Filomat 33:13 (2019), 4261–4280
https://doi.org/10.2298/FIL1913261D

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Classification and Approximation of Solutions to Sylvester Matrix Equation

Bogdan D. Djordjević[a], Nebojša Č. Dinčić[b]

[a]*Department of Mathematics, Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Republic of Serbia*
[b]*Department of Mathematics, Faculty of Sciences and Mathematics, University of Niš, P.O. Box 224, Niš 18000, Republic of Serbia*

Abstract. In this paper we solve Sylvester matrix equation with infinitely-many solutions and conduct their classification. If the conditions for their existence are not met, we provide a way for their approximation by least-squares minimal-norm method.

1. Introduction and preliminaries

For given vector spaces V_1 and V_2, let $A \in L(V_2)$, $B \in L(V_1)$ and $C \in L(V_1, V_2)$ be linear operators. Equations of the form

$$AX - XB = C \tag{1}$$

with solution $X \in L(V_1, V_2)$ are called Sylvester equations, Sylvester-Rosenblum equations or algebraic Ricatti equations. Such equations have various application in vast fields of mathematics, physics, computer science and engineering (see e.g. [5], [19] and references therein). Fundamental results, established by Sylvester and Rosenblum themselves, are now-days the starting point in solving contemporary problems where these equations occur. These results are

Theorem 1.1. [23] *(Sylvester matrix equation)* Let A, B and C be matrices. Equation $AX - XB = C$ has unique solution X iff $\sigma(A) \cap \sigma(B) = \emptyset$.

Theorem 1.2. [22] *(Rosenblum operator equation)* Let A, B and C be bounded linear operators. Equation $AX - XB = C$ has unique solution X if $\sigma(A) \cap \sigma(B) = \emptyset$.

Equations with unique solutions have been extensively studied so far. There are numerous results regarding this case, some of them theoretical (e. g. Lyapunov stability criteria and spectral operators), which can be found in [2], [5] or [9], and some of them computational (matrix sign function, factorization of matrices and operators, various iterative methods etc.). It should be mentioned that matrix eq. (1) with unique solution X has been solved numerically (among others) in [4], [6], [13], [14], [18] and [21]. The case where

2010 *Mathematics Subject Classification*. Primary 47A62; 15A18; Secondary 65F15
Keywords. Sylvester equation, matrix spectrum, matrix approximations, eigenvalues and eigenvectors, Jordan normal form, Frobenius norm, least-squares solution, minimal-norm solution
Received: 13 February 2019; Accepted: 13 May 2019
Communicated by Dijana Mosić
Research supported by Ministry of Science, Republic of Serbia, Grant No. 174007
Email addresses: bogdan.djordjevic93@gmail.com (Bogdan D. Djordjević), ndincic@hotmail.com (Nebojša Č. Dinčić)

A, B and C are unbounded operators but solution X is unique and bounded has been studied in [16] and [20].

Solvability of eq. (1) in matrices, discarding uniqueness of solution, has been studied in [7] and partially in [18]. Main results in [7] are based on the idea that solutions X can be provided as parametric matrices, where number of parameters at hand depends on dimensions of the corresponding eigenspaces for A and B.

The case where A, B and C are unbounded, with infinitely many unbounded solutions X has been studied in [8]. This particular research paper provides insight on new solutions (called the *weak solutions*), which are only defined on the corresponding eigenspaces for A and B.

This research paper concerns the case when A and B are matrices whose spectra intersect, while matrix C is a rectangular matrix of appropriate dimensions. We obtain sufficient conditions for existence of infinitely-many solutions and provide a way for their classification. If the conditions for their existence are not met, we give a way of approximating particular solutions. This study relies on the eigenspace-analysis conducted in [7] and [8].

We assume V_1 and V_2 to be finite dimensional Hilbert spaces over the same scalar filed $\mathbb{C}$ or $\mathbb{R}$, while $A \in \mathcal{B}(V_2)$, $B \in \mathcal{B}(V_1)$ and $C \in \mathcal{B}(V_1, V_2)$ are assumed to be operators which correspond to the afore-mentioned matrices. Further, $\mathcal{N}(L)$ and $\mathcal{R}(L)$ denote null-space and range of the given operator L. Recall that every finite-dimensional subspace W of a Hilbert space V is closed. Consequently, there exists orthogonal projector from V to W, which will be denoted as P_W.

2. Existence and classification of solutions

Through out this paper, we assume that A and B share s common eigenvalues and denote that set by σ:

$$\{\lambda_1, \dots, \lambda_s\} =: \sigma = \sigma(A) \cap \sigma(B).$$

For more elegant notation, we introduce $E_B^k = \mathcal{N}(B - \lambda_k I)$ and $E_A^k = \mathcal{N}(A - \lambda_k I)$ whenever $\lambda_k \in \sigma$. Different eigenvalues generate mutually orthogonal eigenvectors, so the spaces E_B^k form an orthogonal sum. Put $E_B := \sum_{k=1}^{s} E_B^k$. It is a closed subspace of V_1 and there exists $E_B^\perp$ such that $V_1 = E_B \oplus E_B^\perp$. Take $B = B_E \oplus B_1$ with respect to that decomposition and denote $C_1 = CP_{E_B^\perp}$.

Proposition 2.1. Let V be a Hilbert space and $L \in \mathcal{B}(V)$. If W is $L-$invariant subspace of V, then $W^\perp$ is L^*-invariant subspace of V.

Theorem 2.1. *(Existence of solutions)* For every $k \in \{1, \dots, s\}$, let λ_k, E_A^k and E_B^k be provided as in the previous paragraph. If

$$\mathcal{N}(C_1)^\perp = \mathcal{R}(B_1) \qquad \text{and} \qquad C\left(E_B^k\right) \subset \mathcal{R}(A - \lambda_k I), \tag{2}$$

then there exist infinitely many solutions X to the equation (1).

Proof. For every $1 \le k \le s$, let E_B^k, E_B, $E_B^\perp$, B_E and B_1 be provided as in the previous paragraph. Note that $\mathcal{N}(C_1)^\perp = \mathcal{R}(C_1^*)$, where $C_1^* \in \mathcal{B}(V_2, E_B^\perp)$.

Step 1: solutions on $E_B^\perp$.

We first conduct analysis on $E_B^\perp$. Space E_B is $BP_{E_B^\perp}-$invariant subspace of V_1 and Proposition 2.1 yields $E_B^\perp$ to be $(BP_{E_B^\perp})^*-$ invariant subspace of V_1, so without loss of generality we can observe B_1^* as $B_1^* : E_B^\perp \to E_B^\perp$. Since $\sigma(B_E) = \{\lambda_1, \dots, \lambda_s\}$, it follows that

$$\sigma(B_1^*) \subseteq \{0\} \cup \sigma(B^*) \setminus \{\bar{\lambda}_1, \dots, \bar{\lambda}_s\}.$$

Case 1. Assume that $\sigma(B_1^*) \cap \sigma(A^*) = \emptyset$. Then there exists unique $X_1^* \in \mathcal{B}(V_2, E_B^\perp)$ such that

$$X_1^* A^* - B_1^* X_1^* = C_1^*,$$

that is, there exists unique $X_1 \in \mathcal{B}(E_B^\perp, V_2)$ such that

$$AX_1 - X_1 B_1 = C_1$$

holds.

Case 2. Assume that $\sigma(A^*) \cap \sigma(B_1^*) \neq \emptyset$. It follows that $\sigma(A^*) \cap \sigma(B_1^*) = \{0\}$. But then A^* cannot be nilpotent. Truly, if $\sigma(A^*) = \{0\} = \sigma(A)$, then by assumption, $\sigma(B) \cap \sigma(A) \neq \emptyset$, therefore, $0 \in \sigma(B)$, that is, $0 \in \sigma$. If $u \in \mathcal{N}(B_1)$, it follows that $B_1 u = 0$ and $u \in E_B^\perp$, but then $Bu = B_1 u = 0$, so $u \in \mathcal{N}(B) \subset E_B$, therefore $u \in E_B \cap E_B^\perp = \{0\}$. Hence contradiction, implying that A^* is not nilpotent, but rather has finite ascend, $\operatorname{asc}(A^*) = m \geq 1$, where $\mathcal{N}((A^*)^m)$ is a proper subspace of V_2.

Now observe $B_1^* : E_B^\perp \to E_B^\perp$, which is not invertible by assumption. Take arbitrary $Z_0^* \in \mathcal{B}(\mathcal{N}(A^*), \mathcal{N}(B_1^*))$ operator. Then for every $d \in \mathcal{N}(A^*)$, there exists (by (2)) unique $u \in \mathcal{N}(B_1^*)^\perp$ such that

$$B_1^* u = C_1^* d.$$

Define $X_1^*(Z_0^*)$ on $\mathcal{N}(A^*)$ as $X_1^*(Z_0^*)d := Z_0^* d + u$. Since $\operatorname{asc}(A^*) = m$, the following recursive formula applies.

Assume that $m = 1$. Precisely, decompose $V_2 = \mathcal{N}(A^*) \oplus \mathcal{N}(A^*)^\perp$ and $A^* = 0 \oplus A_1^*$. Then A_1^* is injective from $\mathcal{N}(A^*)^\perp$ to $\mathcal{N}(A^*)^\perp$ and X_1^* can be defined on $\mathcal{N}(A^*)^\perp$ as restriction of X_1^* from Case 1.

