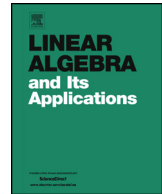




Contents lists available at ScienceDirect

## Linear Algebra and its Applications

journal homepage: [www.elsevier.com/locate/laa](http://www.elsevier.com/locate/laa)

# Minimum-norm solutions of the non-symmetric semidefinite Procrustes problem <sup>☆</sup>

Nicolas Gillis, Stefano Sicilia <sup>\*,1</sup>

Department of Mathematics and Operational Research, University of Mons, Mons, Belgium

## ARTICLE INFO

## Article history:

Received 13 November 2024

Received in revised form 20 March 2025

Accepted 8 June 2025

Available online 16 June 2025

Submitted by V. Mehrmann

Dataset link: [https://github.com/StefanoSicilia/NS\\_Procr\\_min\\_norm](https://github.com/StefanoSicilia/NS_Procr_min_norm)

## MSC:

15A29

65F20

90C22

## Keywords:

Procrustes problem

Positive semidefinite matrices

Minimum-norm solutions

Minimum-rank solutions

## ABSTRACT

Given two matrices  $X, B \in \mathbb{R}^{n \times m}$  and a set  $\mathcal{A} \subseteq \mathbb{R}^{n \times n}$ , a Procrustes problem consists in finding a matrix  $A \in \mathcal{A}$  such that the Frobenius norm of  $AX - B$  is minimized. When  $\mathcal{A}$  is the set of the matrices whose symmetric part is positive semidefinite, we obtain the so-called non-symmetric positive semidefinite Procrustes (NSPSDP) problem. The NSPSDP problem arises in the estimation of compliance or stiffness matrix in solid and elastic structures. If  $X$  has rank  $r$ , Baghel et al. (2022) [4] proposed a three-step semi-analytical approach: (1) construct a reduced NSPSDP problem in dimension  $r \times r$ , (2) solve the reduced problem by means of a fast gradient method with a linear rate of convergence, and (3) post-process the solution of the reduced problem to construct a solution of the larger original NSPSDP problem. In this paper, we revisit this approach of Baghel et al. and identify an unnecessary assumption used by the authors leading to cases where their algorithm cannot attain a minimum and produces solutions with unbounded norm. In fact, revising the post-processing phase of their semi-analytical approach, we show that the infimum of the NSPSDP problem is always attained, and we show how to compute a minimum-norm solution. We also prove that

<sup>☆</sup> We acknowledge the support by the European Union (ERC consolidator, eLinoR, no 101085607).

<sup>\*</sup> Corresponding author.

E-mail addresses: [Nicolas.GILLIS@umons.ac.be](mailto:Nicolas.GILLIS@umons.ac.be) (N. Gillis), [Stefano.SICILIA@umons.ac.be](mailto:Stefano.SICILIA@umons.ac.be) (S. Sicilia).

<sup>1</sup> Part of this work was done when SS was with the Division of Mathematics, Gran Sasso Science Institute, L'Aquila, Italy. SS is member of the Gruppo Nazionale Calcolo Scientifico-Istituto Nazionale di Alta Matematica (GNCS-INdAM).

the symmetric part of the computed solution has minimum rank bounded by  $r$ , and that the skew-symmetric part has rank bounded by  $2r$ . Several numerical examples show the efficiency of this algorithm, both in terms of computational speed and of finding optimal minimum-norm solutions.

© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

## 1. Introduction

A structured mapping problem associated to the set  $\mathcal{A} \subseteq \mathbb{R}^{n \times n}$  consists in finding a matrix  $A \in \mathcal{A}$  such that  $AX = B$ , where  $X, B \in \mathbb{R}^{n \times m}$  are two fixed matrices; see, e.g., [1,18] and the references therein. The corresponding Procrustes problem aims to solve a mapping structured problem when it is inconsistent, that is, when there is no matrix  $A \in \mathcal{A}$  such that  $AX = B$ , leading to the following optimization problem

$$\inf_{A \in \mathcal{A}} \|AX - B\|, \quad \text{where } \|\cdot\| \text{ is some norm.} \quad (1)$$

Many constraints have been studied for the set  $\mathcal{A}$ , such as the orthogonal matrices arising in computer graphics [9,12], Jordan algebra and Lie algebra for which Adhikari and Alam have provided an analytical solution [2], positive semidefiniteness in [5,22] arising in elastic structures, and rank- $r$  matrices with additional constraints in [17], to name just a few.

In this paper we focus on a set  $\mathcal{A}$  requiring a semidefinite property of the matrix. There are two main cases and the set  $\mathcal{A}$  is one of the following

$$\mathcal{S}_{\succeq}^n := \{M \in \mathbb{R}^{n \times n} : M \succeq 0\}, \quad \text{or} \quad \mathcal{N}_{\succeq}^n := \{M \in \mathbb{R}^{n \times n} : (M + M^\top) \succeq 0\},$$

where  $M \succeq N$  means that  $M - N$  is positive semidefinite, and 0 is the zero matrix of appropriate dimension. The set  $\mathcal{S}_{\succeq}^n$  denotes the set of all symmetric and positive semidefinite matrices, while the set  $\mathcal{N}_{\succeq}^n$  contains all the matrices whose symmetric part is positive semidefinite. Given  $X, B \in \mathbb{R}^{n \times m}$ , the positive semidefinite Procrustes (PSDP) problem is

$$\inf_{A \in \mathcal{S}_{\succeq}^n} \|AX - B\|_F^2, \quad (\mathcal{P})$$

while the non-symmetric positive semidefinite Procrustes (NSPSDP) problem is

$$\inf_{A \in \mathcal{N}_{\succeq}^n} \|AX - B\|_F^2, \quad (\mathcal{NP})$$

where  $\|\cdot\|_F$  denotes the Frobenius norm. These problems are both convex, since they minimize a convex function over a convex set. While  $(\mathcal{P})$  is well-known in the literature

and has been studied in many works [5,8,13,15,23],  $(\mathcal{NP})$  is more recent and has been considered in [4,14,15]. Both PSDP and NSPSDP problems arise in the estimate of local compliance matrices during deformable object in various engineering applications; see [4,14]. The compliance matrix represents the Green function matrix that corresponds to the unknown  $A$  in  $(\mathcal{NP})$ . The estimation of  $A$  is based on the observations of the forces, stored in the columns of  $X$ , and their associated displacements, stored in the columns of  $B$ , for a total of  $m$  measurements. When the deformable object is a passive object, that is, it does not generate energy in deformation, the physical properties of the model impose that  $p^\top A p \geq 0$ , where  $p$  is the point load, which is implied by  $A \in \mathcal{N}_\Sigma^n$ . Thus, when estimating  $A$  under noisy measurements, it is crucial to impose the constraint  $A \in \mathcal{N}_\Sigma^n$  in the optimization problem; see [16] and [14, Section III] for more details about the physical model.

### 1.1. Contributions and outline of the paper

For both  $(\mathcal{P})$  and  $(\mathcal{NP})$ , the original  $n \times m$  dimensional problem can be reduced to a problem of dimension  $r \times r$ , where  $r$  is the rank of  $X$ ; see [8] for  $(\mathcal{P})$ , and [4] for  $(\mathcal{NP})$ . This is referred to as a semi-analytical approach because the smaller  $r \times r$  dimensional problem still needs to be solved with iterative methods, and [4,8] relied on an optimal first-order method with linear convergence. It is important to note that the solution of  $(\mathcal{P})$  is not always attained [8]; see Section 2.1 where we recall this result. The approach in [4] follows closely that of [8], and concludes that the solution of  $(\mathcal{NP})$  is also not always attained. However, it turns out this result is not correct, unless one uses an additional constraint within the NSPSDP problem, which is not explicitly stated in [4].

In this work, inspired by the approach of [4], we provide an updated theorem about the NSPSDP problem, namely, we show that the infimum is always attained and provide a semi-analytical solution (Theorem 3.1). Moreover, we show how to compute a minimum-norm solution for which we also prove low-rank properties, namely the symmetric part has rank at most  $r$ , and the skew-symmetric part at most  $2r$  (Theorem 3.4). This allows us to propose a new algorithmic approach for the NSPSDP problem, which always provides optimal and minimum-norm solutions, as opposed to [4].

The paper is organized as follows. In Section 2, we recall the semi-analytical methods proposed in [8] for solving  $(\mathcal{P})$  and in [4] for solving  $(\mathcal{NP})$ . We discuss in detail the main issue of the latter approach, namely the introduction of an unnecessary additional constraint in the problem, which leads to non-optimal and unbounded solutions. In Section 3, we show how to fix this issue, and prove that the infimum of the NSPSDP problem is always attained. We also discuss how to compute the minimum-norm solution and show its low-rank properties. In Section 4, we develop an algorithm relying on the theoretical results from Section 3. Section 5 shows the effectiveness of this new algorithm compared to that of [4] in several numerical examples.

## 2. An overview of Procrustes problems with positive semidefinite constraints

In this section, we briefly recall the existing approaches by Gillis and Sharma [8] for solving  $(\mathcal{P})$ , and by Baghel et al. [4] for solving  $(\mathcal{NP})$ . For the latter, we provide a revised version of their main theorem, stating explicitly the hidden constraint in their result.

### 2.1. The PSDP problem

Let us briefly recall the existence of optimal solutions in the PSDP problem; this will allow us to compare and shed light on the NSPSD case.

Before doing this, let us recall an important result about positive semidefinite characterization that will be used throughout the paper.

**Lemma 2.1.** [3] *Let  $R \in \mathbb{R}^{n \times n}$  be a symmetric matrix partitioned as*

$$R = \begin{pmatrix} B & C^\top \\ C & D \end{pmatrix}.$$

*Then  $R \succeq 0$  if and only if the following three conditions hold: (1)  $B \succeq 0$ , (2)  $\ker(B) \subseteq \ker(C)$ , and (3)  $D - CB^\dagger C^\top \succeq 0$ , where  $B^\dagger$  denotes the Moore-Penrose pseudo-inverse of  $B$ .*

When the matrix  $X$  has full row rank, the solution to problem  $(\mathcal{P})$  is unique, as ensured by the next lemma.

**Lemma 2.2.** [23, Theorem 2.2] *Let  $X, B \in \mathbb{R}^{n \times m}$  and assume that  $X$  has rank  $n$ . Then the infimum of problem  $(\mathcal{P})$  is attained for a unique solution  $A_\star \in \mathcal{S}_\Sigma^n$ .*

**Proof.** This follows from the fact that the problem is strongly convex.  $\square$

Now we state the major result of [8], where a minimum-norm solution of the Procrustes problem  $(\mathcal{P})$  is provided via a semi-analytical approach.

**Theorem 2.3.** *Let  $X, B \in \mathbb{R}^{n \times m}$  and consider the SVD*

$$X = U\Sigma V^\top := (U_1 \ U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (V_1 \ V_2)^\top,$$

*where  $\Sigma_1 \in \mathbb{R}^{r \times r}$  and  $r = \text{rank}(X)$ . Then*

$$\inf_{A \in \mathcal{S}_\Sigma^n} \|AX - B\|_F^2 = \min_{A_{11} \in \mathcal{S}_\Sigma^r} \|A_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 + \|B V_2\|_F^2. \quad (2)$$

*Moreover, letting*

$$\hat{A}_{11} := \arg \min_{A_{11} \in \mathcal{S}_{\leq}^r} \|A_{11}\Sigma_1 - U_1^\top B V_1\|_F^2,$$

and  $Z := U_2^\top B V_1 \Sigma_1^{-1}$ , it holds that

1. If  $\ker(\hat{A}_{11}) \subseteq \ker(Z)$ , the matrix

$$A_\star = U_1 \hat{A}_{11} U_1^\top + U_2 Z U_1^\top + U_1 Z^\top U_2^\top + U_2 K U_2^\top,$$

where  $K = Z \hat{A}_{11}^\dagger Z^\top$ , is the unique matrix that attains the infimum in (2) and whose rank, Frobenius norm and spectral norm are minimum.

