the condition that the eigenvalues of A are pair-wise distinct, seem serious, but in fact their violation can make the perturbation analysis of the Schur decomposition meaningless. If e.g. A has two or more equal eigenvalues then the Schur basis of the perturbed Schur problem may be discontinuous as a function of the perturbation in A, see e.g. [8] and [12].

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

When the matrix A is real (cases 1 and 2), i.e. $A \in \mathbb{R}(n)$, then 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 $U \in \mathbf{U}(n)$ is taken as orthogonal, i.e. $U \in \mathbf{O}(n)$, and both the Schur canonical form and the Schur condensed forms 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 [11, 14], 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 $\alpha_k + \mathrm{i}\beta_k, \alpha_k - \mathrm{i}\beta_k$ of A, where $\alpha_k, \beta_k \in \mathbb{R}$ and $\beta_k > 0$.

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

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

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

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

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

6. Perturbations

Let $(U, T) \in \mathbf{U}(n) \times \mathbf{T}(n)$ be a fixed solution of the Schur problem for the matrix $A \in \mathbb{C}(n)$ with the convention that if $A \in \mathbf{T}(n)$ then $U = I_n$ and $T = A$. 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. $\|\delta A\| \le \rho c \|A\|$, where ρ is the rounding unit of the binary floating-point arithmetic (FPA) used in the computations [2] and c is a small positive constant. In double precision FPA we have $\rho = 2^{-53} \simeq 1.1102 \times 10^{-16}$.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

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 1.

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

1. If the PSP for a slightly perturbed matrix $A + \delta A$ has a *slightly perturbed* solution $(U + \delta U, T + \delta T)$ with
$$\|\delta U\|,\ \|\delta T\| = \mathrm{O}(\delta),\ \delta = \|\delta A\| \to 0,$$
then it also has a *significantly perturbed* solution $(U + \Delta U, T + \delta T)$ with $\Delta U = -2U - \delta U$ and $\|\Delta U\| = 2 + \mathrm{O}(\delta)$. More precisely, we have
$$2 - \|\delta U\| \le \|2U + \delta U\| = \|\Delta U\| = \|U + (U + \delta U)\| \le 2.$$
2. The solution of the PSP for the matrix $A + \delta A$ may have the form $(U, T + \delta T)$ with $\delta U = 0$. This will happen if and only if $U^{\mathrm{H}}\delta A U \in \mathbf{T}(n)$ and in this case $\|\delta T\| = \|\delta A\|$.

Let $(U_0, T_0) \in \mathbf{U}(n) \times \mathbf{T}(n)$ be a solution of the Schur problem for the matrix $A_0 \in \mathbb{C}(n)$ under the convention that if $A_0 \in \mathbf{T}(n)$ then $U_0 = I_n$ and $T_0 = A_0$. Consider for simplicity the case $\delta A = \varepsilon A_1$, where $\varepsilon \in [0, \varepsilon_0)$, $\varepsilon_0 > 0$, is a small parameter and $A_1 \in \mathbb{C}(n)$ is a fixed matrix with $\|A_1\| = 1$. Let $(U(\varepsilon), T(\varepsilon)) \in \mathbf{U}(n) \times \mathbf{T}(n)$ be a solution to the PSP for the matrix $A(\varepsilon) := A_0 + \varepsilon A_1$, i.e.
$$T(\varepsilon) = U^{\mathrm{H}}(\varepsilon)A(\varepsilon)U(\varepsilon),\ U(\varepsilon) \in \mathbf{U}(n),\ T(\varepsilon) \in \mathbf{T}(n). \tag{5}$$
Since the solution of the PSP always exists, we have defined functions $U(\cdot) : [0, \varepsilon_0) \to \mathbf{U}(n)$ and $T(\cdot) : [0, \varepsilon_0) \to \mathbf{T}(n)$ through the 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 15. The pair $(U(\varepsilon), T(\varepsilon))$ is said to be a *regular solution* of the PSP for the matrix $A(\varepsilon) = A_0 + \varepsilon A_1$ if the functions $U(\cdot)$ and $T(\cdot)$ are continuous on the interval $[0, \varepsilon_0)$ and $(U(0), T(0))$ is the principal solution of the SP for A_0. □

A number of examples of condensed Schur forms presented in the next section illustrate the structure of these forms and the behavior of their perturbations, see also [7].