Assume that $m > 1$. Then proceed to decompose $\mathcal{N}(A^*)^\perp = \mathcal{N}(A_1^*) \oplus \mathcal{N}(A_1^*)^\perp$ and and define X_1^* on $\mathcal{N}(A_1^*)$ as $X_1^*(N_1^*)u := N_1^* u + d$, where $Z_1^* \in \mathcal{B}(\mathcal{N}(A_1^*), \mathcal{N}(B_1^*))$ is arbitrary operator and

$$B_1^* u = C_1^* d.$$

If A_1^* is injective on $\mathcal{N}(A_1^*)^\perp$, i.e. if $m = 2$, then X_1 can be defined on $\mathcal{N}(A_1^*)^\perp$ as restriction of X_1 from Case 1. If not, then proceed to decompose $\mathcal{N}(A_1^*)^\perp = \mathcal{N}(A_2^*) \oplus \mathcal{N}(A_2^*)^\perp$ and so on. Eventually, one would get to iteration no. m, in a manner that

$$V_2 = \mathcal{N}(A^*) \oplus \mathcal{N}(A_1^*) \oplus \mathcal{N}(A_2^*) \oplus \ldots \oplus \mathcal{N}(A_m^*) \oplus \mathcal{N}(A_m^*)^\perp$$

and $A_m^* : \mathcal{N}(A_m^*)^\perp \to \mathcal{N}(A_m^*)^\perp$ is injective. Then $\sigma(B_1^*) \cap \sigma(A_m^*) = \emptyset$, ergo define X_1^* on $\mathcal{N}(A_m^*)^\perp$ as restriction of X_1^* from Case 1 to $\mathcal{N}(A_m^*)^\perp$. Further, for $0 \leq n \leq m$, let $Z_n^* \in \mathcal{B}(\mathcal{N}(A_n^*), \mathcal{N}(B_1^*))$ be arbitrary operators. Then define X_1^* on $\mathcal{N}(A_n^*)$ as

$$X_1^*(Z_n^*)d := Z_n^* d + u,$$

where once again $u \in \mathcal{N}(B_1^*)^\perp$ is unique element such that $B_1^* u = C_1^* d$. Equivalently, there exists $X_1 \in \mathcal{B}(E_B^\perp, V_2)$ such that

$$AX_1 - X_1 B_1 = C_1,$$

where

$$X_1 = X_1(Z_0^*, Z_1^*, \dots, Z_m^*).$$

Condition $\mathcal{R}(C_1^*) = \mathcal{N}(B_1^*)^\perp = \mathcal{R}(B_1)$ yields X_1 to be well defined on the entire $E_B^\perp$.

Step 2: solutions on E_B.

We now conduct our analysis on E_B. Define $E_A = \sum_{k=1}^{s} E_A^k$ and split V_2 into orthogonal sum $V_2 = E_A \oplus E_A^\perp$. Decompose $A = A_E \oplus A_1$ with respect to that sum. Then A_1 is injective on $E_A^\perp$ and $A_1 v = Av$, for every $v \in E_A^\perp$. For every $k \in \{1, \dots, s\}$ let $N_k \in \mathcal{B}(E_B^k, E_A^k)$ be arbitrary. For every $u \in E_B^k$, by assumption (2), there exists unique $d(u) \in \left(E_A^k\right)^\perp$ such that

$$(A - \lambda_k I)d(u) = Cu.$$

Define

$$X_E^k : u \mapsto N_k u + d(u), \qquad u \in E_B^k.$$

Then $X_E^k : E_B^k \to E_A^k \oplus \left(P_{E_A^k{}^\perp}(A_1 - \lambda_k I)^{-1} C E_B^k\right)$ defines a linear map. What is left is to check whether $X_E := \sum_{k=1}^{s} X_E^k$ is a solution to the equation

$$AX_E - X_E B_E = CP_{E_B}$$

restricted to E_B. However, this is directly verifiable. For any $u \in E_B$ there exist unique $\alpha_1, \dots, \alpha_s \in \mathbb{C}(\mathbb{R})$ and unique $u_k \in E_B^k$, $1 \le k \le s$, such that $u = \sum \alpha_k u_k$. Then

$$(AX_E - X_E B)u = A\sum_{k=1}^{s} \alpha_k X_E^k u_k - \sum_{k=1}^{s} \lambda_k \alpha_k X_E^k u_k = \sum_{k=1}^{s} (\alpha_k (A - \lambda_k I))\,(N_k u_k + d(u_k)) = \sum_{k=1}^{s} \alpha_k C u_k = Cu.$$

It follows that

$$X = \begin{pmatrix} X_E & 0 \\ 0 & X_1 \end{pmatrix}. \tag{3}$$

is a solution to eq. (1). □

Remark. Notice that in proof of Theorem 2.1 Case 2. only emerges if $\sigma(A) \cap \sigma(B_1) = \{0\}$ while $0 \notin \sigma(B)$. Because of special circumstances under which this problem takes place, this situation will be analyzed separately from the standard problem (see Corollary 2.3).

Theorem 2.1 naturally inquires answers to the following questions:

Question 1. Is every solution to the equation (1) of the form (3)?

Question 2. Under which conditions is the solution to (1) unique?

Both of these questions have affirmative answers, which is justified by analysis of the following *eigen-problem associated with given Sylvester equation*:

Assume that $0 \in \sigma = \sigma(A) \cap \sigma(B)$ and let $N_\lambda \in \mathcal{B}(E_B^\lambda, E_A^\lambda)$, for every $\lambda \in \sigma$ be arbitrary. Define $N_\sigma := \oplus_{\lambda \in \sigma} N_\lambda$. Find a solution X to Sylvester equation such that the following eigen-problem is uniquely solved

$$\begin{cases} AX - XB = C \\ Xu_\lambda := P_{(E_A^\lambda)^\perp}(A - \lambda I)^{-1} C u_\lambda + N_\lambda u_\lambda, \qquad u_\lambda \in E_B^\lambda, \qquad \lambda \in \sigma \cup \{0\}. \end{cases} \tag{4}$$

Theorem 2.2. *(Uniqueness of the solution to the eigen-problem)* With respect to the previous notation, assume that $0 \in \sigma$.
1) If the condition (2) holds for every shared eigenvalue $\lambda \in \sigma$, then solution X depends only on the choice of operator N_σ, that is, for fixed N_σ, there exists unique solution X such that (4) holds.
2) Conversely, for every solution X to (1) and for every shared eigenvalue λ for matrices A and B, there

exists unique quotient class $(A-\lambda I)^{-1}C(\mathcal{N}(B-\lambda I))\oplus\mathcal{N}(A-\lambda I)$ such that X is unique solution to the quotient eigen-problem

$$\begin{cases} AX-XB=C \\ X:\mathcal{N}(B-\lambda I)\to(A-\lambda I)^{-1}C(\mathcal{N}(B-\lambda I))\oplus\mathcal{N}(A-\lambda I). \end{cases} \tag{5}$$

Proof. Recall notation from proof of Theorem 2.1.
1) The first statement of the theorem is proved directly. Namely, take $V_1=E_B\oplus E_B^{\perp}$, $B=B_E\oplus B_1$, $V_2=E_A\oplus E_A^{\perp}$, $A=A_E\oplus A_1$ like in Theorem 2.1. Then there exists $X=X_E\oplus X_1$, which is a solution to (1). By construction, since $\sigma(B_1)\cap\sigma(A)=\emptyset$, Case 1. applies and X_1 is uniquely determined in $\mathcal{B}(E^{\perp},V_2)$ while X_E^{λ} is uniquely determined in the class $\mathcal{B}(E_B/E_B^{\lambda},V_2/E_A^{\lambda})$ for every $\lambda\in\sigma$. Varying λ in σ completes the proof.
2) Conversely, let X be a solution to the eq. (1). Let λ be one of the shared eigenvalues for A and B and fix u as a corresponding eigenvector for B. Then $XBu=\lambda Xu$. Hence

$$AXu-XBu=(A-\lambda I)Xu=Cu.$$

Split Xu into the orthogonal sum $Xu=v_1+v_2$, where $v_1\in\mathcal{N}(A-\lambda I)$ and $v_2\in(\mathcal{N}(A-\lambda I))^{\perp}$. Then v_2 is the sought expression $P_{\mathcal{N}(A-\lambda I)^{\perp}}(A-\lambda I)^{-1}Cu$ and $Xu\in v_2+\mathcal{N}(A-\lambda I)$. Condition (2) follows immediately. Repeating the same procedure for every shared eigenvalue for A and B completes the proof. □

Corollary 2.1. (*Number of solutions*) let Σ be the set of all N_σ introduced in the eigen-problem associated with given Sylvester equation (1), that is

$$\Sigma=\Big\{N_\sigma:\quad N_\sigma=\oplus_{\lambda\in\sigma}N_\lambda,\quad N_\lambda\in\mathcal{B}(E_B^{\lambda},E_A^{\lambda}),\quad \lambda\in\sigma(A)\cap\sigma(B)=\sigma\ni\{0\}\Big\}.$$

Let S be the set of all solutions to (1) which satisfy condition (2). Then $|\Sigma|=|S|$.

Proof. For arbitrary $N_\sigma\in\Sigma$, there exits unique $X\in S$ such that (4) holds. Further, for arbitrary $X\in S$ and arbitrary $\lambda\in\sigma$ there exist quotient classes E_A^{λ} and E_B^{λ} such that (5) holds. Define $N_\lambda:E_B^{\lambda}\to E_A^{\lambda}$ to be bounded. Then $N_\sigma=\oplus_{\lambda\in\sigma}N_\lambda$. It follows that $N_\sigma\in\Sigma$. There is one-to-one surjective correspondence $S\leftrightarrow\Sigma$. □

Remark. Due to Corollary 2.1, solution $X(N_\sigma)\in S$, $(N_\sigma\in\Sigma)$, can be referred to as *particular solution*.

Corollary 2.2. (*Size of particular solution*) With the assumptions and notation from Theorem 2.1, Theorem 2.2 and Corollary 2.1, norm of $X(N_\sigma)$ is given as

$$\|X(N_\sigma)\|^2=\|X_E\|^2+\|X_1\|^2=\|N_\sigma\|^2+\sum_{k=1}^{s}\|P_{(E_A^k)^{\perp}}(A-\lambda_k I)^{-1}CP_{E_B^k}\|^2+\|X_1\|^2. \tag{6}$$

Proof. Taking the same decomposition as in Theorem 2.1, let $X=X_E+X_1$. Since X_E annihilates $E_B^{\perp}$ and X_1 annihilates E_B, it follows that

$$\|X\|^2=\|X_E+X_1\|^2=\|X_E\|^2+\|X_1\|^2.$$

By the same argument, taking

$$\|X_E\|^2=\|N_\sigma\|^2+\sum_{k=1}^{s}\|P_{(E_A^k)^{\perp}}(A-\lambda_k I)^{-1}CP_{E_B^k}\|^2$$

completes the proof. □

Corollary 2.3. (*Singularities on $E_B^{\perp}$*) Assume that $0\notin\sigma$ but $0\in\sigma(A)\cap\sigma(B_1)$ and let $\mathrm{dsc}(A)=m\geq 1$. For every $0\leq n\leq m$, define $Z_n\in\mathcal{B}(R(B_1)^{\perp},\mathcal{R}(A^{n+1})^{\perp}\cap\mathcal{R}(A^n))$ and let $Z=\sum_{n=0}^{m}Z_n$. If $\mathcal{N}(C_1)^{\perp}=\mathcal{R}(B_1)$, then there are infinitely many solutions to (1) on $E_B^{\perp}$. Those solutions depend only on choice for Z, that is, if Z is fixed then there exists unique solution $X_1(Z)$ on $E_B^{\perp}$.

