On the consistency of the matrix equation $X^\top AX = B$ when B is symmetric: the case where CFC(A) includes skew-symmetric blocks

Alberto Borobia[*], Roberto Canogar[†] and Fernando De Terán[‡]

March 17, 2022

Abstract

In this paper, which is a follow-up to [A. Borobia, R. Canogar, F. De Terán, *Mediterr. J. Math.* 18, 40 (2021)], we provide a necessary and sufficient condition for the matrix equation $X^\top AX = B$ to be consistent when B is symmetric. The condition depends on the canonical form for congruence of the matrix A, and is proved to be necessary for all matrices A, and sufficient for most of them. This result improves the main one in the previous paper, since the condition is stronger than the one in that reference, and the sufficiency is guaranteed for a larger set of matrices (namely, those whose canonical form for congruence, CFC(A), includes skew-symmetric blocks).

Keywords: Matrix equation, transpose, congruence, $\top$-Riccati equation, Canonical Form for Congruence, symmetric matrix, bilinear form.

Mathematics subject classification MSC2020: 15A21, 15A24, 15A63.

1 Introduction

Let $A \in \mathbb{C}^{n\times n}$ and let $B \in \mathbb{C}^{m\times m}$ be a symmetric matrix. We are interested in the consistency of the matrix equation

$$X^\top AX = B, \tag{1}$$

where $(\cdot)^\top$ denotes the transpose. To be more precise, we want to obtain necessary and sufficient conditions for (1) to be consistent. The main tool to get these conditions is the *canonical form for congruence*, CFC (see Theorem 1), because (1) is consistent if and only if the equation that we obtain after replacing the matrices A and/or B by their CFCs is consistent. The CFC is a direct sum of three kinds of blocks of different sizes, named Type-0, Type-I, and Type-II, and the idea is to take advantage of this structure to analyze Eq. (1). In particular, the only symmetric canonical blocks are $I_1 = [1]$ and $0_1 = [0]$, so the CFC of the symmetric matrix B is of the form $\mathrm{CFC}(B) = I_m \oplus 0_k$ (where I_m and 0_k are, respectively, a direct sum of m and k copies of I_1 and 0_1). With the help of Lemma 2 we can get rid of the null block 0_k, so the equation we are interested in is

$$X^\top AX = I_m \tag{2}$$

with $m \geqslant 1$.

In [4] we introduced $\tau(A)$, a quantity that depends on the number of certain Type-0, Type-I, and Type-II blocks appearing in the CFC of A, and we proved in [4, Th. 2] that if Eq. (2) is consistent then $m \leqslant \tau(A)$. Moreover, the main result of that paper, [4, Th. 8], establishes that if the CFC of A contains neither $H_2(-1)$ nor $H_4(1)$ blocks (which are specific Type-II blocks) then Eq. (2) is consistent if and only if $m \leqslant \tau(A)$. This is not necessarily true if we allow the CFC of A to contain blocks $H_2(-1)$ and/or $H_4(1)$ (for instance, it is not true for $A = H_2(-1)$ nor $A = H_4(1)$).

In the present work we introduce a new quantity $\upsilon(A)$, that depends also on the number of certain Type-0, Type-I, and Type-II blocks appearing in the CFC of A. In Theorem 7 we will prove that if Eq. (2) is consistent then $m \leqslant \min\{\tau(A), \upsilon(A)\}$. Moreover, according to the main result in the present work (Theorem 12), if the CFC of A does not contain $H_4(1)$ blocks, then Eq. (2) is consistent if and only if $m \leqslant \min\{\tau(A), \upsilon(A)\}$. However, this is not necessarily true if the CFC contains blocks $H_4(1)$ (it is not true, for instance, for $A = H_4(1)$).

Note that the main result of this paper improves the main one in [4] in two senses: (i) the condition here is stronger than the one there; and (ii) the characterization is guaranteed for a larger set of matrices.

[*]Facultad de Ciencias, Universidad Nacional de Educación a Distancia (UNED), email: aborobia@mat.uned.es

[†]Facultad de Ciencias, Universidad Nacional de Educación a Distancia (UNED), email: rcanogar@mat.uned.es

[‡]Departamento de Matemáticas, Universidad Carlos III de Madrid, email: fteran@math.uc3m.es

In the title we have referred to "the case where CFC(A) includes skew-symmetric blocks". This highlights the fact that, compared to [4], in the present work the main result is applied to matrices whose CFC contains $H_2(-1)$ blocks, which are the only nonzero skew-symmetric blocks in a CFC.

The interest on Eq. (1) goes back to, at least, the 1920's [16], and it has been mainly devoted to describing the solution, X, for matrices A, B over finite fields and when A and/or B have some specific structure [7–11, 13, 17]. More recently, some related equations have been analyzed [15] and, in particular, in connection with applications [1–3]. In [5] we have addressed the consistency of Eq. (1) when B is skew-symmetric, where it is emphasized the connection between the consistency of (1) and the dimension of the largest subspace of $\mathbb{C}^n$ for which the bilinear form represented by A is skew-symmetric and non-degenerate. The same connection holds after replacing skew-symmetric by symmetric, which is the structure considered in the present work.

The paper is organized as follows. In Section 2 we introduce the basic notation and definitions (like the CFC), and we also recall some basic results that are used later. In Section 3 the quantities $\tau(A)$ and $\upsilon(A)$ are introduced. Section 4 presents the necessary condition for Eq. (2) to be consistent (Theorem 7), whereas in Section 6 we show that when the CFC of A does not contain blocks $H_4(1)$ this condition is sufficient as well (Theorem 12). In between these two sections, Section 5 is devoted to introduce the tools (by means of several technical lemmas) that are used to prove the sufficiency of the condition. Finally, in Section 7 we summarize the main contributions of this work and indicate the main related open question.

2 Basic approach and definitions

Throughout the manuscript, I_n and 0_n denote, respectively, the identity and the null matrix with size $n \times n$. By $0_{m\times n}$ we denote the null matrix of size $m \times n$. By $\mathtt{i}$ we denote the imaginary unit (namely, $\mathtt{i}^2 = -1$), and by e_j we denote the jth canonical vector (namely, the jth column of the identity matrix) of the appropriate size. The notation $M^{\oplus k}$ stands for a direct sum of k copies of the matrix M.

Following the approach in [4] and [5], a key tool in our developments is the *canonical form for congruence* (CFC). For the ease of reading we first recall the CFC, that depends on the following matrices:

- $J_k(\mu) := \begin{bmatrix} \mu & 1 & & \\ & \ddots & \ddots & \\ & & \mu & 1 \\ & & & \mu \end{bmatrix}$ is a $k \times k$ *Jordan block associated with* $\mu \in \mathbb{C}$;
- $\Gamma_k := \begin{bmatrix} 0 & & & & (-1)^{k+1} \\ & & & \iddots & (-1)^k \\ & & -1 & \iddots & \\ & 1 & 1 & & \\ -1 & -1 & & & \\ 1 & 1 & & & 0 \end{bmatrix}_{k\times k}$ for $k \geqslant 1$ (note that $\Gamma_1 = I_1 = [1]$); and
- $H_{2k}(\mu) := \begin{bmatrix} 0 & I_k \\ J_k(\mu) & 0 \end{bmatrix}$, for $k \geqslant 1$, where $J_k(\mu)$ is a $k \times k$ Jordan block associated with $\mu \in \mathbb{C}$.