Example 2. Let $A \in \mathbb{C}(n)$ be a scalar matrix, i.e. $A = \lambda I_n$, $\lambda \in \mathbb{C}$. Then the general solution of the Schur problem 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 is a scalar matrix and $\mathcal{T}(A) = \{A\}$.
2. If $\mathcal{T}(A) = \{A\}$ then A is a scalar matrix and $\mathcal{U}(A) = \mathbf{U}(n)$. □

Example 3. Let $A = \lambda I_n + J_n$, where $\lambda \in \mathbb{C}$ and $J_n = [0, I_{n-1}; 0, 0]$ is the $n \times n$ Jordan block with zero eigenvalue. Then
$$\mathcal{U}(A) = \mathbf{U}(n) \cap \mathbf{D}(n),\ \mathcal{T}(A) = \{\lambda I_n + X : |X| = J_n\}. \quad \square$$

7. Examples of Real 2×2 Matrices

In this section we consider several examples illustrating the concepts introduced so far. In what follows "Schur form" means "condensed Schur form". The examples are for Schur problem and PSP for matrices $A \in \mathbb{R}(2)$ with real spectra for which the transformation group is $\mathbf{O}(2)$. This is the most

 doi:10.20944/preprints202410.0785.v1

simple non-trivial case. However, the effects observed are in fact valid for matrices $A \in \mathbb{C}(n)$, $n > 2$, e.g. of the form $A = [A_{1,1}, A_{1,2}; 0, A_{2,2}]$, where $A_{1,1} \in \mathbb{R}(2)$ and $A_{2,2} \in \mathbb{C}(n-2)$.

Matrices $A \in \mathbb{R}(2)$ correspond to linear operators $\mathbb{R}^2 \to \mathbb{R}^2$ 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. [4] and the examples below.

Example 4. Let the matrix $A \in \mathbb{R}(2)$ has eigenvalues λ_1, λ_2 and set

$$r = \sqrt{\|A\|_F^2 - |\lambda_1|^2 - |\lambda_2|^2}.$$

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 Schur form λI_2 of A.
- If $\lambda_1 = \lambda_2 = \lambda$ and $r > 0$ then there exist two Schur forms $\lambda I_2 \pm r E_{1,2}$ of A.
- If $\lambda_1 \neq \lambda_2$ and $r = 0$ then there exist two Schur forms $\mathrm{diag}(\lambda_1, \lambda_2)$ and $\mathrm{diag}(\lambda_2, \lambda_1)$ of A.
- If $\lambda_1 \neq \lambda_2$ and $r > 0$ then there exist four Schur forms

$$\mathrm{diag}(\lambda_1, \lambda_2) \pm r E_{1,2},\ \mathrm{diag}(\lambda_2, \lambda_1) \pm r E_{1,2}$$

  of A. □

Example 5. Let $A = \lambda I_2$, $\lambda \in \mathbb{R}$. We have $\mathcal{U}(\lambda I_2) = \mathbf{O}(2)$ and $\mathcal{T}(\lambda I_2) = \{\lambda I_2\}$. Since λI_2 is in Schur form, the principal solution of the Schur problem is $(I_2, \lambda I_2)$. Let the matrix λI_2 be perturbed to $\lambda I_2 + \varepsilon E_{2,1}$, $\varepsilon \neq 0$. Then the Schur decomposition of the perturbed matrix is

$$U = I_2 + \delta U,\ T = \lambda I_2 + \delta T.$$

The set of transformation matrices $\mathcal{U}(\lambda I_2 + \varepsilon E_{2,1})$ consists of 4 matrices $U_1,\ -U_1,\ U_2,\ -U_2$, where