Proof. Proof is the same as part 1) in Theorem 2.2. Note that $\mathrm{dsc}(A)=\mathrm{asc}(A^*)=m$ and $\mathcal{R}(A^{n+1})^{\perp}\cap\mathcal{R}(A^n)=\mathcal{N}((A^*)^{n+1})\cap\mathcal{N}((A^*)^n)^{\perp}$. Then proceed to Case 2. of proof of Theorem 2.1. □

3. Fourier approximation minimal norm solution

As illustrated in Theorem 2.2, the system (4) has unique solution provided that all the input parameters are known and satisfy conditions (2). However, if the only input information is $\sigma(A) \cap \sigma(B) = \sigma \neq \emptyset$, the condition (2) is in general not easy or possible to verify. Thus approximation analysis requires more detailed approach.

The easiest assumption is that there exist eigenvalues for A and B

$$\lambda_{k_1}, \dots, \lambda_{k_w} \in \sigma$$

such that

$$C(E_B^\ell) \cap \mathcal{R}(A - \lambda_\ell I) = \emptyset, \quad \ell \in \{k_1, \dots, k_w\}.$$

Ergo any eigenvector u_ℓ of B that corresponds to λ_{k_ℓ} does not obey the condition (2) ($\ell = \overline{k_1, k_w}$), that is, $Cu_\ell \notin \mathcal{R}(A - \lambda_\ell I)$. There exists an orthonormed basis $(e_k)_k$ for $\mathcal{R}(A - \lambda_\ell I)$, such that Cu_ℓ can be approximated by $\overline{C}u_\ell \in \mathcal{R}(A - \lambda_\ell I)$ and this approximation is the best possible, where

$$\overline{C}u_\ell = \sum_k \langle Cu, e_k \rangle e_k.$$

This way, operator $\overline{C}$ is directly defined on $\sum_{\ell=k_1}^{k_w} E_B^\ell \equiv E_B^w$. The space E_B^w is finite-dimensional and therefore has an orthogonal complement in V_1, denoted as $W = E_B^{w\perp}$. Thus the extension of $\overline{C}$ on V_1 is admissible and we define

$$\overline{C} := \overline{C} \oplus CP_W.$$

Now we solve the approximate Sylvester equation $AX - XB = \overline{C}$, and the solutions X (which exist from Theorem 2.1) are *approximate solutions* to the initial eq. $AX - XB = C$. Combining Corollary 2.2, we see that the error of approximation is derived from

$$\sup_{\|u\|=1} \|(AX - XB - C)u\| = \sup_{\|u\|=1} \|(\overline{C} - C)u\|$$

and this approximation is the best possible, for given u_ℓ. However, note that $\overline{C}$ is not uniquely determined, but still depends on the input parameters: the corresponding eigenvectors for B and the choice for bases in the spaces $\mathcal{R}(A - \lambda_\ell I)$. Hence we try to extract one particular $\overline{C}$ which is the best suited for our approximation problem:

Problem 0. Find those (or that one) approximations for C such that solutions have the smallest possible norm

$$\widetilde{C} = \{\overline{C} : \quad AX - XB = \overline{C} \Rightarrow \|X\| \text{ is the smallest possible}\}.$$

This transfers our problem into minimum function problem, which is solvable in terms of numerical analysis.

4. Least-squares Minimum-norm solutions

When it comes to applications of matrix Sylvester equation, Frobenius norm seems to play more important role than the operator sup −norm. Hence we continue our approximation analysis with that norm.

Frobenius norm, sometimes also called Euclidean norm, of a matrix $A \in \mathbb{C}^{m\times n}$ is defined as

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A) \right)^{1/2},$$

where $\sigma_i(A)$ are the singular values of A. Also, $\|A\|_F = tr(AA^*)^{1/2}$ (recall that $A^* = \overline{A}^T$ is conjugate transpose). Recall that Frobenius norm is sub-multiplicative, i.e. $\|AB\|_F \leq \|A\|_F\|B\|_F$ and unitarily invariant i.e. $\|U_1AU_2\|_F = \|A\|_F$ for some unitary U_1, U_2. From the very definition, it also follows that for matrix A partitioned on block-matrices, $A = [A_{ij}]_{p\times q}$, it follows that $\|A\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\|A_{ij}\|_F^2$. We mention that on any finite dimensional space any two norms are equivalent.

Now we state two problems about Sylvester equation $AX - XB = C$ we are dealing with. Recall that $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times n}$ and $C \in \mathbb{C}^{m\times n}$ are given, and $X \in \mathbb{C}^{m\times n}$ is an unknown matrix. Now Problem 0. can be broken down into two separate problems

Problem 1. Find the set $\mathcal{S}$ of all $\widehat{X}$ such that $\|A\widehat{X} - \widehat{X}B - C\|_F$ is the smallest possible, i.e.

$$\min_{X\in\mathbb{C}^{m\times n}} \|AX - XB - C\|_F = \|A\widehat{X} - \widehat{X}B - C\|_F.$$

Problem 2. Among all $\widehat{X}$ find the one with the smallest Frobenius norm, i.e.

$$\min_{\widehat{X}\in\mathcal{S}} \|\widehat{X}\|_F = \|\widehat{X}_0\|_F.$$

Definition 4.1. Matrices $\widehat{X}$ which are solutions for Problem 1 are **least-squares solutions**. Matrices $\widehat{X_0}$ which are solutions for Problem 2 are **minimal-norm least-squares solution**.

Before we continue our analysis, we remark the following facts:

- if Sylvester equation is consistent for given C, then set $\mathcal{S}$ from Problem 1 consists of all solutions of the Sylvester equation, and norm of approximation error is zero. If there is unique solution, then it solves Problem 2 as well. In the case when there are infinitely many solutions, Problem 2 gives those solutions with the smallest norm;
- for homogeneous equation (i.e. $C = 0$), set $\mathcal{S}$ consists of all homogeneous solutions, and solution of Problem 2 is unique, namely $X = 0$.

It is well-known fact that eq. (1) can, by Kronecker product and vectorization operation, be transformed into

$$(I_n \otimes A - B^T \otimes I_m)\, vec(X) = vec(C).$$

The matrix $I_n \otimes A - B^T \otimes I_m$ is often called "nivellateur" in the literature, and the least-squares minimal-norm solution is unique and is given by

$$\widehat{vec(X)} = (I_n \otimes A - B^T \otimes I_m)^{\dagger} vec(C),$$

where $T^{\dagger}$ denotes the unique Moore-Penrose inverse of in general rectangular complex matrix T. For more on the topic of the generalized inverses reader is referred to [3]. Remark that effective calculation of the Moore-Penrose inverse of nivellateur appears to be very difficult. The authors are unaware of such method, but for the group inverse there is a recent paper of Hartwig and Patrício [11].

Our aim is to reduce the problem for original Sylvester equation to the simplest Sylvester equation case, similar to the approach used in [7]. Suppose that matrices $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{n\times n}$ have the following Jordan canonical forms (for some invertible matrices S and T):

$$A = SJ_AS^{-1},\ B = TJ_BT^{-1}.$$

Without loss of generality, we may assume that $\emptyset \neq \sigma(A) \cap \sigma(B) = \{\lambda_1, ..., \lambda_s\}$, hence we excluded the unique solution case, so

$$J_A = diag\{J(\lambda_1; p_{11}, p_{12}, ..., p_{1,k_1}), ..., J(\lambda_s; p_{s1}, p_{s2}, ..., p_{s,k_s})\} \in \mathbb{C}^{m\times m},$$
$$J_B = diag\{J(\lambda_1; q_{11}, q_{12}, ..., q_{1,\ell_1}), ..., J(\lambda_s; q_{s1}, q_{s2}, ..., q_{s,\ell_s})\} \in \mathbb{C}^{n\times n},$$

where $p_{ij},\ j=\overline{1,k_i},\ i=\overline{1,s}$, and $q_{ij},\ j=\overline{1,\ell_i},\ i=\overline{1,s}$, are natural numbers, and k_i and ℓ_i are geometric multiplicities of the eigenvalue $\lambda_i,\ i=\overline{1,s}$, of A and B, respectively. Notation $J(\lambda;t_1,...,t_k)$ stands for

$$J(\lambda;t_1,...,t_k)=diag\{J_{t_1}(\lambda),...,J_{t_k}(\lambda)\}=J_{t_1}(\lambda)\oplus...\oplus J_{t_k}(\lambda),$$

where $J_{t_1}(\lambda)$ is the Jordan block matrix of dimension $t_1\times t_1$ with λ on its main diagonal.

If we put those Jordan forms in the equation, we have (we denoted $Y=S^{-1}XT$ and $D=S^{-1}CT$):

$$\begin{aligned}\|AX-XB-C\|_F^2&=\|SJ_AS^{-1}X-XTJ_BT^{-1}-C\|_F^2=\\&=\|S\left(J_A(S^{-1}XT)-(S^{-1}XT)J_B-(S^{-1}CT)\right)T^{-1}\|_F^2=\\&=\|S(J_AY-YJ_B-D)T^{-1}\|_F^2\leq\\&\leq\|S\|_F^2\|T^{-1}\|_F^2\|J_AY-YJ_B-D\|_F^2=\\&=\alpha^2(S,T)\,\|J_AY-YJ_B-D\|_F^2.\end{aligned}$$

We used the notation $\alpha(S,T)=\|S\|_F\|T^{-1}\|_F$. Remark that if S and T are unitary matrices, then equality is attained in the previous formula.