2. Otherwise the infimum is not attained, and we can construct a solution  $A_\star^{(\varepsilon)} \in \mathcal{S}_{\leq}^n$  with minimal rank, minimum Frobenius norm and minimum spectral norm, such that

$$\|A_\star^{(\varepsilon)} X - B\|_F^2 \leq \inf_{A \in \mathcal{S}_{\leq}^n} \|AX - B\|_F^2 + \varepsilon,$$

for any  $\varepsilon > 0$  sufficiently small.

**Proof.** We refer to [8, Theorem 1] and [8, Corollary 1] for the proof.  $\square$

**Remark 1.** The explicit expressions for  $\varepsilon$  and  $A_\star^{(\varepsilon)}$  in Theorem 2.3 in the case the infimum is not attained can be found in [8]. These formulas are analogous to the ones of Theorem 2.5 presented in the next section.

## 2.2. The NSPSDP problem

The result of Theorem 2.3 for problem ( $\mathcal{P}$ ) has been adapted by Baghel et al. [4] for ( $\mathcal{NP}$ ), but the approach they proposed contains some incorrect statements. Before showing the issues in the main result of [4], we start by recalling an analogous result to Lemma 2.2 for the NSPSDP problem, which guarantees a unique solution when  $X$  has full row rank.

**Lemma 2.4.** [15, Theorem 2.4.6] Let  $X, B \in \mathbb{R}^{n \times m}$  and assume that  $X$  has rank  $n$ . Then the infimum of problem ( $\mathcal{NP}$ ) is attained for a unique solution  $A_\star \in \mathcal{N}_{\leq}^n$ .

**Proof.** This follows from the fact that the problem is strongly convex.  $\square$

Now we discuss the major result in [4]. Let us define the set

$$\mathcal{C} = \left\{ \begin{pmatrix} C_{11} & 0 \\ C_{12} & C_{22} \end{pmatrix} : C_{11} \in \mathbb{R}^{r \times r}, C_{12} \in \mathbb{R}^{(n-r) \times r}, C_{22} \in \mathbb{R}^{(n-r) \times (n-r)} \right\} \subseteq \mathbb{R}^{n \times n}, \quad (3)$$

where there is a  $r \times (n - r)$  block of zeros at the top right of any matrix in  $\mathcal{C}$ . For a given matrix  $U \in \mathbb{R}^{n \times n}$ , let us also introduce the sets

$$UCU^\top = \{UCU^\top : C \in \mathcal{C}\}, \quad \text{and} \quad \mathcal{N}_{\Sigma}^{UCU^\top} = \mathcal{N}_{\Sigma}^n \cap UCU^\top.$$

The main issue with [4, Theorem 2] is that it considers a different version of the NSPSDP problem, where it is implicitly assumed that the solution  $A$  belongs to  $UCU^\top$ , where the matrix  $U$  comes from the SVD of  $X = U\Sigma V^\top$ . Hence the problem considered in [4] is actually

$$\inf_{A \in \mathcal{N}_{\Sigma}^{UCU^\top}} \|AX - B\|_F^2, \quad (\mathcal{N}\mathcal{P}\mathcal{C})$$

which is different than the original NSPSDP problem  $(\mathcal{N}\mathcal{P})$ . Although the infimums of  $(\mathcal{N}\mathcal{P}\mathcal{C})$  and  $(\mathcal{N}\mathcal{P})$  coincide, the optimal solution of  $(\mathcal{N}\mathcal{P}\mathcal{C})$  might not be attained in some cases, see Theorem 2.5 below, while that of  $(\mathcal{N}\mathcal{P})$  always is; see Theorem 3.1 in the next section.

In the following we provide a revised version of [4, Theorem 2] with this additional constraint which is not present in the original paper but was accidentally imposed in the proof of the result.

**Theorem 2.5.** [4, Theorem 2, revised] Let  $X, B \in \mathbb{R}^{n \times m}$  and consider the SVD

$$X = U\Sigma V^\top := (U_1 \quad U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (V_1 \quad V_2)^\top, \quad (4)$$

where  $\Sigma_1 \in \mathbb{R}^{r \times r}$  has full rank. Then

$$\inf_{A \in \mathcal{N}_{\Sigma}^{UCU^\top}} \|AX - B\|_F^2 = \inf_{A \in \mathcal{N}_{\Sigma}^n} \|AX - B\|_F^2 = \min_{A_{11} \in \mathcal{N}_{\Sigma_1}^r} \|A_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 + \|B V_2\|_F^2.$$

Moreover, let us introduce

$$\hat{A}_{11} = \arg \min_{A_{11} \in \mathcal{N}_{\Sigma_1}^r} \|A_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 := \hat{H}_{11} + \hat{S}_{11},$$

where  $\hat{H}_{11}$  and  $\hat{S}_{11}$  are, respectively, the symmetric and skew-symmetric part of  $\hat{A}_{11}$ . Let  $Z = U_2^\top B V_1 \Sigma_1^{-1}$ , the following holds:

1. If  $\ker(\hat{H}_{11}) \subseteq \ker(Z)$ , then a solution of the problem  $(\mathcal{N}\mathcal{P}\mathcal{C})$  is

$$A_\star = U \begin{pmatrix} \hat{A}_{11} & 0 \\ Z & K + R \end{pmatrix} U^\top = U_1 \hat{A}_{11} U_1^\top + U_2 Z U_1^\top + U_2 (K + R) U_2^\top,$$

where  $R \in \mathbb{R}^{(n-r) \times (n-r)}$  is skew-symmetric and  $K \in \mathbb{R}^{(n-r) \times (n-r)}$  is symmetric such that

$$K \succeq \frac{1}{4} Z H_{11}^\dagger Z^\top.$$

Since the infimum of  $(\mathcal{NP})$  and  $(\mathcal{NPC})$  coincide (see (4)),  $A_\star$  is also an optimal solution of  $(\mathcal{NP})$ .

2. Otherwise the infimum of  $(\mathcal{NPC})$  is not attained, and we can construct a solution  $A_\star^{(\varepsilon)} \in \mathcal{N}_\Sigma^n$  such that

$$\|A_\star^{(\varepsilon)} X - B\|_F^2 \leq \inf_{A \in \mathcal{N}_\Sigma^n} \|AX - B\|_F^2 + \varepsilon$$

for any

$$0 < \varepsilon < \begin{cases} \min\{1, \|\hat{A}_{11}\Sigma_1 - U_1^\top B V_1\|_F^2\} & \text{if } \|\hat{A}_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Consider the two SVDs

$$\begin{aligned} \hat{H}_{11} &= (W_1 \quad W_2) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} (W_1 \quad W_2)^\top, \\ \hat{H}_{11}^{(\varepsilon)} &= (W_1 \quad W_2) \begin{pmatrix} \Lambda & 0 \\ 0 & \Gamma^{(\varepsilon)} \end{pmatrix} (W_1 \quad W_2)^\top, \end{aligned}$$

where  $\Lambda \in \mathbb{R}^{s \times s}$  is invertible and  $\Gamma^{(\varepsilon)} \in \mathbb{R}^{r-s \times r-s}$  is the diagonal matrix with entries  $\varepsilon/\beta$ , with

$$\beta = \begin{cases} 4\sqrt{r-s} \|\Sigma_1\|_F \|\hat{A}_{11}\Sigma_1 - U_1^\top B V_1\|_F & \text{if } \|\hat{A}_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 \neq 0, \\ 4\sqrt{r-s} \|\Sigma_1\|_F & \text{otherwise.} \end{cases}$$

Defining  $\hat{A}_{11}^{(\varepsilon)} = \hat{H}_{11}^{(\varepsilon)} + \hat{S}_{11}$ , we construct  $A_\star^{(\varepsilon)}$  as follows

$$A_\star^{(\varepsilon)} = U \begin{pmatrix} \hat{A}_{11} & 0 \\ Z & K^{(\varepsilon)} + R \end{pmatrix} U^\top = U_1 \hat{A}_{11} U_1^\top + U_2 Z U_1^\top + U_2 (K^{(\varepsilon)} + R) U_2^\top, \quad (5)$$

where  $R \in \mathbb{R}^{(n-r) \times (n-r)}$  is skew-symmetric and  $K^{(\varepsilon)} \succeq \frac{1}{4} Z (\hat{H}_{11}^{(\varepsilon)})^{-1} Z^\top$ .

**Proof.** We report the proof of [4, Theorem 2] until the additional constraint  $A \in UCU^\top$  arises, where the set  $\mathcal{C}$  is defined in (3). Let  $A \in \mathcal{N}_\Sigma^n$  and set

$$\hat{A} := U^\top A U = U^\top \begin{pmatrix} H_{11} & H_{21}^\top \\ H_{21} & H_{22} \end{pmatrix} U + U^\top \begin{pmatrix} S_{11} & -S_{21}^\top \\ S_{21} & S_{22} \end{pmatrix} U := H + S,$$

where  $H$  and  $S$  are, respectively, the symmetric and skew-symmetric part of  $\hat{A}$  that have been decomposed and partitioned in blocks such that  $H_{11}, S_{11} \in \mathbb{R}^{r \times r}$ . By definition  $S_{11}$  and  $S_{22}$  are skew-symmetric and, since  $H \succeq 0$ , Lemma 2.1 implies that

$$H_{11} \succeq 0, \quad \ker(H_{11}) \subseteq \ker(H_{21}), \quad H_{22} - H_{21}H_{11}^\dagger H_{21}^\top \succeq 0.$$

We rewrite the objective function and we get

$$\begin{aligned} \|AX - B\|_F^2 &= \|U\hat{A}U^\top X - B\|_F^2 = \|(H + S)U^\top X - U^\top B\|_F^2 \\ &= \|(H_{11} + S_{11})U_1^\top X - U_1^\top B\|_F^2 + \|(H_{21} + S_{21})U_1^\top X - U_2^\top B\|_F^2 \\ &= \left\| \begin{pmatrix} (H_{11} + S_{11})\Sigma_1 - U_1^\top BV_1 & -U_1^\top BV_2 \end{pmatrix} \right\|_F^2 \\ &\quad + \left\| \begin{pmatrix} (H_{21} + S_{21})\Sigma_1 - U_2^\top BV_1 & -U_2^\top BV_2 \end{pmatrix} \right\|_F^2 \\ &= \|(H_{11} + S_{11})\Sigma_1 - U_1^\top BV_1\|_F^2 + \|(H_{21} + S_{21})\Sigma_1 - U_1^\top BV_1\|_F^2 + \|U_1^\top BV_2\|_F^2 \\ &\quad + \|U_2^\top BV_2\|_F^2, \end{aligned}$$

which implies

$$\|AX - B\|_F^2 = \|(H_{11} + S_{11})\Sigma_1 - U_1^\top BV_1\|_F^2 + \|(H_{21} + S_{21})\Sigma_1 - U_2^\top BV_1\|_F^2 + \|BV_2\|_F^2. \quad (6)$$

By taking the infimum of (6),  $\inf_{A \in \mathcal{N}_\Sigma^n} \|AX - B\|_F^2$  is equal to

$$\inf_{\substack{H_{11} \succeq 0, S_{11} = -S_{11}^\top, \\ \ker(H_{11}) \subseteq \ker(H_{21})}} \|(H_{11} + S_{11})\Sigma_1 - U_1^\top BV_1\|_F^2 + \|(H_{21} + S_{21})\Sigma_1 - U_2^\top BV_1\|_F^2 + \|BV_2\|_F^2.$$