$$U_1 = E_{1,2} + E_{2,1},\ U_2 = E_{1,2} - E_{2,1}.$$

In view of the equalities $U_k = I_2 + \delta U_k$, for two of these matrices we have

$$\|\delta U_1\| = \|I_2 \pm U_1\| = 2$$

and for the other two we have

$$\|\delta U_2\| = \|I_2 \pm U_2\| = \sqrt{2}.$$

At the same time the set of Schur forms $\mathcal{T}(\lambda I_2 + \varepsilon E_{2,1})$ consists of two matrices $\lambda I_2 \pm \varepsilon E_{1,2}$. Thus the transformation matrix $U(\lambda I_2 + \varepsilon E_{2,1})$ is discontinuous (or infinitely sensitive) as a function of the perturbation parameter ε at the point $\varepsilon = 0$.

Consider also the multivalued function $\Psi : \mathbb{R} \to 2^{\mathbf{O}(2)}$, where $2^{\mathbf{O}(2)}$ is the set of subsets of $\mathbf{O}(2)$, defined by

$$\varepsilon \mapsto \Psi(\varepsilon) = \mathcal{U}(\lambda I_2 + \varepsilon E_{2,1}).$$

We have $\Psi(0) = \mathbf{O}(2)$ and

$$\Psi(\varepsilon) = \{U_1, -U_1, U_2, -U_2\}$$

for $\varepsilon > 0$. Hence the function Ψ, i.e. the Schur basis for $\mathbb{R}(2,1)$ relative to the matrix $\lambda I_2 + \varepsilon E_{2,1}$, is discontinuous at the point $\varepsilon = 0$, while the Schur forms of $\lambda I_2 + \varepsilon E_{2,1}$ are continuous and well conditioned in ε. □

Example 6. 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

$$T_{0,1} = \lambda I_2 + E_{1,2},\ T_{0,2} = \lambda I_2 - E_{1,2},$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

while the set $\mathcal{U}(A_0)$ contains four matrices

$$I_2,\ -I_2,\ E_{1,1}-E_{2,2},\ E_{2,2}-E_{1,1}.$$

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

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

we see that there are four Schur forms

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

and

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

where $t(\varepsilon) = 1 - \varepsilon$. The orthonormal matrices $U_k(\varepsilon)$ that transform $A(\varepsilon)$ into Schur forms $T_k(\varepsilon)$, respectively, are

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

and

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

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 diagonally different Schur forms $T_{0,1}$ and $T_{0,2}$, respectively. □

Example 7. Let $A_0 = \mathrm{diag}(\lambda_1, \lambda_2)$, $\lambda_1 \neq \lambda_2$. Here the set $\mathcal{T}(A_0)$ contains two diagonally different Schur forms

$$T_{0,1} = \mathrm{diag}(\lambda_1, \lambda_2),\ T_{0,2} = \mathrm{diag}(\lambda_2, \lambda_1)$$

of A_0, while the set $\mathcal{O}(A_0)$ has 8 elements, namely

$$\pm E_{1,1} \pm E_{2,2},\ \pm E_{1,2} \pm E_{2,1}.$$

Let us again 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:

$$\widetilde{T}_1(\varepsilon) = T_{0,1} + \varepsilon E_{1,2},\ \widetilde{T}_2(\varepsilon) = T_{0,1} - \varepsilon E_{1,2}$$

and

$$\widetilde{T}_3(\varepsilon) = T_2 + \varepsilon E_{1,2},\ \widetilde{T}_4(\varepsilon) = T_2 - \varepsilon E_{1,2}.$$

The matrices $U_1, -U_1$ from Example 5 transform the perturbed matrix $A + \delta A(\varepsilon)$ into the Schur form $\widetilde{T}_3(\varepsilon)$ and the matrices $U_2, -U_2$ transform $A + \delta A(\varepsilon)$ into the Schur form $\widetilde{T}_4(\varepsilon)$ since U_1 and U_2 transform $\delta A(\varepsilon)$ in $\pm\varepsilon E_{1,2} \in \mathbf{T}(2)$, respectively.