Because of:

$$\begin{aligned}J_AY-YJ_B-D&=[J(\lambda_i;p_{i1},...,p_{i,k_i})][Y_{ij}]-[Y_{ij}][J(\lambda_j;q_{j1},...,q_{j,\ell_j})]-[D_{ij}]=\\&=[J(\lambda_i;p_{i1},...,p_{i,k_i})Y_{ij}-Y_{ij}J(\lambda_j;q_{j1},...,q_{j,\ell_j})-D_{ij}]_{s\times s}=\\&=[J_{p_{i,u}}(\lambda_i)Y_{uv}^{(ij)}-Y_{uv}^{(ij)}J_{q_{j,v}}(\lambda_j)-D_{uv}^{(ij)}]_{u=\overline{1,k_i},\,v=\overline{1,\ell_j},\,i,j=\overline{1,s}},\end{aligned}$$

we have

$$\begin{aligned}\|J_AY-YJ_B-D\|_F^2&=\sum_{i,j=1}^{s}\|J(\lambda_i;p_{i1},...,p_{i,k_i})Y_{ij}-Y_{ij}J(\lambda_j;q_{j1},...,q_{j,\ell_j})-D_{ij}\|_F^2=\\&=\sum_{i,j=1}^{s}\sum_{u=1}^{k_i}\sum_{v=1}^{\ell_j}\|J_{p_{i,u}}(\lambda_i)Y_{uv}^{(ij)}-Y_{uv}^{(ij)}J_{q_{j,v}}(\lambda_j)-D_{uv}^{(ij)}\|_F^2.\end{aligned}$$

Now, we distinguish two cases:

- if $\lambda_i\neq\lambda_j$, then by Theorem 1.1 the equation $J_{p_{i,u}}(\lambda_i)Y_{uv}^{(ij)}-Y_{uv}^{(ij)}J_{q_{j,v}}(\lambda_j)=D_{uv}^{(ij)}$ has unique solution $\widehat{Y_{uv}^{(ij)}}$, so $\|J_{p_{i,u}}(\lambda_i)\widehat{Y_{uv}^{(ij)}}-\widehat{Y_{uv}^{(ij)}}J_{q_{j,v}}(\lambda_j)-D_{uv}^{(ij)}\|_F^2=0$.
- if $\lambda_i=\lambda_j$, then the equation $J_{p_{i,u}}(\lambda_i)Y_{uv}^{(ij)}-Y_{uv}^{(ij)}J_{q_{j,v}}(\lambda_j)=D_{uv}^{(ij)}$, after translation for λ_i, reduces to $J_{p_{i,u}}(0)Y_{uv}^{(ii)}-Y_{uv}^{(ii)}J_{q_{i,v}}(0)=D_{uv}^{(ii)}$. From Theorem 1.1 we already know that there may be either infinitely many solutions, or no solutions at all.

Therefore,

$$\begin{aligned}\|AX-XB-C\|_F^2&\leq\alpha^2(S,T)\|J_AY-YJ_B-D\|_F^2=\\&=\alpha^2(S,T)\sum_{i=1}^{s}\sum_{u=1}^{k_i}\sum_{v=1}^{\ell_i}\|J_{p_{i,u}}(\lambda_i)Y_{uv}^{(ii)}-Y_{uv}^{(ii)}J_{q_{i,v}}(\lambda_i)-D_{uv}^{(ii)}\|_F^2=\\&=\alpha^2(S,T)\sum_{i=1}^{s}\sum_{u=1}^{k_i}\sum_{v=1}^{\ell_i}\|J_{p_{i,u}}(0)Y_{uv}^{(ii)}-Y_{uv}^{(ii)}J_{q_{i,v}}(0)-D_{uv}^{(ii)}\|_F^2.\end{aligned}$$

Now we can take minimum over all X (equivalently, over all Y, since they are similar matrices):

$$\min_X \|AX - XB - C\|_F^2 \leq \alpha^2(S,T) \min_Y \|J_A Y - Y J_B - D\|_F^2 =$$
$$= \alpha^2(S,T) \sum_{i=1}^{s} \sum_{u=1}^{k_i} \sum_{v=1}^{\ell_i} \min_{Y_{uv}^{(ii)}} \|J_{p_{i,u}}(0) Y_{uv}^{(ii)} - Y_{uv}^{(ii)} J_{q_{i,v}}(0) - D_{uv}^{(ii)}\|_F^2.$$

Therefore, in order to solve the Problem 1, we need to investigate the following simpler versions of original problems:

Problem 1′. Find the set $\mathcal{S}$ of all least-squares solutions $\widehat{X}$, i.e.

$$\min_{X \in \mathbb{C}^{m\times n}} \|J_m(0)X - XJ_n(0) - C\|_F = \|J_m(0)\widehat{X} - \widehat{X}J_n(0) - C\|_F.$$

Problem 2′. Among all $\widehat{X} \in \mathcal{S}$ find the one, $\widehat{X_0}$, with the smallest Frobenius norm, i.e.

$$\min_{\widehat{X} \in \mathcal{S}} \|\widehat{X}\|_F = \|\widehat{X}_0\|_F.$$

We will prove that such $\widehat{X}$ is unique, and give a method for its explicit finding.

5. Least-squares Solutions for the Simplest Case

Let us denote $p = \min\{m, n\}$ for $m, n \in \mathbb{N}$. For given matrix $A \in \mathbb{C}^{m\times n}$, the set

$$d_k(A) := \{a_{ij} : j - i = k\},\ k = \overline{-m+1, n-1},$$

will be called *k−th small diagonal*. For $k = 0$ we have "the" diagonal, i.e. the set $\{a_{ii} : i = \overline{1,p}\}$. When we refer to some small diagonal, we assume that its elements are ordered accordingly to increase of the index i. For example, if we are dealing with 0−th small diagonal, we assume that the set $d_0(A) = \{a_{11}, a_{22}, ..., a_{pp}\}$ is ordered. We will denote by σ_{m+k} sum of all elements along the k−th small diagonal d_k:

$$\sigma_{m+k}(A) = \sum_{a_{ij} \in d_k} a_{ij},\ k = \overline{-m+1, n-1}.$$

Theorem 5.1. Sylvester equation $J_m(0)X - XJ_n(0) = C$ has a least-squares solution $\widehat{X}$ given by

$$\widehat{X} = X_h + X_p + X_c, \tag{7}$$

where X_h denotes the solution of appropriate homogeneous equation:

$$X_h = \begin{cases} \begin{bmatrix} \frac{p_{n-1}(J_n(0))}{0_{(m-n)\times n}} \end{bmatrix}, & m \geq n, \\ \begin{bmatrix} 0_{m\times(n-m)} & q_{m-1}(J_m(0)) \end{bmatrix}, & m \leq n, \end{cases}$$

X_p is an expression given by

$$X_p = \begin{cases} \sum_{k=0}^{n-1} \left(J_m(0)^T\right)^{k+1} C J_n(0)^k, & m \geq n, \\ -\sum_{k=0}^{m-1} J_m(0)^k C \left(J_n(0)^T\right)^{k+1}, & m \leq n, \end{cases}$$

and

$$X_c = \begin{cases} \begin{bmatrix} 0_{(m-n)\times n} \\ W \end{bmatrix}, & m \geq n, \\ \begin{bmatrix} -W^T & 0_{m\times(n-m)} \end{bmatrix}, & m \leq n, \end{cases}$$

where we denoted

$$W = -\begin{bmatrix} 0 & 0 & ... & 0 & 0 \\ \sigma_p/p & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ \sigma_3/3 & 2\sigma_4/4 & ... & 0 & 0 \\ \sigma_2/2 & 2\sigma_3/3 & ... & (p-1)\sigma_p/p & 0 \end{bmatrix}_{p\times p}.$$

The magnitude of the deviation is:

$$\Delta(\widehat{X}; C) = \min_X \Delta(X; C) = \sum_{k=1}^{p} \frac{\sigma_k^2}{k},$$

where σ_k is a sum of elements over the $(m+k)-$th small diagonal from the matrix C.

Proof. In expanded form, matrix expression $R \equiv R(X) = J_m(0)X - XJ_n(0) - C = [r_{ij}]$ is:

$$\begin{bmatrix} x_{21} - c_{11} & x_{22} - x_{11} - c_{12} & ... & x_{2n} - x_{1,n-1} - c_{1n} \\ x_{31} - c_{21} & x_{32} - x_{21} - c_{22} & ... & x_{3n} - x_{2,n-1} - c_{2n} \\ ... & ... & ... & ... \\ x_{m1} - c_{m-1,1} & x_{m2} - x_{m-1,1} - c_{m-1,2} & ... & x_{mn} - x_{m-1,n-1} - c_{m-1,n} \\ -c_{m1} & -x_{m1} - c_{m2} & ... & -x_{m,n-1} - c_{mn} \end{bmatrix}.$$

As we can see, any fixed small diagonal contains some unknowns and parameters, and those unknowns and parameters cannot be found in any other small diagonal. Therefore, we will make summation over all small diagonals, and then of all elements on each od small diagonals. Since all those summations of the elements can have one of possible 3 forms, described by functions M_1, M_2 and M_4 for $m \geq n$ (or M_1, M_3 and M_4 for $m \leq n$) from Proposition 8.1 (see Appendix), it is customary to separate into three sums. So we have:

$$\begin{aligned} \|R(X)\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}^2 = \sum_{k=-m+1}^{n-1}\sum_{r_{ij}\in d_k} r_{ij}^2 = \\ &= r_{m1}^2 + \sum_{k=-m+2}^{-m+p}\sum_{r_{ij}\in d_k} r_{ij}^2 + \sum_{k=-m+p+1}^{n-p}\sum_{r_{ij}\in d_k} r_{ij}^2 + \sum_{k=n-p+1}^{n-1}\sum_{r_{ij}\in d_k} r_{ij}^2 = \\ &= c_{m1}^2 + \sum_{k=-m+2}^{-m+p} M_1(x_{ij}\in d_k; a_{ij}\in d_k) + \sum_{k=-m+p+1}^{n-p} M_{2\vee 3}(x_{ij}\in d_k; a_{ij}\in d_k) + \\ &+ \sum_{k=n-p+1}^{n-1} M_4(x_{ij}\in d_k; a_{ij}\in d_k) \end{aligned}$$