Theorem 1. (Canonical form for congruence, CFC) [14, Th. 1.1]. *Each square complex matrix is congruent to a direct sum, uniquely determined up to permutation of addends, of canonical matrices of the following three types*

Type 0	$J_k(0)$
Type I	Γ_k
Type II	$H_{2k}(\mu)$, $0 \neq \mu \neq (-1)^{k+1}$ (*μ is determined up to replacement by μ^{-1}*)

Following [5], the notation $A \rightsquigarrow B$ means that the equation $X^\top AX = B$ is consistent, and $A \overset{X_0}{\rightsquigarrow} B$ means that $X_0^\top AX_0 = B$. The following result, that was presented in [5, Lemma 4], includes some basic laws of consistency that are straightforward to check.

Lemma 2. (**Laws of consistency**). *For any complex square matrices A, B, C, A_i, B_i, the following properties hold:*

(i) **Addition law.** *If $A_i \overset{X_i}{\rightsquigarrow} B_i$, for $1 \leqslant i \leqslant k$, then $\bigoplus_{i=1}^k A_i \overset{X}{\rightsquigarrow} \bigoplus_{i=1}^k B_i$, with $X = \bigoplus_{i=1}^k X_i$.*

(ii) **Transitivity law.** *If $A \overset{X_0}{\rightsquigarrow} B$ and $B \overset{Y_0}{\rightsquigarrow} C$, then $A \overset{X_0Y_0}{\rightsquigarrow} C$.*

(iii) **Permutation law.** *$\bigoplus_{i=1}^\ell A_i \rightsquigarrow \bigoplus_{i=1}^\ell A_{\sigma(i)}$, for any permutation σ of $\{1, \ldots, \ell\}$.*

(iv) **Elimination law.** $A \oplus B \overset{X_0}{\rightsquigarrow} A$, *with* $X_0 = \left[\begin{smallmatrix} I_n \\ 0 \end{smallmatrix}\right]$, *and where* n *is the size of* A.

(v) **Canonical reduction law.** *If* A *and* B *are congruent to, respectively,* $\widetilde{A}$ *and* $\widetilde{B}$, *then* $A \rightsquigarrow B$ *if and only if* $\widetilde{A} \rightsquigarrow \widetilde{B}$.

(vi) $J_1(0)$**-law.** *For* $k, \ell \geqslant 0$ *we have* $A \oplus J_1(0)^{\oplus k} \rightsquigarrow B \oplus J_1(0)^{\oplus \ell}$ *if and only if* $A \rightsquigarrow B$.

By the Canonical reduction law, in Eq. (1) we will assume without loss of generality that A and B are given in CFC.

When B is symmetric, the CFC of B is $I_{m_1} \oplus 0_{m_2}$. Then, as a consequence of the Canonical reduction law, we may restrict ourselves to the case where the right-hand side of (1) is of this form. Moreover, as a consequence of the $J_1(0)$-law, in our developments we will consider $B = I_m$ in Eq. (1) (leading to Eq. (2)). Therefore, our goal is to characterize those matrices A such that $A \rightsquigarrow I_m$, for a fixed $m \geqslant 1$. This will be done by concatenating several equations $A \rightsquigarrow A_1 \rightsquigarrow \cdots \rightsquigarrow A_k \rightsquigarrow I_m$, since the Transitivity law allows us to conclude that $A \rightsquigarrow I_m$. For this reason, we will use the word "transformation" for a single equation $A \rightsquigarrow B$.

One way to determine the CFC of an invertible matrix A is by means of its *cosquare*, $A^{-\top}A$ (see [14]), where $(\cdot)^{-\top}$ denotes the transpose of the inverse. Moreover, the cosquare will be used to determine whether two given invertible matrices are congruent, using the following result.

Lemma 3. ([14, Lemma 2.1]). *Two invertible matrices are congruent if and only if their cosquares are similar.*

2.1 The matrices $\widetilde{\Gamma}_k$ and $\widetilde{H}_{2k}(\mu)$

Instead of the blocks Γ_k and $H_{2k}(\mu)$ we will use the following blocks, for $k \geqslant 1$:

$$\widetilde{\Gamma}_k := \begin{bmatrix} 1 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & 1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & (-1)^k & 0 & 1 \\ & & & & (-1)^{k+1} & 0 \end{bmatrix}_{k\times k} \quad \text{and} \quad \widetilde{H}_{2k}(\mu) := \begin{bmatrix} 0 & 1 & & & & & \\ \mu & 0 & 1 & & & \mathbf{0} & \\ & 0 & 0 & 1 & & & \\ & & \mu & 0 & 1 & & \\ & & & 0 & \ddots & \ddots & \\ & \mathbf{0} & & & \ddots & 0 & 1 \\ & & & & & \mu & 0 \end{bmatrix}_{2k\times 2k}.$$

We claim that $\widetilde{\Gamma}_k$ and $\widetilde{H}_{2k}(\mu)$ are congruent to, respectively, Γ_k and $H_{2k}(\mu)$.

In order to prove that Γ_k and $\widetilde{\Gamma}_k$ are congruent, we give an indirect proof. Two matrix pairs (A, B) and (A', B') are *strictly equivalent* if there are invertible matrices R and S such that $RAS = A'$ and $RBS = B'$. It is known (see, for instance, [6, Lemma 1]) that two matrices $A, B \in \mathbb{C}^{n\times n}$ are congruent if and only if $(A, A^\top)$ and $(B, B^\top)$ are strictly equivalent. Since $(\Gamma_k, \Gamma_k^\top)$ and $\left(J_k\left((-1)^{k+1}\right), I_k\right)$ are strictly equivalent (see [6, Th. 4]) and $\left(J_k\left((-1)^{k+1}\right), I_k\right)$ and $(\widetilde{\Gamma}_k, \widetilde{\Gamma}_k^\top)$ are strictly equivalent as well (see Eq. (5) in [12]), the pairs $(\Gamma_k, \Gamma_k^\top)$ and $(\widetilde{\Gamma}_k, \widetilde{\Gamma}_k^\top)$ are strictly equivalent, so Γ_k and $\widetilde{\Gamma}_k$ are congruent. Another alternative to show that Γ_k and $\widetilde{\Gamma}_k$ are congruent is by checking that their cosquares are similar to $J_k((-1)^{k+1})$ and then using Lemma 3.

To see that $H_{2k}(\mu)$ and $\widetilde{H}_{2k}(\mu)$ are congruent, consider the permutation matrix

$$P_{2k} = \begin{bmatrix} e_1 & e_{k+1} & e_2 & e_{k+2} & \cdots & e_k & e_{2k} \end{bmatrix},$$

and note that

$$\widetilde{H}_{2k}(\mu) = P_{2k}^\top H_{2k}(\mu) P_{2k} = P_{2k}^\top \begin{bmatrix} 0 & I_k \\ J_k(\mu) & 0 \end{bmatrix} P_{2k}.$$

Therefore, the congruence by P_{2k} is actually a simultaneous permutation of rows and columns of $H_{2k}(\mu)$. More precisely, we start with $\left[\begin{smallmatrix} 0 & I_k \\ J_k(\mu) & 0 \end{smallmatrix}\right]$ and move rows (and columns) $(k+1, k+2, \ldots, 2k)$ to, respectively, rows (and columns) $(2, 4, \ldots, 2k)$; and we also move rows (and columns) $(1, 2, \ldots, k)$ to rows (and columns) $(1, 3, \ldots, 2k-1)$, respectively. So the 1's coming from the block I_k and the 1's coming from the superdiagonal of the block $J_k(\mu)$ in $H_{2k}(\mu)$, get shuffled to form the superdiagonal of $P_{2k}^\top H_{2k}(\mu) P_{2k}$. Moreover, the μ's from the block $J_k(\mu)$ in $H_{2k}(\mu)$ are taken to the positions $(2,1), (4,3), \ldots, (2k, 2k-1)$ in $\widetilde{H}_{2k}(\mu)$.