We note that in this last infimum the matrices  $H_{22}$  and  $S_{22}$  do not appear, meaning that the conditions  $S_{22} = -S_{22}^\top$  and  $H_{22} - H_{21}H_{11}^\dagger H_{21}^\top \succeq 0$  can be ignored. By dropping the condition  $\ker(H_{11}) \subseteq \ker(H_{21})$  and by denoting  $E = H_{21} + S_{21}$ , we obtain

$$\begin{aligned} \inf_{A \in \mathcal{N}_\Sigma^n} \|AX - B\|_F^2 &\geq \inf_{\substack{H_{11} \succeq 0 \\ S_{11} = -S_{11}^\top}} \|(H_{11} + S_{11})\Sigma_1 - U_1^\top BV_1\|_F^2 \\ &\quad + \inf_{E \in \mathbb{R}^{(n-r) \times r}} \|E\Sigma_1 - U_2^\top BV_1\|_F^2 + \|BV_2\|_F^2. \end{aligned}$$

Since the second infimum in the right hand side is 0 and it is uniquely attained for  $E_\star = H_{21} + S_{21} = U_2^\top BV_1 \Sigma_1^{-1} := Z$ , it holds that

$$\inf_{A \in \mathcal{N}_\Sigma^n} \|AX - B\|_F^2 \geq \|\hat{A}_{11}\Sigma_1 - U_1^\top BV_1\|_F^2 + \|BV_2\|_F^2,$$

where  $\hat{A}_{11} = \hat{H}_{11} + \hat{S}_{11} \in \mathcal{N}_\Sigma^r$  is the unique solution of the first infimum (Lemma 2.4), and  $\hat{H}_{11}, \hat{S}_{11}$  are its symmetric and skew-symmetric part, respectively. Thus the conditions found are



$$\begin{aligned} H_{11} &= \hat{H}_{11}, \quad S_{11} = \hat{S}_{11}, \quad H_{21} + S_{21} = Z, \\ \ker(\hat{H}_{11}) &\subseteq \ker(H_{21}), \quad H_{22} - H_{21}\hat{H}_{11}^\dagger H_{21}^\top \succeq 0, \end{aligned} \quad (7)$$

together with the underlying constraint that  $S_{22}$  is skew-symmetric. The restriction of  $A \in UCU^\top$  is introduced here and it implies that  $H_{21} - S_{21} = 0$ . Thus all the constraints yield that the solution is of the form

$$A_\star = U \left( \begin{pmatrix} \hat{H}_{11} & \frac{1}{2}Z^\top \\ \frac{1}{2}Z & K \end{pmatrix} + \begin{pmatrix} \hat{S}_{11} & -\frac{1}{2}Z^\top \\ \frac{1}{2}Z & R \end{pmatrix} \right) U^\top = U \begin{pmatrix} \hat{A}_{11} & 0 \\ Z & K + R \end{pmatrix} U^\top,$$

where  $K \succeq \frac{1}{4}ZH_{11}^\dagger Z^\top$  and  $R \in \mathbb{R}^{(n-r) \times (n-r)}$  is skew-symmetric. The imposition of the constraint  $A \in UCU^\top$  implies that the infimum of the problem is not always attained. Indeed the condition of (7) that  $\ker(\hat{H}_{11}) \subseteq \ker(Z)$  is not fulfilled in general, especially if  $\ker(\hat{H}_{11}) \neq \{0\}$ . For the continuation of the proof, we refer to [4, Theorem 2], where all the sub-cases (infimum attained or not) are discussed in detail.  $\square$

**Remark 2.** Besides solving a different NSPSDP problem with the additional constraint  $A \in \mathcal{N}_{\succeq}^{UCU^\top}$ , the issue of the method by Baghel et al. concerns the norm of the solution found when the infimum of  $(\mathcal{N}\mathcal{P}\mathcal{C})$  is not attained, that is, when  $X$  has low-rank and  $\ker(\hat{H}_{11}) \neq \{0\}$  is not contained in  $\ker(Z)$ . In that case, the matrix  $K^{(\varepsilon)} \succeq \frac{1}{4}Z(\hat{H}_{11}^{(\varepsilon)})^{-1}Z^\top$  used to define  $A_\star^{(\varepsilon)}$  has a very large norm, going to infinity as  $\varepsilon$  goes to zero. This leads to a huge norm solution, as it will be shown in the numerical examples.

### 3. A new approach for the NSPSDP problem

In Section 2, we have detected that an additional constraint  $A \in UCU^\top$  had been introduced in the optimization problem studied in [4], changing it from  $(\mathcal{N}\mathcal{P})$  to  $(\mathcal{N}\mathcal{P}\mathcal{C})$ . Thus in the following result, we fix this issue and uncover the remarkable property that, as opposed to  $(\mathcal{P})$ ,

*the infimum of the NSPSDP problem  $(\mathcal{N}\mathcal{P})$  is always attained.*

We characterize the set of minimizers of the problem and we determine the family of the solutions for the original NSPSDP problem.

**Theorem 3.1.** *Let  $X, B \in \mathbb{R}^{n \times m}$  and consider the SVD*

$$X = U\Sigma V^\top := (U_1 \quad U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (V_1 \quad V_2)^\top,$$

*where  $\Sigma_1 \in \mathbb{R}^{r \times r}$  has full rank. Then*

$$\inf_{A \in \mathcal{N}_{\succeq}^n} \|AX - B\|_F^2 = \min_{A_{11} \in \mathcal{N}_{\succeq}^r} \|A_{11}\Sigma_1 - U_1^\top B V_1\|_F^2 + \|B V_2\|_F^2.$$

Moreover, let us introduce

$$\hat{A}_{11} = \arg \min_{A_{11} \in \mathcal{N}_{\Sigma}^r} \|A_{11}\Sigma_1 - U_1^\top BV_1\|_F^2 := \hat{H}_{11} + \hat{S}_{11},$$

where  $\hat{H}_{11}$  and  $\hat{S}_{11}$  are, respectively, the symmetric and skew-symmetric part of  $\hat{A}_{11}$ . Then any solution of the problem  $(\mathcal{NP})$  is of the form

$$A_\star = U \begin{pmatrix} \hat{A}_{11} & M^\top \\ Z & N + G \end{pmatrix} U^\top = U_1 \hat{A}_{11} U_1^\top + U_2 Z U_1^\top + U_1 M^\top U_2^\top + U_2 (N + G) U_2^\top,$$

where  $Z = U_2^\top BV_1 \Sigma_1^{-1} \in \mathbb{R}^{(n-r) \times r}$ ,  $M \in \mathbb{R}^{(n-r) \times r}$ ,  $G \in \mathbb{R}^{(n-r) \times (n-r)}$  is skew-symmetric,  $N \in \mathbb{R}^{(n-r) \times (n-r)}$  is symmetric, and they fulfill the conditions

$$\ker(\hat{H}_{11}) \subseteq \ker(Z + M), \quad N \succeq \frac{1}{4}(Z + M)H_{11}^\dagger(Z + M)^\top,$$

so that  $A_\star \in \mathcal{N}_{\Sigma}^n$ .

**Proof.** We proceed as in Theorem 2.5, until the introduction of the unnecessary constraint that the solution belongs to  $UCU^\top$ . With the same notation, let  $A \in \mathcal{N}_{\Sigma}^n$  and set

$$\hat{A} := U^\top AU = U(H + S)U^\top := U \left( \begin{pmatrix} H_{11} & H_{21}^\top \\ H_{21} & H_{22} \end{pmatrix} + \begin{pmatrix} S_{11} & -S_{21}^\top \\ S_{21} & S_{22} \end{pmatrix} \right) U^\top.$$

Let  $\hat{A}_{11} \in \mathcal{N}_{\Sigma}^r$  be the unique solution (provided by Lemma 2.4) of

$$\inf_{A \in \mathcal{N}_{\Sigma}^n} \|AX - B\|_F^2 = \|\hat{A}_{11}\Sigma_1 - U_1^\top BV_1\|_F^2 + \|BV_2\|_F^2,$$

and denote by  $\hat{H}_{11}, \hat{S}_{11}$  its symmetric and skew-symmetric part, respectively. The conditions on the sub-matrices of the solution found in the proof of Theorem 2.5 are

$$H_{11} = \hat{H}_{11}, \quad S_{11} = \hat{S}_{11}, \quad H_{21} + S_{21} = U_2^\top BV_1 \Sigma_1^{-1} = Z, \quad \ker(\hat{H}_{11}) \subseteq \ker(H_{21}).$$

In particular it is not required that  $H_{21} = S_{21} = \frac{1}{2}Z$ , but only that their sum is equal to  $Z$ . Thus this constraint simply implies that  $A_\star$  is of the form

$$A_\star = U \begin{pmatrix} \hat{A}_{11} & M^\top \\ Z & N + G \end{pmatrix} U^\top = U_1 \hat{A}_{11} U_1^\top + U_2 Z U_1^\top + U_1 M^\top U_2^\top + U_2 (N + G) U_2^\top, \quad (8)$$

where  $M \in \mathbb{R}^{(n-r) \times r}$ ,  $G \in \mathbb{R}^{(n-r) \times (n-r)}$  is skew-symmetric,  $N \in \mathbb{R}^{(n-r) \times (n-r)}$  is symmetric and they fulfill the conditions

$$\ker(\hat{H}_{11}) \subseteq \ker(Z + M), \quad N \succeq \frac{1}{4}(Z + M)H_{11}^\dagger(Z + M)^\top. \quad (9)$$

This ensures that all  $A_\star$  of the form of (8) belong to  $\mathcal{N}_\Sigma^n$ . Since they satisfy (6), they are all solutions of the optimization problem  $(\mathcal{NP})$ .  $\square$

Theorem 3.1 shows that

$$UFU^\top := \{UFU^\top : F \in \mathcal{F}\} = \arg \min_{A \in \mathcal{N}_\Sigma^n} \|AX - B\|_F^2, \quad (10)$$

where  $\mathcal{F}$  is the matrix family defined as

$$\mathcal{F} = \left\{ \begin{pmatrix} \hat{A}_{11} & M^\top \\ Z & N + G \end{pmatrix} : G = -G^\top, N \succeq \frac{1}{4}(Z + M)H_{11}^\dagger(Z + M)^\top, \right. \\ \left. Z = U_2^\top BV_1 \Sigma_1^{-1}, \ker(\hat{H}_{11}) \subseteq \ker(Z + M) \right\}.$$

We notice that  $\mathcal{F}$  is not empty, since for instance

$$F := \begin{pmatrix} \hat{A}_{11} & -Z^\top \\ Z & 0 \end{pmatrix} \in \mathcal{F},$$

by choosing  $M = -Z$  and  $N = 0$  that fulfill the constraints given in (9). This provides a solution  $A = UFU^\top$  of  $(\mathcal{NP})$  that attains the infimum, but  $A \notin UCU^\top$  and hence it cannot be found by the method proposed in [4].

Now that we have introduced a new family of solutions for the NSPSDP problem  $(\mathcal{NP})$ , we provide a simplified characterization of the matrix family  $\mathcal{F}$ , in which the constraint concerning the kernels in (9) is removed. In this way it will be also easier to highlight low-norm and low-rank properties of matrices in the family.