Since all summands are nonnegative, the minimization works independently for each of the small diagonal in some of three sums by using Proposition 8.1. We denote by "hat" the elements on which the minimum

is attained. Therefore:

$$\begin{aligned}\min_X \|R(X)\|_F^2 &= c_{m1}^2 + \sum_{k=-m+2}^{-m+p} \min M_1(x_{ij} \in d_k; a_{ij} \in d_k) + \sum_{k=-m+p+1}^{n-p} \min M_{2\vee 3}(x_{ij} \in d_k; a_{ij} \in d_k) + \\ &+ \sum_{k=n-p+1}^{n-1} \min M_4(x_{ij} \in d_k; a_{ij} \in d_k) = \\ &= c_{m1}^2 + \sum_{k=-m+2}^{-m+p} M_1(\widehat{x_{ij}} \in d_k; a_{ij} \in d_k) + \sum_{k=-m+p+1}^{n-p} M_{2\vee 3}(\widehat{x_{ij}} \in d_k; a_{ij} \in d_k) + \\ &+ \sum_{k=n-p+1}^{n-1} M_4(\widehat{x_{ij}} \in d_k; a_{ij} \in d_k) = \\ &= c_{m1}^2 + \sum_{k=-m+2}^{-m+p} M_1(\widehat{x_{ij}} \in d_k; a_{ij} \in d_k) + 0 + 0 = \\ &= \sum_{k=1}^{p-1} \frac{\sigma_k^2}{k}.\end{aligned}$$

During the minimization process of the functions M_1, M_2 (or M_3) and M_4, we have obtained the following:

- the unique $\widehat{x_{ij}}$ lying on the diagonals $d_k,\ k = \overline{-m+1, -m+p-1}$;
- the unique $\widehat{x_{ij}}$ lying on the diagonals $d_k,\ k = \overline{-m+p, n-p-1}$;
- the elements on each of diagonals $d_k,\ k = \overline{n-p, n-1}$, depend on one real parameter, and we assume this element is from the first row in the case $m \geq n$, and in the last column if $m \leq n$.

If we rearrange the matrix $\widehat{X}$ whose elements are known, we obtain precisely (7). Such rearrangement looks rather cumbersome in general case, and some insight can be brought after looking at Examples 6.1 and 6.2.

Let us prove that any $\widehat{X} = X_h + X_p + X_c$ given by (7) is indeed a least-square solution:

$$\begin{aligned}J_m(0)X - XJ_n(0) - C &= (J_m(0)X_h - X_hJ_n(0)) + \Big(J_m(0)X_p - X_pJ_n(0) - C\Big) + \\ &+ J_m(0)X_c - X_cJ_n(0) = \\ &= \Big(J_m(0)X_p - X_pJ_n(0) - C\Big) + J_m(0)X_c - X_cJ_n(0)\end{aligned}$$

It is not hard to check that $J_m(0)X_p - X_pJ_n(0) - C$ is a matrix whose all entries are zero, except the $m-$th row (case $m \geq n$) or the first column (case $m \leq n$):

$$J_m(0)X_p - X_pJ_n(0) - C = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ -\sigma_1 & -\sigma_2 & \dots & -\sigma_n \end{bmatrix} \text{ or } \begin{bmatrix} -\sigma_n & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -\sigma_2 & 0 & \dots & 0 \\ -\sigma_1 & 0 & \dots & 0 \end{bmatrix}.$$

On the similar way it can be shown that for $m \geq n$

$$J_m(0)X_c - X_cJ_n(0) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ -\sigma_n/n & 0 & \dots & \dots & \dots & \dots \\ -\sigma_{n-1}/(n-1) & -\sigma_n/n & \dots & \dots & \dots & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -\sigma_3/3 & -\sigma_4/4 & \dots & \dots & \dots & \\ -\sigma_2/2 & -\sigma_3/3 & -\sigma_4/4 & \dots & -\sigma_n/n & 0 \\ 0 & \sigma_2/2 & 2\sigma_3/3 & & (n-2)\sigma_{n-1}/(n-1) & (n-1)\sigma_n/n \end{bmatrix},$$

(and just transposed matrix if $m \leq n$), so we conclude that

$$J_m(0)X - XJ_n(0) - C = -\left[\begin{array}{c|cccc} & 0_{(m-n)\times n} & & & \\ \hline \sigma_n/n & 0 & \dots & 0 & 0 \\ \sigma_{n-1}/(n-1) & \sigma_n/n & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_2/2 & \sigma_3/3 & \dots & \sigma_n/n & 0 \\ \sigma_1 & \sigma_2/2 & \dots & \sigma_{n-1}/(n-1) & \sigma_n/n \end{array}\right] = -\left[\begin{array}{c} 0_{(m-n)\times n} \\ \hline q_{p-1}(J_p(0))^T \end{array}\right]$$

if $m \geq n$, and transpose of this matrix if $m \leq n$. The Frobenius norm of this matrix is

$$\|J_m(0)X - XJ_n(0) - C\|_F^2 = \sum_{k=1}^{n} k\left(\frac{\sigma_k}{k}\right)^2 = \sum_{k=1}^{n} \frac{\sigma_k^2}{k},$$

so we have the proof. □

Corollary 5.1. Sylvester equation $J_m(0)X - XJ_n(0) = C$ is consistent if and only if $\sigma_k = 0,\ k = \overline{1,p}$, i.e. iff $X_c = 0$, and its general solution is given by (7).

Remark that this result agrees with Theorems 2.3 and 2.5 from [7].

Corollary 5.2. The least-squares solution for the equation $J_n(0)X - XJ_n(0) = I_n$ is

$$\widehat{X} = p_{n-1}(J_n(0)),$$

for any complex polynomial p_{n-1}, and $\|J_n(0)\widehat{X} - \widehat{X}J_n(0) - I_n\|_F^2 = n$.

6. Minimal-norm Least-squares Solution for the Simplest Case

In the previous section, we observed that there is a whole class of the least-squares solutions, depending on some parameters, whether or not the equation is consistent. Now we prove that those free parameters can be chosen such that the solution is with minimal Frobenius norm. If we look at the structure of matrices X_h and X_c from Theorem 5.1, we see that they do not have any non-zero entry on the same position (for $m \geq n$ matrix X_h is upper triangular, while X_c is strictly lower triangular; the case $m \leq n$ is just transposed situation), therefore only matrix X_p may have the influence on the minimization. So far, we concluded that

$$\min_{\widehat{X}\in\mathcal{S}} \|\widehat{X}\|_F^2 = \min_{X_h} \|X_h + X_p\|_F^2 + \|X_c\|_F^2.$$

Remark that $\|X_c\|_F$ is a constant, which is zero iff the equation is consistent and strictly positive otherwise.

Theorem 6.1. There is a unique least-squares minimal-norm solution for the Sylvester equation $J_m(0)X - XJ_n(0) = C$, given by

$$\widehat{X_0} = \widetilde{X_h} + X_p + X_c, \tag{8}$$

where

$$\widetilde{X_h} = \begin{cases} \left[\begin{array}{c} p_{n-1}^*(J_n(0)) \\ \hline 0_{(m-n)\times n} \end{array}\right], & m \geq n, \\ \left[\begin{array}{cc} 0_{m\times(n-m)} & q_{m-1}^*(J_m(0)) \end{array}\right], & m \leq n, \end{cases}$$

and p_{n-1}^* and q_{m-1}^* are uniquely determined polynomials.

Proof. According to the Theorem 5.1, such $\widehat{X_0}$ is a least-squares solution. We must show that it has minimal norm among all least-squares solutions.

Let us consider the matrix $T = X_h + X_p = [t_{ij}]_{m\times n}$. The only entries that can be minimized include $x_{11}, ..., x_{1n}$ (when $m \geq n$) or $x_{1m}, ..., x_{nm}$ (when $m \leq n$), and they are precisely along the small diagonals from $n-p$ to $n-1$:

$$\begin{aligned}
\|T\|_F^2 &= \sum_{k=-m+1}^{n-1}\sum_{t_{ij}\in d_k} t_{ij}^2 = \sum_{k=-m+1}^{n-p-1}\sum_{t_{ij}\in d_k} t_{ij}^2 + \sum_{k=n-p}^{n-1}\sum_{t_{ij}\in d_k} t_{ij}^2 = \\
&= \sum_{k=-m+1}^{n-p-1}\sum_{t_{ij}\in d_k} t_{ij}^2 + \sum_{k=-m+1}^{n-1} G(x_{ij}\in d_k; c_{ij}\in d_k), \\
\|\widehat{X_0}\| &= \min_{\widehat{X}\in\mathcal{S}} \|\widehat{X}\|_F^2 = \min_{X_h} \|X_h + X_p\|_F^2 + \|X_c\|_F^2 = \\
&= \sum_{k=-m+1}^{n-p-1}\sum_{t_{ij}\in d_k} t_{ij}^2 + \sum_{k=-m+1}^{n-1} \min G(x_{ij}\in d_k; c_{ij}\in d_k) + \|X_c\|_F^2 = \\
&= \sum_{k=-m+1}^{n-p-1}\sum_{t_{ij}\in d_k} t_{ij}^2 + \sum_{k=-m+1}^{n-1} G(x_{ij}^*\in d_k; c_{ij}\in d_k) + \|X_c\|_F^2.
\end{aligned}$$

Therefore, for such $x_{1i}^*,\ i=\overline{1,n}$ (when $m\geq n$), or $x_{im}^*,\ i=\overline{1,m}$ (when $m\leq n$), which is given by Proposition 8.2, we obtained the unique least-squares minimum-norm solution. Since it is already known that homogeneous solution is block-polynomial matrix, we have the result. □

Corollary 6.1. There is unique minimal-norm solution for consistent Sylvester equation, and it is given by $\widehat{X_0} = \widetilde{X_h} + X_p$, with notations as in the previous Theorem.

In order to clarify and illustrate the constructions given in previous two theorems about least-squares and minimal-norm solutions, we present the following two in-detailed examples. The first is dealing with case when $m \geq n$, and the second one with the case $m \leq n$.

Example 6.1. Let us find the minimum-norm least-squares solution of the equation $J_4(0)X - XJ_3(0) = C,\ C \in \mathbb{R}^{4\times 3}$, i.e. we want to find all $X \in \mathbb{R}^{4\times 3}$ such that $\|\Delta(X)\|_F^2 = \|J_4(0)X - XJ_3(0) - C\|_F^2$ is minimal, and among them such X which is minimal-Frobenius-norm matrix.