The advantage in using the matrices $\widetilde{\Gamma}_k$ and $\widetilde{H}_{2k}(\mu)$ instead of, respectively, Γ_k and $H_{2k}(\mu)$, is that the first ones are tridiagonal, and this structure is more convenient for our proofs. Tridiagonal canonical blocks have been already used in [12] (actually, $\widetilde{\Gamma}_k$ is exactly the one introduced in Eq. (3) for $\varepsilon = 1$ in that reference).

For the rest of the manuscript, we will replace the blocks Γ_k by $\widetilde{\Gamma}_k$ and $H_{2k}(\mu)$ by $\widetilde{H}_{2k}(\mu)$, so, in particular, we will assume that the CFC is a direct sum of blocks $J_k(0)$, $\widetilde{\Gamma}_k$, and $\widetilde{H}_{2k}(\mu)$. The only exceptions to this rule are Γ_1 which is equal to $\widetilde{\Gamma}_1$, and $H_2(-1)$ which is equal to $\widetilde{H}_2(-1)$.

A	conditions	$\tau(A)$	$\upsilon(A)$
$J_1(0)$	–	0	0
$J_3(0)$	–	2	2
Γ_1	–	1	1
$\widetilde{\Gamma}_2$	–	1	1
$J_{2k+1}(0)$	$k \geqslant 2$	$k+1$	$2k$
$J_{2k}(0)$	$k \geqslant 1$	k	$2k$
$\widetilde{\Gamma}_{2k+1}$	$k \geqslant 1$	$k+1$	$2k+1$
$\widetilde{\Gamma}_{2k}$	$k \geqslant 2$	k	$2k-1$
$\widetilde{H}_{4k-2}(-1)$	$k \geqslant 2$	$2k-1$	$4k-4$
$\widetilde{H}_{4k}(1)$	$k \geqslant 1$	$2k+1$	$4k$
$\widetilde{H}_{2k}(\mu)$	$k \geqslant 1, \mu \neq 0, \pm 1$	k	$2k$
$H_2(-1)$	–	1	0

Table 1: Values of τ and υ for any single canonical block.

3 The quantities $\tau(A)$ and $\upsilon(A)$

The main result of this work (Theorem 12) depends on two intrinsic quantities of the matrix A, that we denote by $\tau(A)$ and $\upsilon(A)$. In this section, we introduce them and present some basic properties that will be used later.

Definition 4. Let A be a complex $n \times n$ matrix and consider its CFC, where

(i) j_1 is the number of Type-0 blocks with size 1;

(ii) $j_{\mathcal{O}}$ is the number of Type-0 blocks with odd size at least 3;

(iii) $\gamma_{\mathcal{O}}$ is the number of Type-I blocks with odd size;

(iv) γ_{ε} is the number of Type-I blocks with even size;

(v) $h^-_{2\mathcal{O}}$ is the number of Type-II blocks $\widetilde{H}_{4k-2}(-1)$ for any $k \geqslant 1$; and

(vi) $h^+_{2\varepsilon}$ is the number of Type-II blocks $\widetilde{H}_{4k}(1)$ for any $k \geqslant 1$;

(vii) it has an arbitrary number of other Type-0 and Type-II blocks.

Then we define the quantities

$$\tau(A) := \frac{n - j_1 + j_{\mathcal{O}} + \gamma_{\mathcal{O}} + 2h^+_{2\varepsilon}}{2} \quad \text{and} \quad \upsilon(A) := n - j_1 - j_{\mathcal{O}} - \gamma_{\varepsilon} - 2h^-_{2\mathcal{O}}. \tag{3}$$

The quantities τ and υ satisfy the following essential additive properties (the proof is straightforward):

$$\tau(A_1 \oplus \cdots \oplus A_k) = \tau(A_1) + \cdots + \tau(A_k) \quad \text{and} \quad \upsilon(A_1 \oplus \cdots \oplus A_k) = \upsilon(A_1) + \cdots + \upsilon(A_k). \tag{4}$$

The notation for the quantities in Definition 4 follows the one in [5]. In particular, the letters used for the number of blocks in parts (i)–(vi) resemble the notation for the corresponding blocks (see [5, Rem. 6]). In [4] we had not yet adopted this notation. The correspondence between the notation in that paper and the one used here is the following: $d \to j_1, r \to j_{\mathcal{O}}, s \to \gamma_{\mathcal{O}}, t \to h^+_{2\varepsilon}$. The values γ_{ε} and $h^-_{2\mathcal{O}}$ played no role in [4].

Table 1 contains the values of $\tau(A)$ and $\upsilon(A)$ for A being a single canonical block in the CFC. We have displayed the values in three categories, from top to bottom, namely: first, those with $\tau(A) = \upsilon(A)$; second, those for which $\tau(A) < \upsilon(A)$; and, finally, those with $\tau(A) > \upsilon(A)$.

Notice that $\tau(A) \leqslant \upsilon(A)$ whenever the CFC of A consists of just a single canonical block, except for $H_2(-1)$. This, together with (4), implies the following result.

Lemma 5. *If the* CFC *of A has no blocks of type $H_2(-1)$ then $\tau(A) \leqslant \upsilon(A)$.*

In order for the condition that we obtain (in Theorem 7) to be sufficient, the following notion is key.

Definition 6. The transformation $A \rightsquigarrow B$ is $(\tau, \upsilon)-$*invariant* if the following three conditions are satisfied:

- $X^\top AX = B$ is consistent,
- $\tau(A) = \tau(B)$, and
- $\upsilon(A) = \upsilon(B)$.

4 A necessary condition

In this section, we introduce a necessary condition on the matrix A for $A \rightsquigarrow I_m$ (namely, for Eq. (1) to be consistent when B is symmetric and invertible). This condition improves the one provided in [4, Th. 2], namely $m \leqslant \tau(A)$.

Theorem 7. *If A is a complex square matrix such that $X^\top AX = I_m$ is consistent, then $m \leqslant \min\{\tau(A), \upsilon(A)\}$.*

Proof. In [4, Th. 2] it was proved that $m \leqslant \tau(A)$ (though the notation τ was not used there). Let us see that $m \leqslant \upsilon(A)$ as well. Assuming that the CFC of A is as in Definition 4, in the proof of Theorem 8 of [5] it was showed that

$$n - \operatorname{rank}(A + A^\top) = j_{\mathcal{O}} + \gamma_\varepsilon + 2h^-_{2\mathcal{O}}. \tag{5}$$

By hypothesis, there exists some $X_0 \in \mathbb{C}^{n\times m}$ such that $X_0^\top AX_0 = I_m$. Now, transposing this equation and adding it up, we get $X_0^\top(A + A^\top)X_0 = 2I_m$. From this identity, and using (5), we obtain

$$m = \operatorname{rank}(X_0^\top(A + A^\top)X_0) \leqslant \operatorname{rank}(A + A^\top) = n - j_{\mathcal{O}} - \gamma_\varepsilon - 2h^-_{2\mathcal{O}},$$

so $m \leqslant n - j_{\mathcal{O}} - \gamma_\varepsilon - 2h^-_{2\mathcal{O}} = \upsilon(A)$, as claimed. □