**Lemma 3.2.** *With the same notation as in Theorem 3.1, the set  $\mathcal{F}$  that can be used to characterize optimal solutions of  $(\mathcal{NP})$  (see (10)) is given by*

$$\mathcal{F} = \left\{ \begin{pmatrix} \hat{A}_{11} & (YW_1^\top - Z)^\top \\ Z & N + G \end{pmatrix} : G = -G^\top, N \succeq \frac{1}{4}Y\Lambda^{-1}Y^\top, Y \in \mathbb{R}^{(n-r) \times s}, \right. \\ \left. Z = U_2^\top BV_1 \Sigma_1^{-1} \right\},$$

where

$$\frac{\hat{A}_{11} + \hat{A}_{11}^\top}{2} = \hat{H}_{11} := (W_1 \quad W_2) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} (W_1 \quad W_2)^\top$$

is an SVD of  $\hat{H}_{11}$ ,  $s \leq r$ ,  $\Lambda \in \mathbb{R}^{s \times s}$  is invertible,  $W_1 \in \mathbb{R}^{r \times s}$  and  $W_2 \in \mathbb{R}^{r \times (r-s)}$ .

**Proof.** The condition on the null spaces of  $\hat{H}_{11}$  and  $Z + M$  in (9) is equivalent to state that there exists a matrix  $Y \in \mathbb{R}^{(n-r) \times s}$  such that  $Z + M = YW_1^\top$ . Indeed  $W := (W_1 \ W_2) \in \mathbb{R}^{r \times r}$  is an orthonormal matrix and  $\ker(\hat{H}_{11}) = W_2$  implies

$$(Z + M)W_2 = YW_1^\top W_2 = 0.$$

Thus

$$N \succeq \frac{1}{4}YW_1^\top(W_1\Lambda^{-1}W_1^\top)W_1Y^\top = \frac{1}{4}Y\Lambda^{-1}Y^\top. \quad \square$$

### 3.1. Minimum-norm solution for the NSPSDP problem

Now we look for the matrices in the set of optimal solutions of the NSPSDP problem, namely  $U\mathcal{F}U^\top$ , with the smallest possible Frobenius norm. That is, we want to minimize the Frobenius norm of

$$A = U \begin{pmatrix} \hat{A}_{11} & (YW_1^\top - Z)^\top \\ Z & N + G \end{pmatrix} U^\top \in U\mathcal{F}U^\top$$

where all the matrices fulfill the conditions in Lemma 3.2. Let us introduce the differentiable function

$$f(Y) := \underbrace{\|YW_1^\top - Z\|_F^2}_{=:g(Y)} + \frac{1}{16} \underbrace{\|Y\Lambda^{-1}Y^\top\|_F^2}_{=:h(Y)}, \quad Y \in \mathbb{R}^{(n-r) \times s}. \quad (11)$$

Note that  $f$  implicitly depends on the fixed matrices  $X$  and  $B$ , since  $W_1$ ,  $Z$  and  $\Lambda$  do.

**Lemma 3.3.** *The function  $f$  defined in (11) is strictly convex in  $Y$  and its minimum is unique.*

**Proof.** We first show that  $g$  is strictly convex and that  $h$  is convex, therefore their linear combination  $f$  is strictly convex. Then we prove that  $f$  attains a minimum, which is unique by strict convexity.

Let us define  $g_1(x) = x^2$  and  $g_2(Y) = \|YW_1^\top - Z\|_F$ . The function  $g_1$  is strictly convex and increasing on  $[0, +\infty)$ , while the function  $g_2$  is convex on  $\mathbb{R}^{(n-r) \times s}$ , since for any  $\alpha \in (0, 1)$  and any  $Y_1, Y_2 \in \mathbb{R}^{(n-r) \times s}$ ,

$$g_2(\alpha Y_1 + (1 - \alpha)Y_2) = \|\alpha(Y_1W_1^\top - Z) + (1 - \alpha)(Y_2W_1^\top - Z)\|_F \leq \alpha g_2(Y_1) + (1 - \alpha)g_2(Y_2).$$

Thus  $g = g_1 \circ g_2$  is strictly convex. In fact, its Hessian matrix is the identity, meaning that it is perfectly conditioned. The unique minimizer of  $g(Y)$  is given by  $ZW_1$ .

Let us now prove the convexity of the quartic function

$$h(Y) = \|Y\Lambda^{-1}Y^\top\|_F^2 = \text{trace}((Y\Lambda^{-1}Y^\top)^2).$$

Define the diagonal matrix  $D := \Lambda^{-1}$ , which has, by construction, positive diagonal entries, and let

$$h_1(X) = \text{trace}(X^2), \quad h_2(Y) = YDY^\top.$$

The function  $h_1$  is operator monotone (see [7, Definition 2.1]), that is, for any  $X_1, X_2 \in \mathcal{S}_{\leq}^{n-r}$  such that  $X_1 \succ X_2$ , we have  $h_1(X_1) > h_1(X_2)$ . Indeed, if  $\lambda_i$  and  $\mu_i$  are the eigenvalues of  $X_1$  and  $X_2$ , respectively,

$$\begin{aligned} \text{trace}(X_1^2 - X_2^2) &= \sum_{i=1}^{n-r} \lambda_i^2 - \mu_i^2 \\ &= \sum_{i=1}^{n-r} (\lambda_i + \mu_i)(\lambda_i - \mu_i) \geq (n-r) \min_i (\lambda_i + \mu_i) \text{trace}(X_1 - X_2) > 0 \end{aligned}$$

since  $X_1 - X_2 \succ 0$ . Moreover  $h_1$  is a convex function, since it fulfills the hypothesis of [7, Theorem 2.10], which states that if  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  is convex, then  $X \rightarrow \text{trace}(\varphi(X))$  is convex, where  $\varphi(X)$  is meant as a matrix function; in this case  $\varphi(x) = x^2$ . The function  $h_2$  is operator convex, that is, for any  $\alpha \in (0, 1)$  and any  $Y_1, Y_2 \in \mathbb{R}^{(n-r) \times s}$ ,

$$\begin{aligned} &\alpha g_2(Y_1) + (1-\alpha)g_2(Y_2) - g_2(\alpha Y_1 + (1-\alpha)Y_2) \\ &= \alpha(1-\alpha) (Y_1DY_1^\top + Y_2DY_2^\top - Y_1DY_2^\top - Y_2DY_1^\top) = \alpha(1-\alpha)(Y_1-Y_2)D(Y_1-Y_2)^\top \succeq 0. \end{aligned}$$

Thus, for any  $\alpha \in (0, 1)$  and any  $Y_1, Y_2 \in \mathbb{R}^{(n-r) \times s}$ , the properties of  $h_1$  and  $h_2$  yield the convexity of  $h$ :

$$\begin{aligned} h(\alpha Y_1 + (1-\alpha)Y_2) &= h_1(h_2(\alpha Y_1 + (1-\alpha)Y_2)) \\ &\leq h_1(\alpha h_2(Y_1) + (1-\alpha)h_2(Y_2)) \leq \alpha h(Y_1) + (1-\alpha)h(Y_2). \end{aligned}$$

The function  $f$  is trivially bounded from below by 0, and it is continuous with bounded level sets, since for any  $\rho > 0$

$$f(Y) \leq \rho \quad \Rightarrow \quad \|Y\|_F = \|YW_1^\top\|_F \leq \|YW_1^\top - Z\|_F + \|Z\|_F \leq \sqrt{\rho} + \|Z\|_F.$$

Thus its infimum is attained while strict convexity implies its uniqueness.  $\square$

Thanks to the properties of  $f(Y)$  shown in Lemma 3.3, the following result provides the unique minimum Frobenius norm solution of  $(\mathcal{NP})$ , and it also shows its low-rank properties.

**Theorem 3.4.** *Using the same notation as in Theorem 3.1, Lemma 3.2 and Lemma 3.3, the minimum Frobenius norm solution of  $(\mathcal{NP})$  is*

$$A_{\star} := U \begin{pmatrix} \hat{A}_{11} & (Y_{\star} W_1^{\top} - Z)^{\top} \\ Z & \frac{1}{4} Y_{\star} \Lambda^{-1} Y_{\star}^{\top} \end{pmatrix} U^{\top},$$

where  $Y_{\star}$  is the unique global minimizer of

$$\min_{Y \in \mathbb{R}^{(n-r) \times s}} \|Y W_1^{\top} - Z\|_F^2 + \frac{1}{16} \|Y \Lambda^{-1} Y^{\top}\|_F^2 = \min_{Y \in \mathbb{R}^{(n-r) \times s}} f(Y).$$

Moreover the symmetric part of  $A_{\star}$  has rank  $s$ , which is the smallest possible rank of the symmetric part of a matrix in  $UFU^{\top}$ , while the skew-symmetric part of  $A_{\star}$  has rank not greater than  $2r$ .

**Proof.** Since the Frobenius norm is unitary invariant and by using the characterization of Lemma 3.2, we obtain

$$\|A_{\star}\|_F^2 = \|\hat{A}_{11}\|_F^2 + \|Z\|_F^2 + \|Y_{\star} W_1^{\top} - Z\|_F^2 + \|N + G\|_F^2.$$

The first two addends of the sum are fixed. For the fourth, it holds that

$$\|N + G\|_F^2 = \|N\|_F^2 + 2\text{trace}(NG) + \|G\|_F^2 = \|N\|_F^2 + \|G\|_F^2,$$

since  $\text{trace}(NG) = -\text{trace}(NG) = 0$  and thus the choice  $G = 0$  is optimal. Regarding the choice of  $N$ , recall that the definition of  $\mathcal{F}$  implies that  $N = \Delta + \frac{1}{4} Y \Lambda^{-1} Y^{\top}$  for some  $\Delta \succeq 0$ . As shown in [4, Lemma 3], the unique solution of the minimization problem

$$\min_{\Delta \in \mathcal{N}_{\succeq}^n} \left\| \Delta + \frac{1}{4} Y \Lambda^{-1} Y^{\top} \right\|_F^2$$

is  $\Delta = 0$  and hence the optimal choice  $N = \frac{1}{4} Y \Lambda^{-1} Y^{\top}$  yields

$$\inf_{A \in U\mathcal{F}U^{\top}} \|A\|_F^2 = \|\hat{A}_{11}\|_F^2 + \|Z\|_F^2 + \inf_{Y \in \mathbb{R}^{(n-r) \times s}} \left( \|Y W_1^{\top} - Z\|_F^2 + \frac{1}{16} \|Y \Lambda^{-1} Y^{\top}\|_F^2 \right). \quad (12)$$

Hence, minimizing the norm of  $A \in U\mathcal{F}U^{\top}$  is equivalent, up to constant summands, to minimize  $f(Y)$ . Lemma 3.3 shows that the minimum of  $f$  is uniquely attained and so  $A_{\star}$  is the unique minimizer of  $\min_{A \in U\mathcal{F}U^{\top}} \|A\|_F^2$ .