We can arrange the small diagonals of matrix C as follows (below the matrix there are the indices for small diagonals):

$$\begin{array}{c}
\left[\begin{array}{ccc|c|cc}
 & & c_{21} & c_{11} & c_{12} & c_{13} \\
 & c_{31} & c_{32} & c_{22} & c_{23} & \\
c_{41} & c_{42} & c_{43} & c_{33} & &
\end{array}\right], \\
\begin{array}{cccccc} -3 & -2 & -1 & 0 & 1 & 2 \end{array}
\end{array}$$

and let us denote by $\sigma_k,\ k=\overline{1,3}$, sum over the $(4+k)-$th small diagonal ($k=\overline{-3,2}$) of the matrix C, i.e.

$$\sigma_1 = c_{41},\ \sigma_2 = c_{31}+c_{42},\ \sigma_3 = c_{21}+c_{32}+c_{43}.$$

We can arrange the small diagonal for the matrix $J_4(0)X - XJ_3(0) - C$ as well:

$$\left[\begin{array}{ccc|c|cc}
 & & x_{31}-c_{21} & x_{21}-c_{11} & x_{22}-x_{11}-c_{12} & x_{23}-x_{12}-c_{13} \\
 & x_{41}-c_{31} & x_{42}-x_{31}-c_{32} & x_{32}-x_{21}-c_{22} & x_{33}-x_{22}-c_{23} & \\
-c_{41} & -x_{41}-c_{42} & -x_{42}-c_{43} & x_{43}-x_{32}-c_{33} & &
\end{array}\right]$$

Vertical bars partition the small diagonals according to the entries they have. By using the properties of the Frobenius norm, it is clear that $\|J_4(0)X - XJ_3(0) - C\|_F^2$ can be obtained as a sum of squared entries over each of the columns. For left submatrix, those sums are precisely the function M_1 applied to all its columns except the first one (we enlist unknown variables x_{ij} and parameters c_{ij} from top to the bottom); for central submatrix we have function M_2, while for right submatrix function M_4 is applied for each of its columns. Hence, we have

$$\begin{aligned}
H(x_{11},...,x_{43}) &= (-c_{41})^2 + \Big((x_{41}-c_{31})^2 + (-x_{41}-c_{42})^2\Big) + \\
&+ \Big((x_{31}-c_{21})^2 + (x_{42}-x_{31}-c_{32})^2 + (-x_{42}-c_{43})^2\Big) + \\
&+ \Big((x_{21}-c_{11})^2 + (x_{32}-x_{21}-c_{22})^2 + (x_{43}-x_{32}-c_{33})^2\Big) + \\
&+ \Big((x_{22}-x_{11}-c_{12})^2 + (x_{33}-x_{22}-c_{23})^2\Big) + (x_{23}-x_{12}-c_{13})^2 = \\
&= c_{41}^2 + M_1(x_{41};c_{31},c_{42}) + M_1(x_{31},x_{42};c_{21},c_{32},c_{43}) + \\
&+ M_2(x_{21},x_{32},x_{43};c_{11},c_{22},c_{33}) + \\
&+ M_4(x_{11},x_{22},x_{33};c_{12},c_{23}) + M_4(x_{12},x_{23};c_{13}).
\end{aligned}$$

By the Theorem 8.1, we have unique $\widehat{x}_{21},\widehat{x}_{31},\widehat{x}_{32},\widehat{x}_{41},\widehat{x}_{42},\widehat{x}_{43}$, given by:

$$\begin{aligned}
&\widehat{x}_{41} = \frac{1}{2}(c_{31}-c_{42}),\\
&\widehat{x}_{31} = \frac{1}{3}(2c_{21}-c_{32}-c_{43}),\ \widehat{x}_{42} = \frac{1}{3}(c_{21}+c_{32}-2c_{43}),\\
&\widehat{x}_{21} = c_{11},\ \widehat{x}_{32} = c_{11}+c_{22},\ \widehat{x}_{43} = c_{11}+c_{22}+c_{33},
\end{aligned}$$

such that $M_1(\widehat{x}_{41};c_{31},c_{42}) = \sigma_2^2/2$ and $M_1(\widehat{x}_{31},\widehat{x}_{42};c_{21},c_{32},c_{43}) = \sigma_3^2/3$, while $\widehat{x}_{11},\widehat{x}_{12},\widehat{x}_{13},\widehat{x}_{22},\widehat{x}_{23},\widehat{x}_{33}$ are depending of parameters denoted by x_{11},x_{12},x_{13}:

$$\begin{aligned}
&\widehat{x}_{11} = x_{11},\ \widehat{x}_{22} = x_{11}+c_{12},\ \widehat{x}_{33} = x_{11}+c_{12}+c_{23},\\
&\widehat{x}_{12} = x_{12},\ \widehat{x}_{23} = x_{12}+c_{13},\\
&\widehat{x}_{13} = x_{13},
\end{aligned}$$

for them minima of all M_2 and M_4 are zeros. Therefore, the minimum of the function H is:

$$H(\widehat{x}_{11},...,\widehat{x}_{43}) = \sigma_1^2 + \frac{\sigma_2^2}{2} + \frac{\sigma_3^2}{3} = \sum_{k=1}^{3}\frac{\sigma_k^2}{k}.$$

It is clear that $H(\widehat{x}_{11},...,\widehat{x}_{43}) = 0 \Leftrightarrow \sigma_k = 0,\ k = \overline{1,3}$, which is precisely the consistency condition (Theorem 2.3 from [7]). If we decompose such $\widehat{X}$ as:

$$\begin{aligned}
\widehat{X} &= \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ c_{11} & x_{11}+c_{12} & x_{12}+c_{13} \\ \frac{1}{3}(2c_{21}-c_{32}-c_{43}) & c_{11}+c_{22} & x_{11}+c_{12}+c_{23} \\ \frac{1}{2}(c_{31}-c_{42}) & \frac{1}{3}(c_{21}+c_{32}-2c_{43}) & c_{11}+c_{22}+c_{33} \end{bmatrix} = \\
&= \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ 0 & x_{11} & x_{12} \\ 0 & 0 & x_{11} \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ c_{11} & c_{12} & c_{13} \\ c_{21} & c_{11}+c_{22} & c_{12}+c_{23} \\ c_{31} & c_{21}+c_{32} & c_{11}+c_{22}+c_{33} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\frac{\sigma_3}{3} & 0 & 0 \\ -\frac{\sigma_2}{2} & -\frac{2\sigma_3}{3} & 0 \end{bmatrix} = \\
&= X_h + X_p + X_c,
\end{aligned}$$

where X_h, X_p and X_c are as in the Theorem 7. Also, remark that:

$$J_4(0)\widehat{X} - \widehat{X}J_3(0) - C = \begin{bmatrix} 0 & 0 & 0 \\ -\frac{\sigma_3}{3} & 0 & 0 \\ -\frac{\sigma_2}{2} & -\frac{\sigma_3}{3} & 0 \\ -\sigma_1 & -\frac{\sigma_2}{2} & -\frac{\sigma_3}{3} \end{bmatrix} = -\begin{bmatrix} 0_{1\times 3} \\ \hline p_2(J_3(0))^T \end{bmatrix},$$

where $p_2(t) = \sigma_3/3 + \sigma_2/2 \cdot t + \sigma_1 \cdot t^2$.

In order to obtain minimum-norm solution among all least-squares solutions, let us arrange small diagonals of $\widehat{X}$ as follows (vertical bars separate those small diagonals which take parts in minimization from the others):

$$\left[\begin{array}{cc|c|cc} & c_{21} - \frac{\sigma_3}{3} & c_{11} & x_{11} & x_{12} \quad x_{13} \\ c_{31} - \frac{\sigma_2}{2} & c_{21} + c_{32} - \frac{2\sigma_3}{3} & c_{11} + c_{22} & x_{11} + c_{12} & x_{12} + c_{13} \\ & & c_{11} + c_{22} + c_{33} & x_{11} + c_{12} + c_{23} & \end{array}\right].$$

There are three parameters, x_{11}, x_{12} and x_{13}, which can be used for the minimization of sums over the columns of right submatrix. In order to obtain such solution, we should minimize the following function (by using the Theorem 8.2) $h : \mathbb{R}^3 \to \mathbb{R}$:

$$\begin{aligned} h(x_{11}, x_{12}, x_{13}) &= x_{11}^2 + (x_{11} + c_{12})^2 + (x_{11} + c_{12} + c_{23})^2 + x_{12}^2 + (x_{12} + c_{13})^2 + x_{13}^2 = \\ &= F(x_{11}; c_{12}, c_{12} + c_{23}) + F(x_{21}; c_{13}) + x_{13}^2. \end{aligned}$$

According to Theorem 8.2, the minimum is obtained for

$$x_{11}^* = -\frac{2c_{12} + c_{23}}{3},\ x_{12}^* = -\frac{c_{13}}{2},\ x_{13}^* = 0,$$

and it is $h(x_{11}^*, x_{12}^*, x_{13}^*) = (3c_{13}^2 + 4(c_{12}^2 + c_{12}c_{23} + c_{23}^2))/6$.

Therefore, the least-squares minimal-norm solution is:

$$\widehat{X_0} = \begin{bmatrix} -(2c_{12} + c_{23})/3 & -c_{13}/2 & 0 \\ c_{11} & (c_{12} - c_{23})/3 & c_{13}/2 \\ (2c_{21} - c_{32} - c_{43})/3 & c_{11} + c_{22} & (c_{12} + 2c_{23})/3 \\ (c_{31} - c_{42})/2 & (c_{21} + c_{32} - 2c_{43})/3 & c_{11} + c_{22} + c_{33} \end{bmatrix}.$$

Note that the solution depends on all entries of the matrix C except c_{41}.

Example 6.2. Let us find the minimum-norm least-squares solution of the equation $J_3(0)X - XJ_4(0) = C,\ C \in \mathbb{R}^{3\times 4}$, i.e. we want to find all $X \in \mathbb{R}^{3\times 4}$ such that $\|\Delta(X)\|_F^2 = \|J_3(0)X - XJ_4(0) - C\|_F^2$ is minimal, and among them such X which is minimal-Frobenius-norm matrix.