5 Absorbing the $H_2(-1)$ blocks

The main goal in the rest of the manuscript is to prove that the necessary condition presented in Theorem 7 is also sufficient when the CFC of A does not contain $\widetilde{H}_4(1)$ blocks. If the CFC of A contains neither $H_2(-1)$ nor $\widetilde{H}_4(1)$ blocks, this is already known [4, Th. 8]. In that case, as a consequence of Lemma 5, the condition for $A \rightsquigarrow I_m$ reduces to $m \leqslant \tau(A)$. When the CFC of A does not contain blocks $\widetilde{H}_4(1)$ but contains blocks $H_2(-1)$, this is no longer true (see, for instance, Example 1 in [4]), and then the quantity $\upsilon(A)$ comes into play. This is an indication that the presence of blocks $H_2(-1)$ in the CFC of A deserves a particular treatment. In this section, we show how to deal with this type of blocks. To be more precise, we see that some blocks $H_2(-1)$ can be combined with other type of blocks in order to "eliminate" them by means of a $(\tau, \upsilon)-$invariant transformation. In this case, we say that the block $H_2(-1)$ has been "absorbed". We will consider separately the cases of Type-0, Type-I, and Type-II blocks, in Sections 5.1, 5.2, and 5.3, respectively.

The following notation is used in the proofs of this section: $E_{\alpha\times\beta}$ denotes the $\alpha \times \beta$ matrix whose $(\alpha, 1)$ entry is equal to 1 and the remaining entries are zero.

5.1 The case of Type-0 blocks

In Lemma 8, we show how to "absorb" a block $H_2(-1)$ with a Type-0 block, $J_k(0)$, with $k \neq 3$. In the statement, $J_0(0)$ stands for an empty block.

Lemma 8. *The following transformation is $(\tau, \upsilon)-$invariant:*

$$J_k(0) \oplus H_2(-1) \rightsquigarrow J_{k-2}(0) \oplus \Gamma_1^{\oplus 2}, \qquad \text{for } k = 2 \text{ and } k \geqslant 4. \tag{6}$$

Proof. By considering separately the cases where k in (6) is odd ($k = 2t + 1$) and even ($k = 2t$), using (4) and looking at Table 1, we obtain:

$$\begin{array}{rcccl}
\tau\big(J_{2t+1}(0) \oplus H_2(-1)\big) & = & t+2 & = & \tau\big(J_{2t-1}(0) \oplus \Gamma_1^{\oplus 2}\big), \quad \text{for } t \geqslant 2, \\
\upsilon\big(J_{2t+1}(0) \oplus H_2(-1)\big) & = & 2t & = & \upsilon\big(J_{2t-1}(0) \oplus \Gamma_1^{\oplus 2}\big), \quad \text{for } t \geqslant 2, \\
\tau\big(J_{2t}(0) \oplus H_2(-1)\big) & = & t+1 & = & \tau\big(J_{2t-2}(0) \oplus \Gamma_1^{\oplus 2}\big), \quad \text{for } t \geqslant 1, \\
\upsilon\big(J_{2t}(0) \oplus H_2(-1)\big) & = & 2t & = & \upsilon\big(J_{2t-2}(0) \oplus \Gamma_1^{\oplus 2}\big), \quad \text{for } t \geqslant 1,
\end{array}$$

so both sides of the transformation in (6) have the same τ and υ. Now let us prove the consistency.

The result is true for $k = 2$, since

$$J_2(0) \oplus H_2(-1) \overset{X_2}{\rightsquigarrow} \Gamma_1^{\oplus 2}, \quad \text{for } X_2 = \begin{bmatrix} \mathrm{i} & 1 \\ -\mathrm{i} & 1 \\ 0 & 1 \\ \mathrm{i} & 0 \end{bmatrix}.$$

Let us prove it for $k \geqslant 4$. Note that $J_{a+b}(0) = \left[\begin{smallmatrix} J_a(0) & E_{a\times b} \\ 0 & J_b(0)\end{smallmatrix}\right]$. If $X_3 = \left[\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & & \\ -1 & & X_2 \\ -1 & & \\ 1 & & \end{array}\right]$ then

$$\begin{aligned}
(I_{k-3}\oplus X_3)^\top \Big(J_k(0)\oplus H_2(-1)\Big)(I_{k-3}\oplus X_3) &= \begin{bmatrix} I_{k-3} & 0 \\ 0 & X_3^\top \end{bmatrix}\begin{bmatrix} J_{k-3}(0) & E_{(k-3)\times 5} \\ 0 & J_3(0)\oplus H_2(-1)\end{bmatrix}\begin{bmatrix} I_{k-3} & 0 \\ 0 & X_3\end{bmatrix} \\
&= \begin{bmatrix} J_{k-3}(0) & E_{(k-3)\times 5}X_3 \\ 0 & X_3^\top\,(J_3(0)\oplus H_2(-1))\,X_3\end{bmatrix} \\
&= \begin{bmatrix} J_{k-3}(0) & E_{(k-3)\times 3} \\ 0 & J_1(0)\oplus \Gamma_1^{\oplus 2}\end{bmatrix} \\
&= J_{k-2}(0)\oplus \Gamma_1^{\oplus 2},
\end{aligned}$$

as wanted. □

We will also use the following result, whose proof is straightforward.

Lemma 9. *The following transformation is $(\tau,\upsilon)-$invariant:*

$$J_3(0) \overset{X_0}{\rightsquigarrow} \Gamma_1^{\oplus 2}, \qquad \text{for } X_0 = \left[\begin{smallmatrix} 1 & 0 \\ 1 & \mathtt{i} \\ 0 & -\mathtt{i}\end{smallmatrix}\right].$$

5.2 The case of Type-I blocks

Lemma 10 is the counterpart of Lemma 8 for Type-I blocks, where Γ_k is replaced by $\widetilde{\Gamma}_k$.

Lemma 10. *The following transformation is $(\tau,\upsilon)-$invariant:*

$$\widetilde{\Gamma}_k \oplus H_2(-1) \rightsquigarrow \Gamma_1^{\oplus 2}\oplus \widetilde{\Gamma}_{k-2}, \quad \text{for } k\geqslant 3. \tag{7}$$

Proof. Considering again separately the cases where k in (7) is odd ($k=2t+1$) and even ($k=2t$), using (4) and looking at Table 1, we obtain:

$$\begin{array}{rclcll}
\tau\big(\widetilde{\Gamma}_{2t+1}\oplus H_2(-1)\big) & = & t+2 & = & \tau\big(\Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{2t-1}\big) & \text{for } t\geqslant 1, \\
\upsilon\big(\widetilde{\Gamma}_{2t+1}\oplus H_2(-1)\big) & = & 2t+1 & = & \upsilon\big(\Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{2t-1}\big) & \text{for } t\geqslant 1, \\
\tau\big(\widetilde{\Gamma}_{2t}\oplus H_2(-1)\big) & = & t+1 & = & \tau\big(\Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{2t-2}\big) & \text{for } t\geqslant 2, \\
\upsilon(\widetilde{\Gamma}_{2t}\oplus H_2(-1)) & = & 2t-1 & = & \upsilon(\Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{2t-2}) & \text{for } t\geqslant 2.
\end{array}$$

so both sides of the transformation in (7) have the same τ and υ. Now let us prove the consistency.