The symmetric part of  $A_{\star}$  is

$$H_{\star} = \frac{A_{\star} + A_{\star}^{\top}}{2} = \begin{pmatrix} \hat{H}_{11} & \frac{1}{2}(Y_{\star} W_1^{\top})^{\top} \\ \frac{1}{2} Y_{\star} W_1^{\top} & \frac{1}{4} Y_{\star} \Lambda^{-1} Y_{\star}^{\top} \end{pmatrix} = \begin{pmatrix} W_1 \\ \frac{1}{2} Y_{\star} \end{pmatrix} \begin{pmatrix} \Lambda & I_s \\ I_s & \Lambda^{-1} \end{pmatrix} \begin{pmatrix} W_1^{\top} & \frac{1}{2} Y_{\star}^{\top} \end{pmatrix},$$

where  $I_s$  denotes the  $s \times s$  identity matrix, and hence

$$\text{rank}(H_{\star}) = \text{rank} \begin{pmatrix} \Lambda & I_s \\ I_s & \Lambda^{-1} \end{pmatrix} = s.$$

Moreover any matrix in  $\mathcal{F}$  has symmetric part with rank not smaller than  $s$ , since using the Schur complement yields

$$\text{rank} \begin{pmatrix} \hat{H}_{11} & \frac{1}{2}(Y_{\star}W_1^{\top})^{\top} \\ \frac{1}{2}Y_{\star}W_1^{\top} & N \end{pmatrix} \geq \text{rank} \begin{pmatrix} \hat{H}_{11} & 0 \\ 0 & N - \frac{1}{4}Y_{\star}\Lambda^{-1}Y_{\star}^{\top} \end{pmatrix} \geq s.$$

To conclude the proof, we have for the skew-symmetric part of  $A_{\star}$  that

$$\text{rank} \left( \frac{A_{\star} - A_{\star}^{\top}}{2} \right) = \text{rank} \begin{pmatrix} \hat{S}_{11} & \frac{1}{2}(Y_{\star}W_1^{\top} - 2Z)^{\top} \\ \frac{1}{2}(2Z - Y_{\star}W_1^{\top}) & 0 \end{pmatrix} \leq 2r. \quad \square$$

### 3.2. Computing and approximating the minimum-norm solution

Theorem 3.4 describes the minimum-norm and minimum-rank solution of  $(\mathcal{NP})$ . However a closed-form for the unique solution (Lemma 3.3) of the problem,

$$Y_{\star} = \arg \min_{Y \in \mathbb{R}^{(n-r) \times s}} \|YW_1^{\top} - Z\|_F^2 + \frac{1}{16} \|Y\Lambda^{-1}Y^{\top}\|_F^2 = \arg \min_{Y \in \mathbb{R}^{(n-r) \times s}} f(Y), \quad (13)$$

is not available. We propose two different approaches to overcome this issue.

*Computing the minimum-norm solution* To solve (13), we use a standard non-linear optimization algorithm, namely gradient descent. The gradient of  $f$  can be computed (see, e.g., [21]) explicitly:

$$\nabla f(Y) = \frac{1}{4}Y\Lambda^{-1}Y^{\top}Y\Lambda^{-1} + 2(YW_1^{\top} - Z)W_1 = \frac{1}{4}Y\Lambda^{-1}Y^{\top}Y\Lambda^{-1} + 2Y - 2ZW_1.$$

Note that  $f(Y)$  is not globally Lipschitz because of the quartic term  $\|Y\Lambda^{-1}Y^{\top}\|_F^2$ . Although gradient descent is a relatively naive method, it worked very well for our purpose, because the problem is typically well conditioned (the Hessian is the identity plus a semidefinite matrix). In particular a simple backtracking line search will work well in our case. We update the current approximation  $Y$  by choosing a suitable step size  $\gamma$  that satisfies the Armijo condition:

$$f(Y - \gamma \nabla f(Y)) \leq f(Y) - c\gamma \|\nabla f(Y)\|^2, \quad (14)$$

where  $0 < c < 1$ . This allows us to get global convergence towards  $Y_{\star}$ , since  $f$  is trivially bounded from below by 0 and  $\nabla f$  is locally Lipschitz continuous (see [20, Theorem 3.2]). We will provide the details in Algorithm 2.

*Replacing the minimum-norm solution using Cardano's formula* Instead of approximating  $Y_{\star}$  using gradient descent applied on (13), we propose the following approximation:

$$Y = \alpha ZW_1, \quad \alpha \in [0, 1].$$

This choice represents a convex combination of the minimizer of the first term  $\|YW_1^\top - Z\|_F^2$  in (13), namely  $Y = ZW_1$ , and of the minimizer of the second term, namely  $Y = 0$ . The following lemma provides the optimal choice for the parameter  $\alpha$ , which minimizes  $f(\alpha ZW_1)$ .

**Lemma 3.5.** *Let*

$$\varphi(\alpha) := f(\alpha ZW_1) = \frac{\alpha^4}{16} \|ZW_1 \Lambda^{-1} W_1^\top Z^\top\|_F^2 + \alpha^2 \|ZW_1 W_1^\top\|_F^2 - 2\alpha \langle ZW_1 W_1^\top, Z \rangle + \|Z\|_F^2.$$

*Then, the unique minimum of  $\varphi$  is*

$$\alpha_\diamond := \arg \min_{\alpha \in \mathbb{R}} \varphi(\alpha) = \sqrt[3]{\frac{p}{2} + \sqrt{\frac{p^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{\frac{p}{2} - \sqrt{\frac{p^2}{4} + \frac{p^3}{27}}} \in (0, 1), \quad (15)$$

*where*

$$p = \frac{8\|ZW_1\|_F^2}{\|ZW_1 \Lambda^{-1} W_1^\top Z^\top\|_F^2} > 0.$$

**Proof.** The best value for  $\alpha$  can be computed by finding a real zero of the derivative of  $\varphi$ , given by

$$\varphi'(\alpha) = \frac{\alpha^3}{4} \|ZW_1 \Lambda^{-1} W_1^\top Z^\top\|_F^2 + 2\alpha \|ZW_1\|_F^2 - 2\|ZW_1\|_F^2.$$

Since  $\varphi'(0) < 0$  and  $\varphi'(1) > 0$ , there must be a zero in  $(0, 1)$ . The equation  $\varphi'(\alpha) = 0$  is equivalent to

$$\alpha^3 + p\alpha - p = 0, \quad (16)$$

which is a cubic equation with discriminant  $\Delta = -(4p^2 + 27p^3)$ . Since  $\Delta$  is negative, Cardano's formula (see, e.g., [6]) implies that the unique solution of (16) is  $\alpha_\diamond$  defined in (15).  $\square$

**Remark 3.** We have observed numerically that the solutions  $Y_\star$  and  $Y_\diamond = \alpha_\diamond ZW_1$  are typically close. Moreover it is possible to give a simple bound for the difference of the functional evaluated in these two points:

$$\|ZW_1 W_1^\top - Z\|_F^2 \leq f(Y_\star) \leq f(Y_\diamond) \leq f(\alpha ZW_1), \quad \forall \alpha \in \mathbb{R}$$



and the choices  $\alpha = 0$  and  $\alpha = 1$  imply

$$|f(Y_\star) - f(Y_\diamond)| \leq \min \left\{ \frac{1}{16} \|ZW_1 \Lambda^{-1} W_1^\top Z^\top\|_F^2, \|Z\|_F^2 - \|ZW_1 W_1^\top - Z\|_F^2 \right\}.$$

We will compare these two solutions in Section 5.

#### 4. An algorithm for the NSPSDP problem

Relying on Lemma 3.2 and Theorem 3.4, we now propose an algorithm for the NSPSDP problem ( $\mathcal{NP}$ ). The method follows the approach from [4], but it performs a different post-processing of the solution of the reduced  $r \times r$  dimensional problem, in order to avoid the unnecessary constraint introduced in ( $\mathcal{NPC}$ ), and hence avoid unbounded solutions when the solution of ( $\mathcal{NPC}$ ) is not attained. Algorithm 1 describes our proposed algorithm, which attains small Frobenius norm solutions with low-rank properties.

---

##### Algorithm 1 Algorithm for the NSPSDP problem ( $\mathcal{NP}$ ).

---

**Input:** Two matrices  $X, B \in \mathbb{R}^{n \times m}$ .

**Output:** An optimum minimum-norm solution  $A_\star$  of ( $\mathcal{NP}$ ); see Theorem 3.4.

---

- 1: Compute a rank-revealing SVD of  $X = (U_1 \ U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (V_1 \ V_2)^\top$ , with  $\Sigma_1 \in \mathbb{R}^{r \times r}$  where  $r = \text{rank}(X)$ .
- 2: Solve the reduced  $r \times r$  NSPSDP problem

$$\hat{A}_{11} = \arg \min_{A_{11} \in \mathcal{N}_{\Sigma}^c} \|A_{11} \Sigma_1 - U_1^\top B V_1\|_F^2 := \hat{H}_{11} + \hat{S}_{11},$$

by means of a fast gradient method [19], exactly as done in [4].

- 3: Compute  $Z = U_2^\top B V_1 \Sigma_1^{-1}$  and a thin SVD of  $\hat{H}_{11} = W_1 \Lambda W_1^\top$ , with  $\Lambda \in \mathbb{R}^{s \times s}$  where  $s = \text{rank}(\hat{H}_{11})$ .
- 4: Solve the optimization problem

$$Y_\star = \arg \min_{Y \in \mathbb{R}^{(n-r) \times s}} \|Y W_1^\top - Z\|_F^2 + \frac{1}{16} \|Y \Lambda^{-1} Y^\top\|_F^2,$$

using gradient descent with backtracking line search; see Algorithm 2.

- 5: Return the matrix

$$A_\star = (U_1 \ U_2) \begin{pmatrix} \hat{A}_{11} & (Y_\star W_1^\top - Z)^\top \\ Z & \frac{1}{4} Y_\star \Lambda^{-1} Y_\star^\top \end{pmatrix} (U_1 \ U_2)^\top.$$


---

The first two steps of Algorithm 1 involving the solution of the reduced problem are the same as in [4]. The novelty of our algorithm is the post-processing procedure described in Steps 3, 4 and 5, which ensures that the algorithm always provides a minimum of problem ( $\mathcal{NP}$ ), while [4] was only providing a large-norm approximate solution (see (5)), in the case the infimum of ( $\mathcal{NPC}$ ) is not attained.

*Computation of  $Y_\star$*  Step 4 in Algorithm 1 is solved by means of a gradient descent procedure, as shown in Algorithm 2.

---

**Algorithm 2** Backtracking line search with Armijo condition for solving problem (13).

---

**Input:** The matrices  $Z \in \mathbb{R}^{(n-r) \times r}$ ,  $W_1 \in \mathbb{R}^{r \times s}$ ,  $\Lambda \in \mathbb{R}^{s \times s}$ , a stepsize parameter  $\theta > 1$ , a parameter  $c > 0$ , maximum number of iterations  $i_{\max}$ , maximum number of checks for the Armijo condition  $j_{\max}$  and an error tolerance  $\tau$

**Output:** A solution  $Y_\star$  of (13), that minimizes  $f(Y) = \|YW_1^\top - Z\|_F^2 + \frac{1}{16} \|Y\Lambda^{-1}Y^\top\|_F^2$ .

- 1: Initialize  $Y_0 = \alpha_0 ZW_1$  defined by the Cardano approximation (15).
- 2: Set  $i = 0$  and  $e_0 = 1$ ; compute  $f(Y_0)$  and

$$\nabla f(Y_0) = \frac{1}{4} Y_0 \Lambda^{-1} Y_0^\top Y_0 \Lambda^{-1} + 2Y_0 - 2ZW_1.$$

- 3: **while**  $i < i_{\max}$  **and**  $\|\nabla f(Y_i)\|_F \leq \tau$  **and**  $e_i \leq 10^{-15}$  **do**
- 4:     Set

$$\gamma = \frac{1}{10} \cdot \frac{\|Y_i\|_F}{\|\nabla f(Y_i)\|_F}.$$

- 5:     **for**  $j = 1, \dots, j_{\max}$  **do**
  - 6:         Compute  $\hat{Y} = Y_i - \gamma \nabla f(Y_i)$ .
  - 7:         **if**  $f(\hat{Y}) \leq f(Y_i) - c\gamma \|\nabla f(Y_i)\|_F^2$  **then**
  - 8:             **break**
  - 9:         **else**
  - 10:              $\gamma = \gamma/\theta$ .
  - 11:         **end if**
  - 12:     **end for**
  - 13:     Update  $Y_{i+1} := \hat{Y}$  and compute  $\nabla f(Y_{i+1})$ .
  - 14:     Set  $i = i + 1$  and  $e_i = \frac{|f(Y_{i+1}) - f(Y_i)|}{f(Y_i)}$ .
  - 15: **end while**
  - 16: **return**  $Y_\star := Y_i$ .
- 

The parameter  $\theta$  introduced in the algorithm determines the reduction rate for the stepsize  $\gamma$ , the constant  $c$  is used in the Armijo condition (14), and the safety parameter  $j_{\max} = 50$  stops the algorithm in the case the line search fails to fulfill (14). In our experiments, we have always used  $\theta = 1.5$  and  $c = 10^{-4}$ . In Section 5, we give more details about the stopping criterion for the algorithm.