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

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2024 doi:10.20944/preprints202410.0785.v1

where

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

We have

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

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

Example 8. Let

$$A_0 = \mathrm{diag}(\lambda_1, \lambda_2) + aE_{1,2} \in \mathbb{R}(2),$$

where $\delta = |\lambda_1 - \lambda_2| > 0$ and $a \neq 0$. The set $\mathcal{T}(A_0)$ of condensed Schur forms of A_0 contains 4 matrices:

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

The Schur canonical form of A_0 is

$$S_0 = \mathrm{diag}(\lambda_{\min}, \lambda_{\max}) + |a|E_{1,2} \in \mathcal{T}(A_0),$$

where

$$\lambda_{\min} = \min\{\lambda_1, \lambda_2\} < \lambda_{\max} = \max\{\lambda_1, \lambda_2\}.$$

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 condensed Schur forms of the matrix $A(\varepsilon)$ are

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

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

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

Among the four condensed forms only the matrix $\widetilde{T}_1 = \mathrm{diag}(\widetilde{\lambda}_1, \widetilde{\lambda}_2) + \widetilde{a}E_{1,2}$ is regular. □

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.

$$\mathrm{spect}(A) = \{A(1,1), A(2,2), \ldots, A(n,n)\} \tag{6}$$

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

$$\det(A - A(k,k)I_n) = 0, \ k \in \mathbb{Z}[1, n] \tag{7}$$

Definition 16. The matrices $A \in \Delta(n)$ which satisfy (6) or (7) are 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 elements of A and is hence a closed algebraic variety of complex dimension up to $n(n-1)$.

 doi:10.20944/preprints202410.0785.v1

Upper triangular matrices and lower triangular matrices are diagonally spectral. Schur condensed form 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 9. The elements of the matrix $A \in \Delta(2)$ satisfy one independent algebraic equation $A(1,2)A(2,1) = 0$. Hence matrices $A \in \Delta(2)$ have the form

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

where * denotes unspecified matrix elements. □

Example 10. The matrices $A_1, A_1^\top, A_2, A_2^\top, A_3, A_3^\top \in \mathbb{C}(3)$, where

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

are diagonally spectral. □

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

Example 11. 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.

$$\mathrm{spect}(A(z)) = \mathrm{spect}(B(z)) = \{1,2,3\}.$$

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

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

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

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

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

Important observation The requirement that the condensed form T 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.

Example 12. For the matrix $A = \lambda I_2$, $\lambda \in \mathbb{C}$, the pair (T,U) is $(\lambda I_2, I_2)$. If we perturb A to $\widetilde{A} = \lambda I_2 + \varepsilon E_{2,1}$, where $\varepsilon > 0$ is arbitrarily small, the pair (T,U) is transformed to $(\widetilde{T}, \widetilde{U})$, where $\widetilde{T} = \lambda I_2 + \varepsilon E_{1,2}$ and $\widetilde{U}$ is any of the four matrices $\pm E_{1,2} \pm E_{2,1}$. Thus $\|U - \widetilde{U}\|_{\mathrm{F}} = 2$ and the transformation matrix $U = U(\varepsilon)$ is even discontinuous at the point $\varepsilon = 0$. □

This high sensitivity may not be relevant to the problem of computing the spectrum $\mathrm{spect}(A)$ of A. To avoid such artificial sensitivity in the next section we introduce the concept of quasi-Schur condensed forms.

9. Quasi-Schur Condensed Forms

Definition 17. A matrix $S = U^{\mathrm{H}} A U \in \mathbb{C}(n)$, $U \in \mathbf{U}(n)$, is said to be a *quasi-Schur condensed form* of $A \in \mathbb{C}(n)$ if it is block-upper triangular with $m \geq 1$ diagonal blocks $S_{k,k} \in \mathbb{C}(n_k)$, $n = n_1 + n_2 + \cdots + n_m$, 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 $S_{k,k}^{\top} \in \mathbf{T}(n_k)$, $k \in \mathbb{Z}[1, m]$. □

Example 13. For $n = 2$ the quasi-Schur condensed forms have the structure $A_1, A_1^{\top}$, where $A_1 \in \mathbf{T}(2)$. For $n = 3$ the quasi-Schur condensed forms are $A_1, A_1^{\top}, A_2, A_2^{\top}$ and $A_3, A_3^{\top}$, where $A_1 \in \mathbf{T}(3)$,

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

and $*$ denotes unspecified quantities. □

Quasi-Schur condensed forms are diagonally spectral but the opposite is not true for $n \geq 3$, see e.g. Example 11.