For $k=3$ we have

$$\widetilde{\Gamma}_3\oplus H_2(-1) \overset{P}{\rightsquigarrow} H_2(-1)\oplus\widetilde{\Gamma}_3 \overset{X_3}{\rightsquigarrow} \Gamma_1^{\oplus 3}, \text{ for } P=\left[\begin{smallmatrix} 0 & I_3 \\ I_2 & 0\end{smallmatrix}\right], \text{ and } X_3 = \left[\begin{smallmatrix} 1 & 0 & \mathtt{i} \\ 0 & -\mathtt{i} & 0 \\ 0 & 1 & 0 \\ -\mathtt{i} & 0 & 1 \\ \frac{\mathtt{i}}{2} & 0 & \frac{1}{2}\end{smallmatrix}\right],$$

as can be directly checked. For $k\geqslant 4$ we are going to prove that

$$\widetilde{\Gamma}_k\oplus H_2(-1) \overset{P}{\rightsquigarrow} H_2(-1)\oplus\widetilde{\Gamma}_k \overset{X_k}{\rightsquigarrow} \Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{k-2}, \text{ with } P=\left[\begin{smallmatrix} 0 & I_k \\ I_2 & 0\end{smallmatrix}\right] \text{ and } X_k = \left[\begin{array}{ccc|c} & & & 0 \\ & & & -\frac{\mathtt{i}}{2} \\ & X_3 & & 0 \\ & & & 0 \\ & & & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right]\oplus I_{k-4},$$

where the first transformation is just a block permutation. So for the rest of the proof we will focus on the second transformation. We use the following notation: $A(i:j)$ is the principal submatrix of A containing the rows and columns from the ith to the jth ones.

If $k=4$ then $H_2(-1)\oplus\widetilde{\Gamma}_4 \overset{X_4}{\rightsquigarrow} \Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_2$, for X_4 as above, as can be directly checked.

If $k>4$ then $H_2(-1)\oplus\widetilde{\Gamma}_k \overset{X_k}{\rightsquigarrow} \Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{k-2}$, for $X_k = X_4\oplus I_{k-4}$. To prove it we will use the identities

$$\widetilde{\Gamma}_k = \begin{bmatrix} \widetilde{\Gamma}_4 & E_{4\times(k-4)} \\ E_{4\times(k-4)}^\top & \widetilde{\Gamma}_k(5:k)\end{bmatrix} \quad \text{and} \quad \widetilde{\Gamma}_{k-2} = \begin{bmatrix} \widetilde{\Gamma}_2 & E_{2\times(k-4)} \\ E_{2\times(k-4)}^\top & \widetilde{\Gamma}_{k-2}(3:k-2)\end{bmatrix},$$

so that

$$
\begin{aligned}
(X_4\oplus I_{k-4})^\top\Big(H_2(-1)\oplus\widetilde{\Gamma}_k\Big)(X_4\oplus I_{k-4}) &= \begin{bmatrix} X_4^\top & 0\\ 0 & I_{k-4}\end{bmatrix}\begin{bmatrix} H_2(-1)\oplus\widetilde{\Gamma}_4 & E_{6\times(k-4)}\\ E_{6\times(k-4)}^\top & \widetilde{\Gamma}_k(5:k)\end{bmatrix}\begin{bmatrix} X_4 & 0\\ 0 & I_{k-4}\end{bmatrix}\\
&= \begin{bmatrix} X_4^\top\Big(H_2(-1)\oplus\widetilde{\Gamma}_4\Big)X_4 & X_4^\top E_{6\times(k-4)}\\ E_{6\times(k-4)}^\top X_4 & \widetilde{\Gamma}_k(5:k)\end{bmatrix}\\
&= \begin{bmatrix} \Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_2 & E_{4\times(k-4)}\\ E_{4\times(k-4)}^\top & \widetilde{\Gamma}_{k-2}(3:k-2)\end{bmatrix}\\
&= \Gamma_1^{\oplus 2}\oplus\widetilde{\Gamma}_{k-2}
\end{aligned}
$$

where in the last-but-one equality we use that $\widetilde{\Gamma}_k(5:k)=\widetilde{\Gamma}_{k-2}(3:k-2)$.

□

5.3 The case of Type-II blocks

Finally, Lemma 11 is the counterpart of Lemmas 8 and 10 for Type-II blocks. Again, instead of the blocks $H_{2k}(\mu)$ we use the tridiagonal version, $\widetilde{H}_{2k}(\mu)$. In the statement, $\widetilde{H}_0(\mu)$ stands for an empty block.

Lemma 11. *The following transformations are $(\tau,\upsilon)-$invariant:*

(i) $\widetilde{H}_{2k}(\mu)\oplus H_2(-1)\rightsquigarrow\widetilde{H}_{2k-2}(\mu)\oplus\Gamma_1^{\oplus 2}$, *for* $\mu\neq\pm1$ *and* $k\geqslant 1$.

(ii) $\widetilde{H}_{4k+2}(-1)\oplus H_2(-1)\rightsquigarrow\widetilde{\Gamma}_{2k}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}$, *for* $k\geqslant 1$.

(iii) $\widetilde{H}_{4k}(1)\oplus H_2(-1)\rightsquigarrow\widetilde{\Gamma}_{2k-1}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}$, *for* $k\geqslant 1$.

Proof. In order to see that all transformations in (i)–(iii) are $(\tau,\upsilon)-$invariant, first note that

$$
\begin{array}{lcllll}
\tau\big(\widetilde{H}_{2k}(\mu)\oplus H_2(-1)\big) & = & k+2 & = & \tau\big(\widetilde{H}_{2k-2}(\mu)\oplus\Gamma_1^{\oplus 2}\big), & \text{for } \mu\neq\pm1 \text{ and } k\geqslant 1\\
\upsilon\big(\widetilde{H}_{2k}(\mu)\oplus H_2(-1)\big) & = & 2k & = & \upsilon\big(\widetilde{H}_{2k-2}(\mu)\oplus\Gamma_1^{\oplus 2}\big), & \text{for } \mu\neq\pm1 \text{ and } k\geqslant 1\\
\tau\big(\widetilde{H}_{4k+2}(-1)\oplus H_2(-1)\big) & = & 2k+2 & = & \tau\big(\widetilde{\Gamma}_{2k}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}\big), & \text{for } k\geqslant 1,\\
\upsilon\big(\widetilde{H}_{4k+2}(-1)\oplus H_2(-1)\big) & = & 4k & = & \upsilon\big(\widetilde{\Gamma}_{2k}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}\big), & \text{for } k\geqslant 1,\\
\tau\big(\widetilde{H}_{4k}(1)\oplus H_2(-1)\big) & = & 2k+2 & = & \tau\big(\widetilde{\Gamma}_{2k-1}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}\big), & \text{for } k\geqslant 1,\\
\upsilon\big(\widetilde{H}_{4k}(1)\oplus H_2(-1)\big) & = & 4k & = & \upsilon\big(\widetilde{\Gamma}_{2k-1}^{\oplus 2}\oplus\Gamma_1^{\oplus 2}\big), & \text{for } k\geqslant 1.
\end{array}
$$

Now let us prove the consistence in (i)–(iii). The following identity is used:

$$
\widetilde{H}_{2k}(\mu)=\begin{bmatrix}\widetilde{H}_{2k-2t}(\mu) & E_{(2k-2t)\times 2t}\\ 0 & \widetilde{H}_{2t}(\mu)\end{bmatrix},\qquad \text{for } t<k.
$$

(i) If $k=1$ then $\widetilde{H}_2(\mu)\oplus H_2(-1)\overset{X_1}{\rightsquigarrow}\Gamma_1^{\oplus 2}$, for $X_1=\begin{bmatrix}\frac{1}{1+\mu} & -\frac{\mathrm{i}}{1+\mu}\\ 0 & 1-\mu\\ -\frac{\mathrm{i}}{1+\mu} & 0\end{bmatrix}$.