*Computing  $Y_\diamond$*  In Step 4 of Algorithm 1 we may replace  $Y_\star$  with  $Y_\diamond = \alpha_\diamond ZW_1$ , where  $\alpha_\diamond$  is defined in (15). This approximate solution allows us to avoid solving (13), while still preserving the low-rank properties of  $A_\star$ , since

$$A_\diamond := U \begin{pmatrix} \hat{A}_{11} & (Y_\diamond W_1^\top - Z)^\top \\ Z & \frac{1}{4} Y_\diamond \Lambda^{-1} Y_\diamond^\top \end{pmatrix} U^\top$$

has a symmetric part of rank  $s$  and skew-symmetric part of rank bounded by  $2r$ , as shown in the proof of Theorem 3.4 for  $Y_\star$ . However the solution  $A_\diamond$  is not guaranteed

to have minimum Frobenius norm. In the numerical experiments in Section 5, we will compare these two approaches, both in terms of computational time and size of the norm of the solution.

#### 4.1. Computational complexity

Given  $X, B \in \mathbb{R}^{n \times m}$  with  $\text{rank}(X) = r$ , let us analyze the computational complexity of Algorithm 1:

1. The SVD in Step 1 requires  $\mathcal{O}(mn \min(m, n))$  flops.
2. As shown in [4], the complexity of the fast gradient method applied to the  $r \times r$  NSPSDP problem is dominated by the computation of the gradient  $AXX^\top - BX^\top$ , which is  $\mathcal{O}(\tau_1 r^3)$ , where  $\tau_1$  is the number of iterations performed. See Section 5 below for a discussion on the stopping criterion used.
3. For Step 3, the cost of the computation of  $Z$  is  $\mathcal{O}(nmr)$ , while the computation of the SVD of  $\hat{H}_{11}$  is  $\mathcal{O}(r^3)$ .
4. The computation of the solution of the optimization problem in Step 4 is performed by means of a gradient descent method, which requires the computation of the matrix  $\nabla f(Y) = Y\Lambda^{-1}(Y^\top Y)\Lambda^{-1} + 2(Y - ZW_1)$  at each iteration, leading to a cost of  $\mathcal{O}(\tau_2((n-r)^2 r + s^3))$ , where  $\tau_2$  is the number of iterations performed. See Section 5 below for a discussion on the stopping criterion used.
5. Step 5 requires matrix multiplications whose cost sums up to  $\mathcal{O}(n^2 r)$  flops.

Thus the overall cost of the algorithm is  $\mathcal{O}(mn \min(m, n) + \tau_2(n-r)^2 r + \tau_1 r^3)$ . This cost reduces to  $\mathcal{O}(mn \min(m, n) + \tau_1 r^3)$  if the solution of the optimization problem in Step 4 of Algorithm 1 is replaced by  $Y_\diamond = \alpha_\diamond ZW_1$ , with  $\alpha_\diamond$  defined as in (15).

Typically,  $m \gg n \geq r$  since the number of variables is smaller than the number of measurements, and hence the most expensive step is the SVD computation in Step 1 in  $\mathcal{O}(mn \min(m, n))$  operations. We will provide numerical experiments on the scalability of Algorithm 1 in Section 5.5.

#### 4.2. The algorithm for nearly low-rank $X$

The solution of the Procrustes problem ( $\mathcal{NP}$ ) may be also modified in order to take advantage of a possible low-rank property of the input data. When the matrix  $X$  is close to be rank-deficient, it is possible to replace it by its  $(k-1)$ -rank approximation  $\tilde{X}$  such that

$$\sigma_k \leq \eta \sigma_1,$$

where  $\sigma_i$  are the singular values of  $X$  and  $0 < \eta \ll 1$  is a small tolerance. The following result gives an idea of how this modification changes the error  $\|AX - B\|_F$  of the problem.

**Theorem 4.1.** Given  $\eta > 0$ , let  $X = \tilde{X} + \Delta X$  be a decomposition of the form

$$X = U \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_{k-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} V^\top + U \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_k & & \\ & & & & \ddots & \\ & & & & & \sigma_q \end{pmatrix} V^\top,$$

where  $q = \min(m, n)$  and  $\sigma_k \leq \eta\sigma_1$ . Define

$$\tilde{\gamma} := \min_{A \in \mathcal{N}_{\Sigma}^n} \|A\tilde{X} - B\|_F := \|\tilde{A}\tilde{X} - B\|_F, \quad \gamma_\star := \min_{A \in \mathcal{N}_{\Sigma}^n} \|AX - B\|_F := \|A_\star X - B\|_F,$$

where  $A_\star$  is the unique minimizer of the original NSPSDP problem, while  $\tilde{A}$  is a minimizer of the perturbed NSPSDP problem with smallest norm. Then

$$\tilde{\gamma} - \xi \|A_\star\|_F \leq \gamma_\star \leq \tilde{\gamma} + \xi \|\tilde{A}\|_F, \quad (17)$$

where  $\xi = \eta\sigma_1\sqrt{q-k+1}$ .

**Proof.** The norm of the perturbation  $\Delta X$  is bounded by  $\xi$  since

$$\|\Delta X\|_F = \sqrt{\sum_{i=k}^q \sigma_i^2} \leq \sqrt{\eta^2(q-k+1)\sigma_1^2} \leq \eta\sigma_1\sqrt{q-k+1} = \xi.$$

Thus

$$\begin{aligned} \gamma_\star &= \|A_\star X - B\|_F \leq \|\tilde{A}(\tilde{X} + \Delta X) - B\|_F \leq \tilde{\gamma} + \|\tilde{A}\|_F \|\Delta X\|_F, \\ \tilde{\gamma} &= \|\tilde{A}\tilde{X} - B\|_F \leq \|A_\star(X - \Delta X) - B\|_F \leq \gamma_\star + \|A_\star\|_F \|\Delta X\|_F \end{aligned}$$

and the claim follows.  $\square$

While the upper bound of (17) contains  $\|\tilde{A}\|_F$ , which is generally small, the lower bound provided depends on  $\|A_\star\|_F$ , which may be very large when  $X$  has small non-zero singular values. Thus even a small perturbation can cause a huge change in the value of the functional  $\|AX - B\|_F$ . We will discuss more in detail this behavior in the numerical experiments in Section 5.

## 5. Numerical experiments

In this section, we compare the performances of four different methods for solving the NSPSDP problem:

- Fast Gradient Method (FGM), applied directly to  $(\mathcal{NP})$ , an optimal first-order method for convex problems developed in [19], as done in [4].
- ANalytic Fast Gradient Method (ANFGM) uses the semi-analytical approach proposed in [4] to reduce the dimension of the problem from  $n \times m$  to  $r \times r$ , solves it via FGM, and then uses the post-processing as in Theorem 2.5. This algorithm solves a more constrained NSPSDP problem, namely  $(\mathcal{NPC})$ .
- MINimum Gradient Descent (MINGD) implements Algorithm 1.
- CARDano (CARD) follows the same approach as MINGD by implementing Algorithm 1, but it replaced the minimum-norm solution in step 4 by the solution obtained via the Cardano's formula; see (15).

In [4], the performances of the interior point method (IPM) proposed in [14] and the convex MATLAB software (CVX) [10,11] was compared to FGM and ANFGM. In most cases, IPM and CVX were slower and less accurate and hence we do not report their results here. The code and experiments are available from [https://github.com/StefanoSicilia/NS\\_Procr\\_min\\_norm](https://github.com/StefanoSicilia/NS_Procr_min_norm).

Before providing a comparison on several types of data sets, let us discuss the stopping criterion and metrics used.

*Stopping criteria* ANFGM, MINGD and CARD reduce the size of the original problem, and then rely on the fast gradient method (FGM) proposed in [19]. We will use the same stopping criterion as in [4], namely when the  $k$ th iterate  $A^{(k)}$  satisfies

$$\|A^{(k)} - A^{(k-1)}\|_F < \delta \|A^{(1)} - A^{(0)}\|_F,$$

with  $\delta = 10^{-6}$ , that is, the modification of  $A$  compared to the first step is less than  $10^{-6}$ , or when the maximum number of iteration  $k_{\max} = 10000$  is reached. We will use this stopping criterion for FGM, ANFGM, MINGD and CARD in all numerical experiments.

Regarding Algorithm 2 for the computation of  $Y_*$  in MINGD, we stop the gradient descent algorithm when the iterate  $Y^{(k)}$  satisfies one of the following conditions

$$\|\nabla(Y^{(k)})\|_F \leq \tau := 10^{-8}, \quad \frac{|f(Y^{(k+1)}) - f(Y^{(k)})|}{f(Y^{(k)})} \leq 10^{-15},$$

or when the number of iteration number exceeds  $i_{\max} = 500$ . The stopping criterion that is most often reached first is the one based on the norm of the gradient. However, in some cases, we observed that the objective function reached machine accuracy (about  $10^{-15}$ ) while the norm of the gradient was significantly larger than  $10^{-8}$ . The reason is the presence of a quartic term in  $f$ .

*Metrics used to compare the algorithms* We will use the following metrics to compare the algorithms: given a solution  $A$  computed by an algorithm, we will report

**Table 5.1**  
Results on a sample of 20 matrices, with  $n = 50$ ,  $m = 70$ , and  $r = 20$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	$0.8674 \pm 0.0114$	$(2.07 \pm 0.48) \cdot 10^7$	$< \mathbf{0.01}$	$20 \pm 0$	$40 \pm 0$
FGM	$0.8674 \pm 0.0114$	$1.2845 \pm 0.0646$	$0.86 \pm 0.44$	$39.40 \pm 0.68$	$40 \pm 0$
MINGD	$0.8674 \pm 0.0114$	$\mathbf{0.9116 \pm 0.0576}$	$< \mathbf{0.01}$	$\mathbf{9.40 \pm 0.68}$	$40 \pm 0$
CARD	$0.8674 \pm 0.0114$	$0.9408 \pm 0.0582$	$< \mathbf{0.01}$	$\mathbf{9.40 \pm 0.68}$	$40 \pm 0$

**Table 5.2**  
Results on a sample of 20 matrices, with  $n = 100$ ,  $m = 100$ , and  $r = 40$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	$0.8007 \pm 0.0060$	$(2.12 \pm 0.22) \cdot 10^7$	$\mathbf{0.03 \pm 0.01}$	$40 \pm 0$	$80 \pm 0$
FGM	$0.8007 \pm 0.0060$	$1.5392 \pm 0.0461$	$5.37 \pm 1.86$	$78.90 \pm 0.64$	$80 \pm 0$
MINGD	$0.8007 \pm 0.0060$	$\mathbf{1.1459 \pm 0.0285}$	$0.05 \pm 0.03$	$\mathbf{18.90 \pm 0.64}$	$80 \pm 0$
CARD	$0.8007 \pm 0.0060$	$1.1960 \pm 0.0408$	$\mathbf{0.03 \pm 0.00}$	$\mathbf{18.90 \pm 0.64}$	$80 \pm 0$

1. The relative error

$$\frac{\|AX - B\|_F}{\|B\|_F},$$

which measures the quality of a solution.