Obviously a Schur condensed form is also a quasi-Schur condensed form but the opposite may not be true (we recall that $n \geq 2$). We stress that high sensitivity of Schur forms as in Example 12 may not be observed for quasi-Schur condensed forms.

10. Conclusions

In this paper we have considered the Schur canonical forms for a square matrix A with pair-wise distinct eigenvalues. Sensitivity of the Schur form relative to perturbations in A was also studied. The concept of regular solution to the perturbed Schur form was introduced and illustrated by a number of examples. We have also introduced the concepts of diagonally spectral matrices (Schur forms are diagonally spectral) and of quasi-Schur condensed forms of a matrix A which may be much less sensitive to perturbations in A.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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. J. Brenner, The problem of unitary equivalence, *Acta Mathematica*, **86** (1951), pp. 297-308, doi:10.1007/BF02392670
2. IEEE Computer Society, *754-2019-IEEE Standard for Floating-Point Arithmetic*, https://ieeexplore.ieee.org/document/8766229, 2019, doi:10.1109/IEEESTD.2019.8766229, ISBN 978-1504459242
3. K. Ikramov, The canonical Schur form of a matrix with simple eigenvalues, *Doklady Mathematics*, **77** (2008), no. 3, pp. 359-360, doi:10.1134/S1064562408030101

4. M. Glazman and Ju. Ljubich, *Finite-Dimensional Linear Analysis: A Systematic Presentation in Problem Form*, New York, Dover Books in Mathematics, Dover Publications, 2006, ISBN 978-0486453323
5. G. Golub and C. Van Loan, *Matrix Computations*, Baltimore, The Johns Hopkins University Press (4th edition), 2013, ISBN 978-1421407944
6. R. Hartshorne, *Algebraic Geometry*, Berlin, Springer-Verlag, 1977, ISBN 978-0387902449
7. M. Konstantinov and P. Petkov, *Perturbation Methods in Matrix Analysis and Control*, New York, Nova Science Publishers, 2020, ISBN 978-1536174700, BISAC:MAT034000, doi:10.52305/EZCI5944
8. M. Konstantinov, P. Petkov and N. Christov, Nonlocal perturbation analysis of the Schur system of a matrix, *SIAM Journal on Matrix Analysis and Applications*, **15** (1994), no. 2, pp. 383-392, doi.org/10.1137/S089547989120267X
9. D. Littlewood, On unitary equivalence, *Journal of the London Mathematical Society*, **28** (1953), pp. 314-322, doi.org/10.1112/jlms/s1-28.3.314
10. MATLAB®, *Math. Graphics. Programming*, Natick, MathWorks Inc. (release R2024b), 2024, www.mathworks.com
11. F. Murnaghan and A. Wintner, A canonical form for real matrices under orthogonal transformations. *Proceedings of the National Academy of Sciences of the USA*, **17**(7) (1931), pp. 417-420, doi:10.1073/pnas.17.7.417
12. P. Petkov, Componentwise perturbation analysis of the Schur decomposition of a matrix, *SIAM Journal on Matrix Analysis and Applications*, **42** (2021), no. 1, pp. 108-133, doi.org/10.1137/20M1330774
13. I. Schur, Beiträge zur Theorie der Gruppen linearer homogener Substitutionen. *Transactions of the American Mathematical Society*, **10** (1909), pp. 159-175, doi:10.2307/1988679.
14. H. Shapiro, A survey of canonical forms and invariants for unitary similarity. *Linear Algebra and its Applications*, **147** (1991), pp. 101-167, doi:10.1016/0024-3795(91)90232-L