We can arrange the small diagonals of matrix C as follows (below the matrix there are the indices for small diagonals):

$$\begin{array}{c} \left[\begin{array}{ccc|c|cc} & & c_{11} & c_{12} & c_{13} & c_{14} \\ & c_{21} & c_{22} & c_{23} & c_{24} & \\ c_{31} & c_{32} & c_{33} & c_{34} & & \end{array}\right], \\ \begin{array}{cccccc} -2 & -1 & 0 & 1 & 2 & 3 \end{array} \end{array}$$

and let us denote by $\sigma_k,\ k = \overline{1,3}$, sum over the $(3+k)-$th small diagonal ($k = \overline{-2,3}$) of the matrix C, i.e.

$$\sigma_1 = c_{31},\ \sigma_2 = c_{21} + c_{32},\ \sigma_3 = c_{11} + c_{22} + c_{33}.$$

We can arrange the small diagonal for the matrix $J_3(0)X - XJ_4(0) - C$ as well:

$$\left[\begin{array}{ccc|c|ccc} & & x_{21}-c_{11} & x_{22}-x_{11}-c_{12} & x_{23}-x_{12}-c_{13} & x_{24}-x_{13}-c_{14} \\ & x_{31}-c_{21} & x_{32}-x_{21}-c_{22} & x_{33}-x_{22}-c_{23} & x_{34}-x_{23}-c_{24} & \\ -c_{31} & -x_{31}-c_{32} & -x_{32}-c_{33} & -x_{33}-c_{34} & & \end{array}\right]$$

Vertical bars partition the small diagonals according to the entries they have. By using the properties of the Frobenius norm, it is clear that $\|J_3(0)X - XJ_4(0) - C\|_F^2$ can be obtained as a sum of squared entries over each of the columns. For left submatrix, those sums are precisely the function M_1 applied to all its columns except the first one (we enlist unknown variables x_{ij} and parameters c_{ij} from top to the bottom); for central submatrix we have function M_3, while for right submatrix function M_4 is applied for each of its columns. Hence, we have

$$\begin{aligned} H(x_{11},\ldots,x_{34}) &= (-c_{41})^2 + \Big((x_{31}-c_{21})^2 + (-x_{31}-c_{32})^2\Big) + \\ &+ \Big((x_{21}-c_{11})^2 + (x_{32}-x_{21}-c_{22})^2 + (-x_{32}-c_{33})^2\Big) + \\ &+ \Big((x_{22}-x_{11}-c_{12})^2 + (x_{33}-x_{22}-c_{23})^2 + (-x_{33}-c_{34})^2\Big) + \\ &+ \Big((x_{23}-x_{12}-c_{13})^2 + (x_{34}-x_{23}-c_{24})^2\Big) + (x_{24}-x_{13}-c_{14})^2 = \\ &= c_{41}^2 + M_1(x_{31};c_{21},c_{32}) + M_1(x_{21},x_{32};c_{11},c_{22},c_{33}) + \\ &+ M_3(x_{11},x_{22},x_{33};c_{12},c_{23},c_{34}) + \\ &+ M_4(x_{12},x_{23},x_{34};c_{13},c_{24}) + M_4(x_{13},x_{24};c_{14}). \end{aligned}$$

By Theorem 8.1, we have unique $\widehat{x}_{11},\widehat{x}_{21},\widehat{x}_{22},\widehat{x}_{31},\widehat{x}_{32},\widehat{x}_{33}$, given by:

$$\begin{aligned} &\widehat{x}_{31} = \frac{1}{2}(c_{21}-c_{32}), \\ &\widehat{x}_{21} = \frac{1}{3}(2c_{11}-c_{22}-c_{33}),\ \widehat{x}_{32} = \frac{1}{3}(c_{11}+c_{22}-2c_{33}), \\ &\widehat{x}_{11} = -(c_{12}+c_{23}+c_{34}),\ x_{22} = -(c_{23}+c_{34}),\ \widehat{x}_{33} = -c_{34}, \end{aligned}$$

such that $M_1(\widehat{x}_{31};c_{21},c_{32}) = \sigma_2^2/2$ and $M_1(\widehat{x}_{21},\widehat{x}_{32};c_{11},c_{22},c_{33}) = \sigma_3^2/3$, while $\widehat{x}_{12},\widehat{x}_{13},\widehat{x}_{14},\widehat{x}_{23},\widehat{x}_{24},\widehat{x}_{34}$ are depending of parameters denoted by x_{14},x_{24},x_{34} (this is important difference from the previous example!):

$$\begin{aligned} &\widehat{x}_{34} = x_{34},\ \widehat{x}_{23} = x_{34}-c_{24},\ \widehat{x}_{12} = x_{34}-c_{13}-c_{24}, \\ &\widehat{x}_{24} = x_{24},\ \widehat{x}_{13} = x_{24}-c_{14}, \\ &\widehat{x}_{14} = x_{14}, \end{aligned}$$

for them minima of all M_3 and M_4 are zeros. Therefore, the minimum of the function H is:

$$H(\widehat{x}_{11},\ldots,\widehat{x}_{34}) = \sigma_1^2 + \frac{\sigma_2^2}{2} + \frac{\sigma_3^2}{3} = \sum_{k=1}^{3}\frac{\sigma_k^2}{k}.$$

It is clear that $H(\widehat{x}_{11},\ldots,\widehat{x}_{34}) = 0 \Leftrightarrow \sigma_k = 0,\ k = \overline{1,3}$, which is precisely the consistency condition (Theorem 2.5 from [7]). If we decompose such $\widehat{X}$ as:

$$\begin{aligned} \widehat{X} &= \begin{bmatrix} -(c_{12}+c_{23}+c_{34}) & x_{34}-c_{13}-c_{24} & x_{24}-c_{14} & x_{14} \\ \frac{1}{3}(2c_{11}-c_{22}-c_{33}) & -(c_{23}+c_{34}) & x_{34}-c_{24} & x_{24} \\ \frac{1}{2}(c_{21}-c_{32}) & \frac{1}{3}(c_{11}+c_{22}-2c_{33}) & -c_{34} & x_{34} \end{bmatrix} = \\ &= \begin{bmatrix} 0 & x_{34} & x_{24} & x_{14} \\ 0 & 0 & x_{34} & x_{24} \\ 0 & 0 & 0 & x_{34} \end{bmatrix} + \begin{bmatrix} -(c_{12}+c_{23}+c_{34}) & -(c_{13}+c_{24}) & -c_{14} & 0 \\ -(c_{22}+c_{33}) & -(c_{23}+c_{34}) & -c_{24} & 0 \\ -c_{32} & -c_{33} & -c_{34} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ \frac{2\sigma_3}{3} & 0 & 0 & 0 \\ \frac{\sigma_2^2}{2} & \frac{\sigma_3}{3} & 0 & 0 \end{bmatrix} = \\ &= X_h + X_p + X_c, \end{aligned}$$

where X_h, X_p and X_c are as in the Theorem 7. Also, remark that:

$$J_3(0)\widehat{X} - \widehat{X}J_4(0) - C = \begin{bmatrix} -\frac{\sigma_3}{3} & 0 & 0 & 0 \\ -\frac{\sigma_2}{2} & -\frac{\sigma_3}{3} & 0 & 0 \\ -\sigma_1 & -\frac{\sigma_2}{2} & -\frac{\sigma_3}{3} & 0 \end{bmatrix} = -\left[\begin{array}{c|c} p_2(J_3(0))^T & 0_{3\times 1} \end{array}\right],$$

where $p_2(t) = \sigma_3/3 + \sigma_2/2 \cdot t + \sigma_1 \cdot t^2$.

In order to obtain minimum-norm solution among all least-squares solutions, let us arrange small diagonals of $\widehat{X}$ as follows (vertical bars separate those small diagonals which take parts in minimization from the others):

$$\left[\begin{array}{ccc|c|ccc} & & & -(c_{12}+c_{23}+c_{34}) & x_{34}-c_{13}-c_{24} & x_{24}-c_{14} & x_{14} \\ & \frac{2\sigma_3}{3}-c_{22}-c_{33} & & -(c_{23}+c_{34}) & x_{34}-c_{24} & x_{24} & \\ \frac{\sigma_2}{2}-c_{32} & & \frac{\sigma_3}{3}-c_{33} & -c_{34} & x_{34} & & \end{array}\right].$$

There are three parameters, x_{14}, x_{24} and x_{34}, which can be used for the minimization of sums over the columns of right submatrix. In order to obtain such solution, we should minimize the following function (by using the Theorem 8.2) $h : \mathbb{R}^3 \to \mathbb{R}$:

$$\begin{aligned} h(x_{14}, x_{24}, x_{34}) &= x_{34}^2 + (x_{34} - c_{24})^2 + (x_{34} - c_{24} - c_{13})^2 + x_{24}^2 + (x_{24} - c_{14})^2 + x_{14}^2 = \\ &= G(x_{34}; -c_{24}, -c_{13}) + G(x_{24}; -c_{14}) + x_{14}^2. \end{aligned}$$

According to Theorem 8.2, the minimum is obtained for

$$x_{34}^* = \frac{2c_{24} + c_{13}}{3},\ x_{24}^* = \frac{c_{14}}{2},\ x_{14}^* = 0,$$

and it is $h(x_{14}^*, x_{24}^*, x_{34}^*) = (3c_{14}^2 + 4(c_{24}^2 + c_{24}c_{13} + c_{13}^2))/6$.

Therefore, the least-squares minimal-norm solution is:

$$\widehat{X_0} = \begin{bmatrix} -(c_{12}+c_{23}+c_{34}) & -(2c_{24}+c_{13})/3 & -c_{14}/2 & 0 \\ (2c_{11}-c_{22}-c_{33})/3 & -(c_{23}+c_{34}) & (c_{13}-c_{24})/3 & c_{14}/2 \\ (c_{21}-c_{32})/2 & (c_{11}+c_{22}-2c_{33})/3 & -c_{34} & (c_{13}+2c_{24})/3 \end{bmatrix}.$$

Note that the solution depends on all entries of the matrix C except c_{31}.