- 2. The Frobenius norm of the solution,  $\|A\|_F$ .
- 3. The time required by the algorithm (in seconds).
- 4. The rank of the symmetric part of the solution,  $\text{rk}(\text{Sym}(A))$ .
- 5. The rank of the skew-symmetric part of the solution,  $\text{rk}(\text{Skew}(A))$ .

5.1. Synthetic data

Similarly to [4,8], we generate some synthetic data for different values of  $n, m$ . For each dimension, we generate 20 random matrices and report the mean and standard deviation of the results. Given  $m, n$  and  $r$ , a rank  $r$  matrix  $X \in \mathbb{R}^{n \times m}$  is generated as the product of two randomly generated matrices of dimension  $n \times r$  and  $r \times m$  whose entries follow the standard normal distribution, that is,  $\mathbf{X} = \text{randn}(\mathbf{n}, \mathbf{r}) * \text{randn}(\mathbf{r}, \mathbf{m})$  in MATLAB, while the entries of  $B$  are generated uniformly at random,  $\mathbf{B} = \text{randn}(\mathbf{n}, \mathbf{m})$  in MATLAB.

We choose  $r \ll \min(m, n)$  in order to highlight the low-rank properties of the algorithms and solutions. Tables 5.1, 5.2 and 5.3 show the results of the four algorithm to solve  $(\mathcal{NP})$  for various dimensions.

For all the choices of  $n, m$  and  $r$ , we observe a similar behavior of the algorithms:

- All the methods provide the same relative error, up to a difference of order  $10^{-9}$ . Although ANFGM does not solve the same problem, it constructs a solution using the parameter  $\varepsilon = 10^{-8}$  in (5), and hence the solution generated has a relative error

**Table 5.3**

Results on a sample of 20 matrices, with  $n = 200$ ,  $m = 200$ , and  $r = 50$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	$0.8776 \pm 0.0028$	$(9.22 \pm 4.84) \cdot 10^6$	<b><math>0.03 \pm 0.00</math></b>	$50 \pm 0$	$100 \pm 0$
FGM	$0.8776 \pm 0.0028$	$1.235 \pm 0.0198$	$16.67 \pm 6.34$	$171.60 \pm 4.07$	$100 \pm 0$
MINGD	$0.8776 \pm 0.0028$	<b><math>0.7859 \pm 0.0154</math></b>	$0.12 \pm 0.08$	<b><math>23.95 \pm 0.83</math></b>	$100 \pm 0$
CARD	$0.8776 \pm 0.0028$	$0.8203 \pm 0.021$	<b><math>0.03 \pm 0.01</math></b>	<b><math>23.95 \pm 0.83</math></b>	$100 \pm 0$

very close to that of the other algorithms, even though MINGD and CARD are preferable since they attain a minimum.

- The norm of the solution provided by ANFGM is very large (of order  $10^7$ ), while, in contrast, FGM, MINGD and CARD have solutions with norms of comparable size, although MINGD provides always the smallest value, as shown in Theorem 3.4.
- The CPU time required by FGM is significantly larger (approximately 50 times) than that required by the other algorithms, because it does not reduce the size of the problem.
- The symmetric part of the solution provided by MINGD and CARD has the lowest rank, followed by ANFGM. On the other hand, the symmetric part of the solution of FGM is almost full rank.
- All the skew-symmetric parts of the solutions have the same rank equal to  $2r$ .

In summary the best methods on our set of synthetic data are MINGD and CARD: they provide a solution with small norm (the one from CARD is generally 4% larger than that of MINGD), the computation is fast (MINGD is a bit slower than CARD; see the next sections for more experiments) and they satisfy the low-rank properties of Theorem 3.4, which makes the storage of the solution more efficient. Instead FGM is quite slow, while ANFGM provides solutions with very large norm and neither of these two methods provides a solution whose symmetric part has the smallest possible rank.

## 5.2. Synthetic data with perturbed low-rank $X$

In order to study the stability of the approaches, we add a full-rank perturbation of order  $\eta = 10^{-8}$  to the previous synthetic data and we observe the behavior of the algorithms. The perturbed matrix  $\mathbf{X} = \text{randn}(\mathbf{n}, \mathbf{r}) * \text{randn}(\mathbf{r}, \mathbf{m}) + \eta * \text{randn}(\mathbf{n}, \mathbf{m})$  is now full rank and this guarantees a unique solution to  $(\mathcal{NP})$ . We also consider two other versions of MINGD and CARD, that will be denoted by MINGD<sub>lr</sub> and CARD<sub>lr</sub> respectively, where the matrix  $X$  is replaced by its rank- $k$  approximation and  $k$  is chosen such that

$$\sigma_{k+1}(X) \leq \eta \sigma_1(X),$$

and we test the results given by Theorem 4.1.

**Table 5.4**  
Perturbed  $X$ : results on a sample of 20 matrices, with  $n = 50$ ,  $m = 70$ , and  $r = 20$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	<b>0.7544 ± 0.0135</b>	$(5.08 \pm 0.21) \cdot 10^8$	0.20 ± 0.01	43.10 ± 0.64	50 ± 0
FGM	0.8703 ± 0.0146	1.2692 ± 0.0486	2.06 ± 0.04	39.45 ± 0.69	50 ± 0
MINGD	<b>0.7544 ± 0.0135</b>	$(5.08 \pm 0.21) \cdot 10^8$	0.22 ± 0.04	43.10 ± 0.64	50 ± 0
CARD	<b>0.7544 ± 0.0135</b>	$(5.08 \pm 0.21) \cdot 10^8$	0.22 ± 0.03	43.10 ± 0.64	50 ± 0
MINGD <sub>lr</sub>	0.8703 ± 0.0146	<b>0.8986 ± 0.0410</b>	<b>0.01 ± 0.00</b>	<b>9.45 ± 0.69</b>	<b>40 ± 0</b>
CARD <sub>lr</sub>	0.8703 ± 0.0146	0.9272 ± 0.0441	<b>0.01 ± 0.00</b>	<b>9.45 ± 0.69</b>	<b>40 ± 0</b>

**Table 5.5**  
Perturbed  $X$ : results on a sample of 20 matrices, with  $n = 100$ ,  $m = 100$ , and  $r = 40$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	<b>0.6506 ± 0.0134</b>	$(2.16 \pm 0.98) \cdot 10^9$	0.70 ± 0.01	93.60 ± 1.82	100 ± 0
FGM	0.8064 ± 0.0073	1.5450 ± 0.0469	8.02 ± 0.51	78.50 ± 0.83	100 ± 0
MINGD	<b>0.6506 ± 0.0134</b>	$(2.16 \pm 0.98) \cdot 10^9$	0.89 ± 0.06	93.60 ± 1.82	100 ± 0
CARD	<b>0.6506 ± 0.0134</b>	$(2.16 \pm 0.98) \cdot 10^9$	0.80 ± 0.01	93.60 ± 1.82	100 ± 0
MINGD <sub>lr</sub>	0.8064 ± 0.0073	<b>1.1504 ± 0.0350</b>	0.04 ± 0.02	<b>18.50 ± 0.83</b>	<b>80 ± 0</b>
CARD <sub>lr</sub>	0.8064 ± 0.0073	1.1940 ± 0.0449	<b>0.03 ± 0.00</b>	<b>18.50 ± 0.83</b>	<b>80 ± 0</b>

**Table 5.6**  
Perturbed  $X$ : results on a sample of 20 matrices, with  $n = 200$ ,  $m = 200$ , and  $r = 50$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	<b>0.6401 ± 0.0098</b>	$(3.62 \pm 1.11) \cdot 10^9$	2.21 ± 0.05	192.25 ± 1.52	200 ± 0
FGM	0.8767 ± 0.0031	1.2380 ± 0.0196	32.07 ± 0.33	171.70 ± 2.83	200 ± 0
MINGD	<b>0.6401 ± 0.0098</b>	$(3.62 \pm 1.11) \cdot 10^9$	2.20 ± 0.06	192.70 ± 1.13	200 ± 0
CARD	<b>0.6401 ± 0.0098</b>	$(3.62 \pm 1.11) \cdot 10^9$	2.19 ± 0.06	192.70 ± 0.98	200 ± 0
MINGD <sub>lr</sub>	0.8767 ± 0.0031	<b>0.7863 ± 0.0169</b>	0.10 ± 0.06	<b>24.30 ± 0.80</b>	<b>100 ± 0</b>
CARD <sub>lr</sub>	0.8767 ± 0.0031	0.8223 ± 0.0215	<b>0.03 ± 0.00</b>	<b>24.30 ± 0.80</b>	<b>100 ± 0</b>

In this way, we analyze if it is possible to exploit the nearly low-rank properties of the problem, see the discussion around Theorem 4.1. Tables 5.4, 5.5 and 5.6 show the results of the six algorithms.

For all the choices of  $n, m$  and  $r$  we observe a similar behavior of the algorithms:

- ANFGM, MINGD and CARD provide the same result, by computing the unique solution of the perturbed NSPSDP problem. Thus relative error, norm of the solution and the rank of its symmetric and skew-symmetric part coincide.
- MINGD<sub>lr</sub> and CARD<sub>lr</sub> compute the solution of the low-rank problem associated to the low-rank approximation of  $X$ , and also FGM seems to do so. The reason is that FGM is a first-order method, with linear convergence of  $\mathcal{O}((1 - \kappa)^t)$ , where  $t$  is the iteration count and  $\kappa = \frac{\sigma_r(X)}{\sigma_1(X)} \in (0, 1)$  is the ratio between the  $r$ th and first singular value of  $X$ . Since the last  $n - r$  are very small, it cannot converge within the allotted number of iterations. These three methods have the same relative error, which is



**Table 5.7**

Results for the matrix NSPSDP problem with  $X = J_n J_m^\top$  and  $B = H$ . Best solution highlighted in bold, if applicable.

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM	0.9605	$5.8633 \cdot 10^{15}$	4.0469	500	498
FGM	0.9605	$1.0172 \cdot 10^4$	207.9887	321	20
MINGD	0.9605	<b><math>8.8618 \cdot 10^3</math></b>	3.8801	<b>5</b>	<b>12</b>
CARD	0.9605	$8.8633 \cdot 10^3$	<b>3.6061</b>	<b>5</b>	<b>12</b>

approximately 30 – 40% larger than the unperturbed solutions. However FGM is much slower than MINGD<sub>lr</sub> and CARD<sub>lr</sub>.

- The norm of the solutions of ANFGM, MINGD and CARD are very large, in contrast to the small norms of the other methods. The reason is that  $X$  has very small non-zero singular values. To be able to scale up these small contribution to approximate  $B$ , the norm of  $A$  in these directions must be of the order of the inverse of these singular values.
- ANFGM, MINGD and CARD are slower than MINGD<sub>lr</sub> and CARD<sub>lr</sub>. This is because they cannot reduce the problem size much since  $X$  has full rank, while MINGD<sub>lr</sub> and CARD<sub>lr</sub> work with a  $r \times r$  subproblem.
- Only the solution computed by MINGD<sub>lr</sub> and CARD<sub>lr</sub> have low-rank properties.

These experiments show that the low-rank approximation improves the speed of the algorithm and provides low-norm solution, with low-rank properties, but at the price of increasing the relative error. This is unavoidable, since the unique solution of the full-rank problem has large norm.