If $k=2$ then $\widetilde{H}_4(\mu)\oplus H_2(-1)\overset{X_2}{\rightsquigarrow}\widetilde{H}_2(\mu)\oplus\Gamma_1^{\oplus 2}$, for $X_2=\left[\begin{array}{cc|cc} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ \hline 0 & 0 & & \\ 0 & -\frac{1}{1+\mu} & & \\ 0 & -\mathrm{i} & & X_1\\ 0 & -\frac{\mathrm{i}}{-1+\mu^2} & & \end{array}\right]$.

If $k>2$ then $\widetilde{H}_{2k}(\mu)\oplus H_2(-1)\overset{X_k}{\rightsquigarrow}\widetilde{H}_{2k-2}(\mu)\oplus\Gamma_1^{\oplus 2}$, for $X_k=I_{2k-4}\oplus X_2$, since

$$
\begin{aligned}
(I_{2k-4}\oplus X_2)^\top\Big(\widetilde{H}_{2k}(\mu)\oplus H_2(-1)\Big)(I_{2k-4}\oplus X_2) &= \begin{bmatrix} I_{2k-4} & 0\\ 0 & X_2^\top\end{bmatrix}\begin{bmatrix}\widetilde{H}_{2k-4}(\mu) & E_{(2k-4)\times 6}\\ 0 & \widetilde{H}_4(\mu)\oplus H_2(-1)\end{bmatrix}\begin{bmatrix} I_{2k-4} & 0\\ 0 & X_2\end{bmatrix}\\
&= \begin{bmatrix}\widetilde{H}_{2k-4}(\mu) & E_{(2k-4)\times 6}X_2\\ 0 & X_2^\top\Big(\widetilde{H}_4(\mu)\oplus H_2(-1)\Big)X_2\end{bmatrix}\\
&= \begin{bmatrix}\widetilde{H}_{2k-4}(\mu) & E_{(2k-4)\times 4}\\ 0 & \widetilde{H}_2(\mu)\oplus\Gamma_1^{\oplus 2}\end{bmatrix}\\
&= \widetilde{H}_{2k-2}(\mu)\oplus\Gamma_1^{\oplus 2}.
\end{aligned}
$$

(ii) Let us prove, for $k \geqslant 1$, that

$$\widetilde{H}_{4k+2}(-1) \oplus H_2(-1) \overset{X_k}{\rightsquigarrow} \widetilde{H}_{4k}(-1) \oplus \Gamma_1^{\oplus 2}, \text{ for } X_k = I_{4k-2} \oplus C, \text{ with } C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\mathrm{i} \\ 0 & 0 & 1 & \mathrm{i} \\ 0 & 0 & 0 & -\mathrm{i} \\ 0 & -1 & 1 & -\mathrm{i} \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

This is because

$$\begin{aligned}
(I_{4k-2} \oplus C)^\top \Big(\widetilde{H}_{4k+2}(-1) \oplus H_2(-1)\Big)(I_{4k-2} \oplus C) &= \begin{bmatrix} I_{4k-2} & 0 \\ 0 & C^\top \end{bmatrix} \begin{bmatrix} \widetilde{H}_{4k-2}(-1) & E_{(4k-2)\times 6} \\ 0 & \widetilde{H}_4(-1) \oplus H_2(-1) \end{bmatrix} \begin{bmatrix} I_{4k-2} & 0 \\ 0 & C \end{bmatrix} \\
&= \begin{bmatrix} \widetilde{H}_{4k-2}(-1) & E_{(4k-2)\times 6} C \\ 0 & C^\top \left(\widetilde{H}_4(-1) \oplus H_2(-1)\right) C \end{bmatrix} \\
&= \begin{bmatrix} \widetilde{H}_{4k-2}(-1) & E_{(4k-2)\times 4} \\ 0 & \widetilde{H}_2(-1) \oplus \Gamma_1^{\oplus 2} \end{bmatrix} \\
&= \widetilde{H}_{4k}(-1) \oplus \Gamma_1^{\oplus 2}.
\end{aligned}$$

Finally, let us see that $\widetilde{H}_{4k}(-1)$ is congruent to $\widetilde{\Gamma}_{2k}^{\oplus 2}$ or, equivalently, that $H_{4k}(-1)$ is congruent to $\Gamma_{2k}^{\oplus 2}$. In order to do this, we are going to prove that the cosquares of $H_{4k}(-1)$ and $\Gamma_{2k}^{\oplus 2}$ are similar, and this immediately implies that $H_{4k}(-1)$ and $\Gamma_{2k}^{\oplus 2}$ are congruent, by Lemma 3.

The cosquare of $H_{4k}(-1)$ is

$$\begin{aligned}
H_{4k}(-1)^{-\top} H_{4k}(-1) &= \begin{bmatrix} 0 & I_{2k} \\ J_{2k}(-1) & 0 \end{bmatrix}^{-\top} \begin{bmatrix} 0 & I_{2k} \\ J_{2k}(-1) & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & J_{2k}(-1)^{-1} \\ I_{2k} & 0 \end{bmatrix}^{\top} \begin{bmatrix} 0 & I_{2k} \\ J_{2k}(-1) & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & I_{2k} \\ J_{2k}(-1)^{-\top} & 0 \end{bmatrix} \begin{bmatrix} 0 & I_{2k} \\ J_{2k}(-1) & 0 \end{bmatrix} \\
&= \begin{bmatrix} J_{2k}(-1) & 0 \\ 0 & J_{2k}(-1)^{-\top} \end{bmatrix},
\end{aligned}$$

and the cosquare of $\Gamma_{2k}^{\oplus 2}$ is

$$\begin{bmatrix} \Gamma_{2k}^{-\top} & 0 \\ 0 & \Gamma_{2k}^{-\top} \end{bmatrix} \begin{bmatrix} \Gamma_{2k} & 0 \\ 0 & \Gamma_{2k} \end{bmatrix} = \begin{bmatrix} \Gamma_{2k}^{-\top} \Gamma_{2k} & 0 \\ 0 & \Gamma_{2k}^{-\top} \Gamma_{2k} \end{bmatrix},$$

with (see [6, p. 13])

$$\Gamma_{2k}^{-\top} \Gamma_{2k} = \begin{bmatrix} -1 & -2 & & \star \\ & \ddots & \ddots & \\ & & -1 & -2 \\ 0 & & & -1 \end{bmatrix},$$

where $\star$ denotes some entries that are not relevant in our arguments. As $J_{2k}(-1)^{-\top}$ is similar to $J_{2k}(-1)$, the previous identities show that $\big(H_{4k}(-1)\big)^{-\top} H_{4k}(-1)$ and $\left(\Gamma_{2k}^{\oplus 2}\right)^{-\top} \Gamma_{2k}^{\oplus 2}$ are similar, since the Jordan canonical form of both them is $J_{2k}(-1)^{\oplus 2}$.