7. Return to the Main Case

We have thoroughly examined the simplest case, so we can return to our original problem.

$$\begin{aligned} \min_X \|AX - XB - C\|_F^2 \le \alpha^2(S,T)\ \min_Y \|J_A Y - Y J_B - D\|_F^2 = \\ = \alpha^2(S,T)\ \sum_{i=1}^{s}\sum_{u=1}^{k_i}\sum_{v=1}^{\ell_i} \min_{Y_{uv}^{(ii)}} \|J_u(0)Y_{uv}^{(ii)} - Y_{uv}^{(ii)}J_v(0) - D_{uv}^{(ii)}\|_F^2 = \\ = \alpha^2(S,T)\ \sum_{i=1}^{s}\sum_{u=1}^{k_i}\sum_{v=1}^{\ell_i}\sum_{k=1}^{\min\{u,v\}?} \frac{\sigma_k^2(D_{uv}^{(ii)})}{k} \end{aligned}$$

If the right hand side is zero, i.e.

$$\sigma_k(D_{uv}^{(ii)}) = 0,\ k = \overline{1, \min\{u,v\}},\ i = \overline{1,s},$$

then the equation is consistent. This result is independent of the choice of matrices S and T (except the request that proper Jordan blocks should be on appropriate positions).

If the right hand side is not zero, at least we have an upper bound (which need not be the best possible!) for the error.

Remark that if $A = SJ_AS^{-1}$, then $A = (\gamma S)J_A(\gamma S)^{-1}$ for any $\gamma \neq 0$, so one can further analyze quantity $\alpha(S,T)$.

8. Appendix: Some Auxiliary Results on the Minimum of the Functions

Suppose that $C \in \mathbb{C}^{m\times n}$ is a complex matrix. It can be written on the unique way as $C = C' + iC''$, where $C', C'' \in \mathbb{R}^{m\times n}$. Note that $\|C\|_F^2 = \|C'\|_F^2 + \|C''\|_F^2$. Because of:

$$\begin{aligned}\Delta(X;C) :&= \|J_m(0)X - XJ_n(0) - C\|_F^2 = \\ &= \|J_m(0)(X' + iX'') - (X' + iX'')J_n(0) - (C' + iC'')\|_F^2 = \\ &= \|J_m(0)X' - X'J_n(0) - C' + i(J_m(0)X'' - X''J_n(0) - C'')\|_F^2 = \\ &= \|J_m(0)X' - X'J_n(0) - C'\|_F^2 + \|J_m(0)X'' - X''J_n(0) - C''\|_F^2 = \\ &= \Delta(X';C') + \Delta(X'';C''),\end{aligned}$$

it is enough to consider minimization of appropriate real function.

The following two results can be easily proven by elementary calculations, but since they are the backbone for Theorems 5.1–6.1, we formulate them as propositions.

Proposition 8.1. 1) The function $M_1 : \mathbb{R}^n \to \mathbb{R}$ depending on $n+1$ real parameters $a_1, ..., a_{n+1}$,

$$M_1(x_1, ..., x_n; a_1, ..., a_n, a_{n+1}) = (x_1 - a_1)^2 + \sum_{k=1}^{n-1}(x_{k+1} - x_k - a_{k+1})^2 + (x_n + a_{n+1})^2$$

attains its minimum for the uniquely determined arguments

$$\widehat{x_k} = \frac{1}{n+1}\left((n+1-k)\sum_{i=1}^{k} a_i - k\sum_{i=k+1}^{n+1} a_i\right) = \sum_{i=1}^{k} a_i - \frac{k}{n+1}\sum_{i=1}^{n+1} a_i,\ k = \overline{1,n},$$

and this minimal value is

$$M_1(\widehat{x_1}, ..., \widehat{x_n}; a_1, ..., a_{n+1}) = \frac{1}{n+1}\left(\sum_{i=1}^{n+1} a_i\right)^2.$$

2) The function $M_2 : \mathbb{R}^n \to \mathbb{R}$, depending on n real parameters $a_1, ..., a_n$,

$$M_2(x_1, ..., x_n; a_1, ..., a_n) = (x_1 - a_1)^2 + \sum_{k=1}^{n-1}(x_{k+1} - x_k - a_{k+1})^2$$

attains its minimum 0 for the uniquely determined arguments

$$\widehat{x_k} = \sum_{i=1}^{k} a_i,\ k = \overline{1,n}.$$

3) The function $M_3 : \mathbb{R}^n \to \mathbb{R}$, depending on n real parameters $a_1, ..., a_n$,

$$M_3(x_1, ..., x_n; a_1, ..., a_n) = \sum_{k=1}^{n-1}(x_{k+1} - x_k - a_k)^2 + (x_n + a_n)^2$$

attains its minimum 0 for the uniquely determined arguments

$$\widehat{x_k} = -\sum_{i=k}^{n} a_i,\ k = \overline{1,n}.$$

4) The function $M_4 : \mathbb{R}^n \to \mathbb{R}$, depending on $n-1$ real parameters $a_1, ..., a_{n-1}$,

$$M_4(x_1, ..., x_n; a_1, ..., a_{n-1}) = \sum_{k=1}^{n-1} (x_{k+1} - x_k - a_k)^2$$

attains its minimum 0 for the one-parameter set of the arguments

$$\widehat{x_k} = x_1 + \sum_{i=2}^{k} a_{i-1},\ k = \overline{2,n},$$

or

$$\widehat{x_k} = x_n - \sum_{i=k}^{n-1} a_i,\ k = \overline{1,n-1}.$$

Proposition 8.2. 1) The function $F : \mathbb{R} \to \mathbb{R}$, depending on n real parameters $a_1, ..., a_n$:

$$F(x; a_1, ..., a_n) = x^2 + (x + a_1)^2 + ... + (x + a_n)^2,$$

attains its minimum at

$$x^* = -\frac{a_1 + ... + a_n}{n+1},$$

and this minimal value is

$$F(x^*; a_1, ..., a_n) = \sum_{k=1}^{n} a_k^2 - \frac{1}{n+1}\left(\sum_{k=1}^{n} a_k\right)^2 = \frac{1}{n+1}\left(n\sum_{j=1}^{n} a_j^2 - 2\sum_{j<k} a_j a_k\right).$$

2) The function $G(x; a_1, ..., a_n) = F(x; a_1, a_1 + a_2, ..., a_1 + a_2 + ... + a_n)$, depending on n real parameters $a_1, ..., a_n$, attains its minimum for

$$x^* = -\sum_{j=1}^{n} \frac{n+1-j}{n+1} a_j,$$

and

$$G(x^*; a_1, ..., a_n) = \sum_{k=1}^{n}\left(\sum_{j=1}^{k} a_j\right)^2 - \frac{1}{n+1}\left(\sum_{k=1}^{n}\sum_{j=1}^{k} a_j\right)^2.$$

References

[1] S. Albeverio, A. K. Motovilov and A. A. Shkalikov, *Bounds on variation of spectral subspaces under J−self-adjoint perturbations*, Int. Equ. Oper. Theory 64 (2009), 455–486.

[2] W. Arendt, F. Räbiger and A. Sourour, *Spectral properties of the operator equation AX + XB = Y*, Quart. J. Math. Oxford (2) 45 (1994), 133–149.

[3] A. Ben-Israel and T. N. E. Greville, *Generalized inverses, theory and applications*, 2nd edition, Springer, 2003.

[4] P. Benner, *Factorized solutions of Sylvester equations wth applications in control*, in Proc. of the 16th International Symposium on Mathematical Theory of Network and Systems (MTNS 2004), 2004.

[5] R. Bhatia and P. Rosenthal, *How and why to solve the operator equation AX − XB = Y*, Bull. London Math. Soc. 29 (1997), 1–21

[6] R. Byers, *Solving the algebraic Ricatti equation with the matrix sign function*, Linear Algebra Appl. 85 (1987), 267–279.

[7] N. Č. Dinčić, *Solving the Sylvester equation AX − XB = C when σ(A) ∩ σ(B) ≠ ∅*, Electron. J. Linear Algebra 35 (2019), 1–23.

[8] B. D. Djordjević and N. Č. Dinčić, *Solving the operator equation $AX - XB = C$ with closed A and B*, Integral Equation and Operator Theory (2018) 90:51
[9] M. P. Drazin, *On a result of J. J. Sylvester*, Linear Algebra Appl. 505 (2016), 361–366.
[10] F. R. Gantmacher, *The theory of matrices, vol 1*, Chelsea Pub. Co., 1959.
[11] R. E. Hartwig and P. Patrício, *Group inverse of the nivellateur*, Linear and Multilinear Algebra 66 (12) (2018), 2409–2420
[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
[13] Q. Hu, D. Cheng, *The polynomial solution to the Sylvester matrix equation*, Applied Mathematics Letters 19 (9) (2006), 859–864.
[14] D. Y. Hu, L. Reichel, *Krylov-subspace methods for the Sylvester equation*, Linear Algebra Appl. 172 (1992), 283–313.
[15] X. Jin, Y. Wei, *Numerical linear algebra and its applications*, SCIENCE PRESS USA Inc. Beijing, 2005.
[16] N. T. Lan, *On the operator equation $AX - XB = C$ with unbounded operators A, B and C*, Abstract and Applied Analysis 6 (6) (2001), 317–328.
[17] G. Lumer, M. Rosenblum, *Linear operator equations*, Proc. Amer. Math. Soc. 10 (1959), 32–41.
[18] E-C. Ma, *A finite series solution of the matrix equation $AX - XB = C$*, SIAM J. Appl. Math. 14 (3) (1966), 490–495.
[19] S. Mecheri, *Why we solve the operator equation $AX - XB = C$*, preprint, (www.researchgate.net/publication/240626459_Why_we_solve_the_operator_equation_AX_XB_C)
[20] V. Q. Phóng, *The operator equation $AX - XB = C$ with unbounded operators A and B and related abstract Cauchy problems*, Math. Z. 208 (1991), 567–588.
[21] J. D. Roberts, *Linear model reduction and solution of the algebraic Ricatti equation by use of the sign function*, Internat. J. Control 32 (1980), 677–687.
[22] M. Rosenblum, *On the operator equation $BX - XA = Q$*, Duke Math. J. 23 (1956), 263–270
[23] J. J. Sylvester, *Sur l'equation en matrices $px = xq$*, C. R. Acad. Sci. Paris, 99 (1884) 67–71 and 115–116.