### 5.3. A large low-rank example

Now we consider an example of large dimension, with  $n = 500$ ,  $m = 10000$  and  $r = 10$ . We define

$$J_n = \begin{pmatrix} 1 & \cdots & 1 \\ 0 & \ddots & \vdots \\ \vdots & \ddots & 1 \\ 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n \times r}, \quad H = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ n & \cdots & 2 & 1 \end{pmatrix} \in \mathbb{R}^{n \times m}, \quad (18)$$

and we solve the NSPSDP problem ( $\mathcal{NP}$ ) for  $X = J_n J_m^\top$  and  $B = H$ . The results are shown in Table 5.7.

We observe similar results as the ones obtained for the synthetic data. In particular all the methods have the same relative error, ANFGM provides a large norm solution, and FGM is by far the slowest method. In contrast MINGD and CARD perform well and their solutions have low-rank properties. This example shows that it is possible to solve the low-rank NSPSDP problem of large dimension relatively fast, we further discuss the scalability in the next section.

**Table 5.8**  
Results for the tiger plush toy example where  $X$  is replaced by its best rank-2 approximation,  $X_2$ .

	Relative error	$\ A\ _F$	Time (sec.)	rk(Sym( $A$ ))	rk(Skew( $A$ ))
ANFGM (lr)	0.2022	6.6785	<0.01	2	2
FGM (lr)	0.2022	7.4655	0.0158	3	2
MINGD (lr)	0.2022	<b>6.6781</b>	0.0143	2	2
CARD (lr)	0.2022	<b>6.6781</b>	<0.01	2	2

5.4. *A real-world application*

As a practical example, we study the local compliance estimation problem already considered in [4,14], concerning a tiger plush toy. The dimensions of the problem are  $n = 3$  and  $m = 12$  and the data matrices (with two digits of accuracy) are

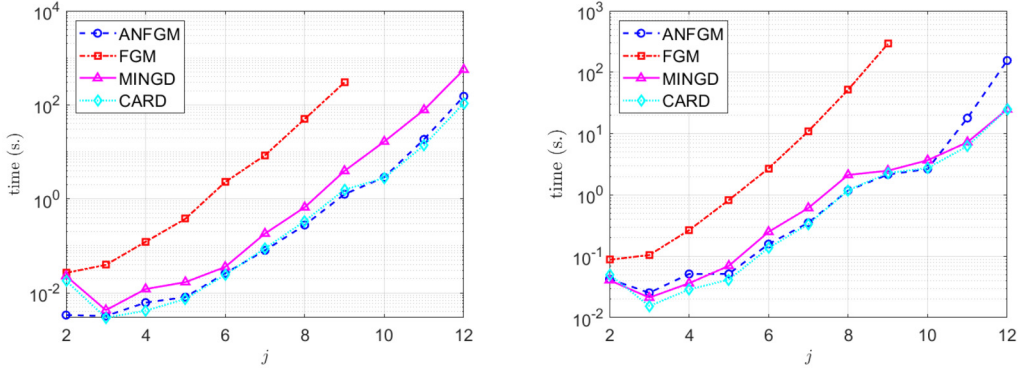
$$X^\top = \begin{pmatrix} -0.32 & 0.03 & 0.06 \\ -0.33 & -0.02 & 0.06 \\ -0.36 & 0.08 & 0.06 \\ -0.30 & 0.03 & 0.05 \\ -0.32 & 0.00 & 0.07 \\ -0.34 & 0.07 & 0.05 \\ -0.24 & 0.07 & 0.05 \\ -0.21 & -0.01 & 0.02 \\ -0.33 & 0.16 & 0.10 \\ -0.25 & 0.09 & 0.06 \\ -0.22 & 0.00 & 0.03 \\ -0.31 & 0.15 & 0.09 \end{pmatrix}, \quad B^\top = \begin{pmatrix} -1.43 & 0.15 & -0.44 \\ -1.40 & -0.31 & -0.42 \\ -1.38 & 0.44 & -0.42 \\ -1.43 & 0.14 & -0.44 \\ -1.40 & -0.31 & -0.42 \\ -1.37 & 0.43 & -0.42 \\ -1.43 & 0.16 & -0.43 \\ -1.40 & -0.32 & -0.42 \\ -1.38 & 0.42 & -0.43 \\ -1.43 & 0.15 & -0.44 \\ -1.40 & -0.33 & -0.42 \\ -1.37 & 0.42 & -0.44 \end{pmatrix}.$$

The matrix  $X$  has rank  $r = n = 3$ , hence the solution to the problem is unique and all the methods converge to the same matrix provided by [4,14], with a relative error of 18.99%. In order to run the algorithms in a non-full rank setting, we replace  $X$  by its best rank-2 approximation,  $X_2$ . Note that  $\|X - X_2\|_F / \|X\|_F = 3.5\%$  so  $X_2$  is a good approximation of  $X$ . Table 5.8 shows the results of the methods applied on  $X_2$ .

All methods find a solution with minimum relative error. However, CARD and MINGD are the only ones to find a minimizer with minimum norm and minimum rank; FGM does not find a solution with minimum rank nor minimum norm, and ANFGM does not find a solution with minimum norm.

5.5. *Scalability*

In order to study the scalability of the algorithm, we run the methods considered in the previous examples with increasing dimension and we show the trend of the computational time required to solve the problem. We study two examples, where the dimensions are fixed as  $n = m$  and  $r = n/2$  for  $n = 2^j$  with  $j = 2, \dots, 12$ . In the first example we consider a sample of 10 random matrices  $X, B \in \mathbb{R}^{n \times m}$  with  $\text{rank}(X) = r$ , as done for



**Fig. 5.1.** Scaling computational times for  $n = m = 2^j$  and  $r = n/2$ . On the left the average time over 10 runs with  $X$  a low-rank random matrix, while on the right the example defined by (18). FGM computation exceeded 500 seconds for  $n \geq 2^{10}$  in both cases and we did not report its result for larger  $n$ .

the synthetic data, while for the second example we consider  $X = J_n J_n^\top$  and  $B = H$ , with  $J_n$  and  $H$  defined as in (18).

Fig. 5.1 shows a similar trend for both examples. For ANFGM, MINGD and CARD, the complexity is dominated by the SVD computation ( $\mathcal{O}(mn^2)$  flops). The trend of the computational time is approximately a power function of  $n$  with exponent between 2 and 3. Also in these cases ANFGM is generating solutions with very large norms (e.g., larger than  $10^{14}$  for  $j \geq 9$ ), as in the previous examples, while FGM is significantly more expensive and its time requirement is larger for  $n = 2^9$  than that of the other three methods in dimension  $n = 2^{11}$ .

Interestingly, for the second example (Fig. 5.1, right plot), MINGD and CARD are faster than ANFGM for large values of  $n$  (namely  $n = 2^{11}, 2^{12}$ ). The reason is the initialization step of ANFGM that does not scale as well. For  $n = 2^{12}$ , the solution of MINGD and CARD coincide, and their CPU time is the same.

## 6. Conclusion

In this paper, we proposed a state-of-the-art semi-analytical approach for the NSPSDP problem. Our approach is inspired by that of Baghel et al. [4], but we resolved an issue when  $X$  is rank deficient. By doing so, we were able to prove that the solution of the NSPSDP problem is always attained (Theorem 3.1). Moreover, we proposed a way to compute the minimum-norm and minimum-rank solution (Theorem 3.4). We illustrated the effectiveness of the new proposed algorithm on several numerical experiments.

## Declaration of competing interest

There is no competing interest.

## Acknowledgement

We are very grateful to the editor and the anonymous reviewers who carefully read the manuscript; their feedback allowed us to improve it significantly.

## Data availability

Code available from [https://github.com/StefanoSicilia/NS\\_Procr\\_min\\_norm](https://github.com/StefanoSicilia/NS_Procr_min_norm).

## References

- [1] B. Adhikari, Backward perturbation and sensitivity analysis of structured polynomial eigenvalue problem, Ph.D. thesis, IIT Guwahati, Dept. of Mathematics, 2008.
- [2] B. Adhikari, R. Alam, Structured Procrustes problem, *Linear Algebra Appl.* 490 (2016) 145–161.
- [3] A. Albert, Conditions for positive and nonnegative definiteness in terms of pseudoinverses, *SIAM J. Appl. Math.* 17 (2) (1969) 434–440.
- [4] M.K. Baghel, N. Gillis, P. Sharma, On the non-symmetric semidefinite Procrustes problem, *Linear Algebra Appl.* 648 (2022) 133–159.
- [5] J.E. Brock, Optimal matrices describing linear systems, *AIAA J.* 6 (7) (1968) 1292–1296.
- [6] G. Cardano, T.R. Witmer, O. Ore, *The Rules of Algebra: Ars Magna*, vol. 685, Courier Corporation, 2007.
- [7] E. Carlen, Trace inequalities and quantum entropy: an introductory course, *Entropy Quantum* 529 (73–140) (2010) 146.
- [8] N. Gillis, P. Sharma, A semi-analytical approach for the positive semidefinite Procrustes problem, *Linear Algebra Appl.* 540 (2018) 112–137.
- [9] J.C. Gower, G.B. Dijkstra, *Procrustes Problems*, vol. 30, OUP, Oxford, 2004.
- [10] M. Grant, S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, Version 2.1, 2014.
- [11] M.C. Grant, S.P. Boyd, Graph implementations for nonsmooth convex programs, in: *Recent Advances in Learning and Control*, Springer, 2008, pp. 95–110.
- [12] B.F. Green, The orthogonal approximation of an oblique structure in factor analysis, *Psychometrika* 17 (4) (1952) 429–440.
- [13] P. Jingjing, W. Qingwen, P. Zhenyun, C. Zhencheng, Solution of symmetric positive semidefinite Procrustes problem, *Electron. J. Linear Algebra* 35 (2019) 543–554.
- [14] N. Krislock, J. Lang, J. Varah, D.K. Pai, H.P. Seidel, Local compliance estimation via positive semidefinite constrained least squares, *IEEE Trans. Robot.* 20 (6) (2004) 1007–1011.
- [15] N.G.B. Krislock, Numerical solution of semidefinite constrained least squares problems, Ph.D. thesis, University of British Columbia, 2003.
- [16] J. Lang, D.K. Pai, R.J. Woodham, Acquisition of elastic models for interactive simulation, *Int. J. Robot. Res.* 21 (8) (2002) 713–733.
- [17] Z. Li, L.H. Lim, Generalized matrix nearness problems, *SIAM J. Matrix Anal. Appl.* 44 (4) (2023) 1709–1730.
- [18] D.S. Mackey, N. Mackey, F. Tisseur, Structured mapping problems for matrices associated with scalar products. Part I: Lie and Jordan algebras, *SIAM J. Matrix Anal. Appl.* 29 (4) (2008) 1389–1410.
- [19] Y. Nesterov, A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ , *Dokl. Akad. Nauk SSSR* 269 (3) (1983) 543.
- [20] J. Nocedal, S.J. Wright, *Numerical Optimization*, Springer, 1999.
- [21] K.B. Petersen, M.S. Pedersen, The matrix cookbook, *Tech. Univ. Den.* 7 (15) (2008) 510.
- [22] T. Suffridge, T. Hayden, Approximation by a Hermitian positive semidefinite Toeplitz matrix, *SIAM J. Matrix Anal. Appl.* 14 (3) (1993) 721–734.
- [23] K.G. Woodgate, Least-squares solution of  $F = PG$  over positive semidefinite symmetric  $P$ , *Linear Algebra Appl.* 245 (1996) 171–190.