(iii) Let us prove that, for $k \geqslant 1$:

$$\widetilde{H}_{4k}(1) \oplus H_2(-1) \overset{X_k}{\rightsquigarrow} \widetilde{H}_{4k-2}(1) \oplus \Gamma_1^{\oplus 2}, \quad \text{for } X_k = I_{4k-4} \oplus C, \text{ with } C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \mathrm{i} \\ 0 & -\frac{1}{2} & \frac{1}{2} & -\frac{\mathrm{i}}{2} \\ 0 & 0 & 1 & \mathrm{i} \\ 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}.$$

For $k=1$ the solution matrix is $X_1 = C$, as it can directly checked. Let us now see it for $k \geqslant 2$:

$$\begin{aligned}(I_{4k-4} \oplus C)^\top \left(\widetilde{H}_{4k}(1) \oplus H_2(-1)\right)(I_{4k-4} \oplus C) &= \begin{bmatrix} I_{4k-4} & 0 \\ 0 & C^\top \end{bmatrix} \begin{bmatrix} \widetilde{H}_{4k-4}(1) & E_{(4k-4)\times 6} \\ 0 & \widetilde{H}_4(1) \oplus H_2(-1) \end{bmatrix} \begin{bmatrix} I_{4k-4} & 0 \\ 0 & C \end{bmatrix} \\ &= \begin{bmatrix} \widetilde{H}_{4k-4}(1) & E_{(4k-4)\times 6} C \\ 0 & C^\top \left(\widetilde{H}_4(1) \oplus H_2(-1)\right) C \end{bmatrix} \\ &= \begin{bmatrix} \widetilde{H}_{4k-4}(1) & E_{(4k-4)\times 4} \\ 0 & \widetilde{H}_2(1) \oplus \Gamma_1^{\oplus 2} \end{bmatrix} \\ &= \widetilde{H}_{4k-2}(1) \oplus \Gamma_1^{\oplus 2}.\end{aligned}$$

It remains to see that $\widetilde{H}_{4k-2}(1)$ is congruent to $\widetilde{\Gamma}_{2k-1}^{\oplus 2}$ or, equivalently, that $H_{4k-2}(1)$ is congruent to $\Gamma_{2k-1}^{\oplus 2}$. To prove this, we can proceed as before, by showing that the cosquares of $H_{4k-2}(1)$ and $\Gamma_{2k-1}^{\oplus 2}$ are similar (in this case, their Jordan canonical form is $J_{2k-1}(1)^{\oplus 2}$), and this implies that $H_{4k-2}(1)$ and $\Gamma_{2k-1}^{\oplus 2}$ are congruent, again by Lemma 3.

□

6 The main result

The following result, which is the main result in this work, improves the main result in [4] (namely, Theorem 8 in that reference) by including the case where the CFC of A contains blocks of type $H_2(-1)$, that were excluded in [4, Th. 8].

Theorem 12. *Let A be a complex square matrix whose* CFC *does not have blocks of type $\widetilde{H}_4(1)$, and B a symmetric matrix. Then $X^\top A X = B$ is consistent if and only if* $\operatorname{rank} B \leqslant \min\{\tau(A), \upsilon(A)\}$.

Proof. The necessity of the condition is already stated in Theorem 7. We are going to prove that it is also sufficient.

By the $J_1(0)$-law and the Canonical reduction law, we may assume that both A and B are given in CFC and that neither A nor B have blocks of type $J_1(0)$. This implies, in particular, that $B = I_m$, for some m, and that A is as in Definition 4, with $j_1 = 0$. We also assume that all blocks Γ_k and $H_{2k}(\mu)$ in A, if present, have been replaced by $\widetilde{\Gamma}_k$ and $\widetilde{H}_{2k}(\mu)$, respectively.

Let us recall that $\Gamma_1^{\oplus m} = I_m$. Throughout the proof, we mainly use the first notation, to emphasize that we are dealing with canonical blocks.

If the CFC of A does not contain blocks $H_2(-1)$, then the result is provided in [4, Th. 8]. Otherwise, we are going to see that it is possible, by means of $(\tau, \upsilon)-$invariant transformations, to either "absorb" all blocks $H_2(-1)$ or to end up with a direct sum of blocks $H_2(-1)$, together with, possibly, other blocks, which are quite specific. More precisely, we can end up with a direct sum of blocks satisfying one of the following conditions:

(C0) There are no blocks $H_2(-1)$.

(C1) There are some blocks $H_2(-1)$ together with, possibly, a direct sum of blocks $J_3(0)$, $\widetilde{\Gamma}_2$, and/or Γ_1.

We are first going to see that, indeed, we can arrive to one of the situations described in cases (C0)–(C1). In the procedure, we may need to permute the canonical blocks, in order to use Lemmas 8, 10, and 11. By Theorem 1, this provides a congruent matrix which has, in particular, the same τ and υ, so these permutations do not affect the consistency. Then, we will prove that in both cases (C0) and (C1) the statement holds.

So let us assume that the CFC of A contains a direct sum of blocks $H_2(-1)$, together with some other Type-0, Type-I, and Type-II blocks (except $\widetilde{H}_4(1)$).

Using Lemma 8, for each block $J_k(0)$ (with $k \neq 3$) we can "absorb" a block $H_2(-1)$ by means of a $(\tau, \upsilon)-$invariant transformation, and we end up with a direct sum of a block $J_{k-2}(0)$ together with two blocks Γ_1. We can keep reducing the size of the Type-0 blocks until either all $H_2(-1)$ blocks have been absorbed (so we end up in case (C0)) or there are no more Type-0 blocks, except maybe blocks $J_3(0)$. Now, we can proceed in the same way with Type-I blocks using Lemma 10. Again, we end up either with a direct sum containing no $H_2(-1)$ blocks (case (C0) again) or no Type-I blocks, except maybe blocks Γ_1 and/or $\widetilde{\Gamma}_2$. Next, we do the same with Type-II blocks using Lemma 11. Note that the reductions in parts (ii) and (iii) in the statement of Lemma 11 produce as an output some Type-I blocks $\widetilde{\Gamma}_k$, with $k \geqslant 1$. In the case when $k > 1$, we can use again Lemma 10, provided that there are still blocks $H_2(-1)$. Therefore, after these reductions, either we have absorbed all blocks $H_2(-1)$ (case (C0) again), or there are blocks $H_2(-1)$, together with, possibly, a direct sum of other blocks that cannot absorb them, namely $J_3(0)$, $\widetilde{\Gamma}_2$, and/or Γ_1 (case (C1)).

Now, it remains to prove that in both cases (C0) and (C1) the statement holds, namely that $A \rightsquigarrow \Gamma_1^{\oplus m}$, for any $m \leqslant \min\{\tau(A), \upsilon(A)\}$, in these two cases. Let $\widehat{A}$ be the matrix obtained after applying to A all the transformations explained in the previous paragraph. By the Transitive law, $A \rightsquigarrow \widehat{A}$. Moreover, since all these transformations are $(\tau, \upsilon)-$invariant, then (4) implies that $\tau(A) = \tau(\widehat{A})$ and $\upsilon(A) = \upsilon(\widehat{A})$. Therefore, it is enough to prove that $\widehat{A} \rightsquigarrow \Gamma_1^{\oplus m}$ for any $m \leqslant \min\{\tau(\widehat{A}), \upsilon(\widehat{A})\}$. By the Elimination law, $\Gamma_1^{\oplus a} \rightsquigarrow \Gamma_1^{\oplus b}$ for any $b < a$, so it will be enough to prove that $\widehat{A} \rightsquigarrow \Gamma_1^{\min\{\tau(\widehat{A}), \upsilon(\widehat{A})\}}$.

In case (C0) the statement is true, as a consequence of [4, Th. 8]. More precisely, in this case, $\min\{\tau(A), \upsilon(A)\} = \tau(A)$, as a consequence of Lemma 5. Then, [4, Th. 8] guarantees that $A \rightsquigarrow \Gamma_1^{\oplus \tau(A)}$ (in [4, Th. 8], however, the notation τ was not used).

In case (C1), we may assume that

$$\widehat{A} = H_2(-1)^{\oplus j} \oplus J_3(0)^{\oplus h} \oplus \widetilde{\Gamma}_2^{\oplus k} \oplus \Gamma_1^{\oplus \ell} \quad \text{for some } j, h, k, \ell \geqslant 0.$$

Note that, in this case, $\min\{\tau(\widehat{A}), \upsilon(\widehat{A})\} = \upsilon(\widehat{A})$, since $\tau(\widehat{A}) = j + 2h + k + \ell > \upsilon(\widehat{A}) = 2h + k + \ell$. Hence, it is enough to prove that $A \rightsquigarrow \Gamma_1^{\upsilon(\widehat{A})}$. In order to do this, we consider the transformations

$$H_2(-1)^{\oplus j} \oplus J_3(0)^{\oplus h} \oplus \widetilde{\Gamma}_2^{\oplus k} \oplus \Gamma_1^{\oplus \ell} \rightsquigarrow J_3(0)^{\oplus h} \oplus \widetilde{\Gamma}_2^{\oplus k} \oplus \Gamma_1^{\oplus \ell} \rightsquigarrow \Gamma_1^{\oplus 2h} \oplus \Gamma_1^{\oplus k} \oplus \Gamma_1^{\oplus \ell} = \Gamma_1^{\oplus \upsilon(\widehat{A})},$$

where the first transformation is a consequence of the Elimination law, and the second transformation is a consequence of the Addition law, together with Lemma 9 (for the first addend) and with $\widetilde{\Gamma}_2 \overset{\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]}{\rightsquigarrow} \Gamma_1$ (for the second addend). □

Remark 13. Unfortunately, when the CFC of A contains at least one block $\widetilde{H}_4(1)$, it is no longer true that, for any $m \leqslant \min\{\tau(A), \upsilon(A)\}$, the equation $X^\top A X = I_m$ is consistent. For instance, $X^\top \widetilde{H}_4(1) X = I_3$ is not consistent (see [4, Th. 7]), but $\tau(\widetilde{H}_4(1)) = 4$ and $\upsilon(\widetilde{H}_4(1)) = 3$, so $\min\{\tau(\widetilde{H}_4(1)), \upsilon(\widetilde{H}_4(1))\} = 3$. Therefore, the case where the CFC of A contains blocks $\widetilde{H}_4(1)$ deserves a further analysis.

Related to this, Theorem 12 can be slightly improved, allowing the CFC of A to contain blocks $\widetilde{H}_4(1)$ provided that the number of these blocks is not larger than the number of blocks $H_2(-1)$. In this case, we can start the reduction procedure described in the proof of Theorem 12 by "absorbing" the blocks $\widetilde{H}_4(1)$ with the blocks $H_2(-1)$ as described in Lemma 11-(iii). More precisely, we can gather each block $\widetilde{H}_4(1)$ with a block $H_2(-1)$, and use the $(\tau, \upsilon)-$invariant transformation $\widetilde{H}_4(1) \oplus H_2(-1) \rightsquigarrow \Gamma_1^{\oplus 4}$. Once we have absorbed all blocks $\widetilde{H}_4(1)$ we can continue with the reduction as explained in the proof of Theorem 12.

7 Conclusions and open questions

In this paper, we have obtained a necessary condition for the equation $X^\top A X = B$ to be consistent, with A, B being complex square matrices and B being symmetric. This condition improves the one obtained in [4, Th. 2]. Moreover, we have proved that the condition is sufficient when the CFC of A does not contain blocks $\widetilde{H}_4(1)$. This result also improves the one in [4, Th. 8], where the case in which the CFC has blocks $H_2(-1)$ was excluded.

As a natural continuation of this work it remains to address the case where the CFC of A contains blocks $\widetilde{H}_4(1)$, in order to fully characterize the consistency of $X^\top A X = B$, with B symmetric, for any matrix A. We have seen that the condition mentioned above is no longer sufficient in this case, so a different characterization is needed. So far, we have been unable to find such a characterization.

Acknowledgments: This research has been funded by the *Agencia Estatal de Investigación* of Spain through grants PID2019-106362GB-I00/AEI/10.13039/501100011033 and MTM2017-90682-REDT.

References

[1] P. Benner, B. Iannazzo, B. Meini, D. Palitta. *Palindromic linearization and numerical solution of nonsymmetric algebraic T-Riccati equations.* (2021) arXiv:2110.03254

[2] P. Benner, D. Palitta. *On the solution of the non-symmetric T-Riccati equation.* Electron. Trans. Numer. Anal. 54 (2021) 66–88.

[3] M. Benzi, M. Viviani. *Solving cubic matrix equations arising in conservative dynamics.* (2021) arXiv:2111.12373

[4] A. Borobia, R. Canogar, F. De Terán. On the consistency of the matrix equation $X^\top A X = B$ when B is symmetric. *Mediterr. J. Math.* 18, 40 (2021). https://doi.org/10.1007/s00009-020-01656-7

[5] A. Borobia, R. Canogar, F. De Terán. The equation $X^\top AX = B$ with B skew-symmetric: How much of a bilinear form is skew-symmetric? *Lin. Multilin. Algebra* (submitted). `http://arxiv.org/abs/2203.07100`

[6] F. De Terán. Canonical forms for congruence of matrices: a tribute to H. W. Turnbull and A. C. Aitken. *SeMA J.* 73 (2016) 7-16.

[7] P. G. Buckhiester. Rank r solutions to the matrix equation $XAX^t = C$, A alternate, over GF(2^y). *Trans. Amer. Math. Soc.* 189 (1974) 201–209.

[8] P. G. Buckhiester. Rank r solutions to the matrix equation $XAX^t = C$, A non-alternate, C alternate, over GF(2^y). *Canad. J. Math.* 26 (1974) 78–90.

[9] P. G. Buckhiester. The number of solutions to the matrix equation $XAX' = C$, A and C nonalternate and of full rank, over GF(2^y). *Math. Nachr.* 63 (1974) 37–41.

[10] L. Carlitz. Representations by skew forms in a finite field. *Arch. Math.* V (1954) 19–31.

[11] J. D. Fulton. Generalized inverses of matrices over fields of characteristic two. *Linear Algebra Appl.* 28 (1979) 69–76.

[12] V. Futorny, R. A. Horn, V. V. Sergeichuk. Tridiagonal canonical matrices of bilinear or sesquilinear forms and of pairs of symmetric, skew-symmetric, or Hermitian forms. *J. Algebra* 319 (2008) 2351–2371.

[13] J. H. Hodges. A skew matrix equation over a finite field. *Math. Nachr.* 17 (1966) 49–55.

[14] R. A. Horn, V. V. Sergeichuk. Canonical forms for complex matrix congruence and $*$-congruence. *Linear Algebra Appl.* 416 (2006) 1010–1032.

[15] Kh. D. Ikramov. On the solvability of a certain class of quadratic matrix equations. *Dokl. Math.* 89 (2014) 162–164.

[16] J. H. M. Wedderburn. *The automorphic transformation of a bilinear form.* Ann. of Math. 2, 23 (1921) 122–134.

[17] H. Wei, Y. Zhang. The number of solutions to the alternate matrix equation over a finite field and a q-identity. *J. Stat. Plan. Infer.* 94 (2001) 349–358.