A semi-analytical approach for the positive semidefinite Procrustes problem

Nicolas Gillis* Punit Sharma*
Department of Mathematics and Operational Research
Faculté Polytechnique, Université de Mons
Rue de Houdain 9, 7000 Mons, Belgium

Abstract

The positive semidefinite Procrustes (PSDP) problem is the following: given rectangular matrices X and B, find the symmetric positive semidefinite matrix A that minimizes the Frobenius norm of $AX - B$. No general procedure is known that gives an exact solution. In this paper, we present a semi-analytical approach to solve the PSDP problem. First, we characterize completely the set of optimal solutions and identify the cases when the infimum is not attained. This characterization requires the unique optimal solution of a smaller PSDP problem where B is square and X is diagonal with positive diagonal elements. Second, we propose a very efficient strategy to solve the PSDP problem, combining the semi-analytical approach, a new initialization strategy and the fast gradient method. We illustrate the effectiveness of the new approach, which is guaranteed to converge linearly, compared to state-of-the-art methods.

1 Introduction

Given $n \times m$ matrices X and $B \in \mathbb{R}^{n,m}$, the corresponding positive semidefinite Procrustes (PSDP) problem is defined by

$$\inf_{A \in \mathcal{S}^n_+} \|AX - B\|_F^2, \qquad \text{where } X, B \in \mathbb{R}^{n,m}, \tag{$\mathcal{P}$}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix and $\mathbb{R}^{n,r}$ denotes the set of $n \times r$ real matrices with the special case $\mathbb{R}^n = \mathbb{R}^{n,1}$. The set of symmetric positive semidefinite matrices of size n is denoted by $\mathcal{S}^n_+$.

This problem occurs for example in the field of structure analysis [6] and in signal processing [17]. For an elastic structure, each column of X consists of generalized forces while each column of B consists of the corresponding displacements, for a set of m-measurements. From this data, it is possible to recover the so-called compliance matrix A that relates these column vectors by $AX = B$ and that must be symmetric positive definite. Such a compliance matrix may not exist for the available measurements and it is therefore desirable to find the matrix A that solves ($\mathcal{P}$) instead [22]. The PSDP problem was first introduced and studied by Allwright [4]; see also [21, 17, 22, 5] for more references.

In the simplest case when X is the identity matrix, the nearest positive semidefinite matrix to B in the Frobenius norm is given by $(C + H)/2$, where H is the symmetric polar factor of the matrix

*The authors acknowledge the support of the ERC (starting grant n° 679515). NG also acknowledges the support of the F.R.S.-FNRS (incentive grant for scientific research n° F.4501.16). Email: {nicolas.gillis, punit.sharma}@umons.ac.be.

$C = (B + B^T)/2$ [12]. Equivalently, the projection $\mathcal{P}_{\succeq}(B)$ of B onto the cone of semidefinite matrices is given by

$$\mathcal{P}_{\succeq}(B) = U\left(\max\left(\Gamma, 0\right)\right)U^T, \tag{1.1}$$

where $U\Gamma U^T$ is an eigenvalue decomposition of the symmetric matrix $\frac{B+B^T}{2}$.

The problem of finding the nearest Hermitian positive semidefinite matrix with a Toeplitz structure is studied in [17]. If the feasible set in $(\mathcal{P})$ is chosen to be the set of orthogonal matrices, then the problem is called the orthogonal Procrustes problem which arises in many applications such as computer vision, factor analysis, multidimensional scaling, and manifold optimization; see [8, 10, 16, 1] and references therein. On the other hand, the symmetric Procrustes problem, where the feasible set in $(\mathcal{P})$ is the set of symmetric matrices arise in applications such as determination of space craft altitudes and the stiffness matrix of an elastic structure [11, 6, 13].

In a recent work, Alam and Adhikari [2] have characterized and determined all solutions of the structured Procrustes problem analytically, where the feasible set in $(\mathcal{P})$ is either a Jordan algebra or a Lie algebra associated with an appropriate scalar product on $\mathbb{R}^n$ or $\mathbb{C}^n$.

We note that no general procedure is known for solving $(\mathcal{P})$ analytically. Many algorithmic solutions have been proposed in the literature, see for example [4, 22, 5].

In the following, $\|\cdot\|_2$ denotes the spectral norm of a vector or a matrix. For $A = A^T \in \mathbb{R}^{n,n}$, we denote $A \succ 0$ and $A \succeq 0$ if A is symmetric positive definite or symmetric positive semidefinite, respectively. By $\sigma_i(X)$ we denote the ith largest singular value of X. We use the acronym SVD for the singular value decomposition.

1.1 Contributions and outline of the paper

In section 2, we present some preliminary lemmas that will be repeatedly used in the paper. We also provide a simpler and shorter proof of the fact that the infimum of $(\mathcal{P})$ is always attained when X has rank n.

In section 3, we describe a semi-analytical approach to solve the PSDP problem $(\mathcal{P})$. We reduce the original problem $(\mathcal{P})$ into a smaller PSDP problem that always has a unique solution and whose size is equal to the rank of X. Assuming the solution for the subproblem is known, we derive the set of positive semidefinite matrices that either solve $(\mathcal{P})$ analytically when the infimum in $(\mathcal{P})$ is attained, or give arbitrary accurate approximations to the infimum in $(\mathcal{P})$ when the infimum is not attained. Further, in the case $\operatorname{rank}(X) = 1$, the subproblem admits a closed-form solution hence our derivations allow us to provide an analytical solution for $(\mathcal{P})$. Note that the fact that $(\mathcal{P})$ can be reduced to a smaller well-posed problem was already known [5]. However, the complete and explicit characterization of the set of optimal solutions of $(\mathcal{P})$ is, to the best of our knowledge, a new contribution.

In section 4, we describe the fast gradient method (FGM) applied to $(\mathcal{P})$. FGM is an optimal first-order method and, in the strongly convex case, is guaranteed to converge linearly with optimal rate among first-order methods. Therefore, combining the semi-analytical approach with FGM allows us to guarantee linear convergence of the objective function value with rate $(1 - 1/\kappa)$, where $\kappa = \frac{\sigma_1(X)}{\sigma_r(X)} \geq 1$ and $r = \operatorname{rank}(X)$. Note that Andersson and Elfving [5] had already introduced the reduction of $(\mathcal{P})$ but did not use it in their algorithms (also, they did not provide an explicit characterization of the solutions of $(\mathcal{P})$ based on the solution of the subproblem, and their first-order algorithm is not optimal). Moreover, to deal with ill-conditioned cases effectively (when κ is large), we propose a very effective initialization strategy for $(\mathcal{P})$ based on a recursive decomposition of the ill-conditioned problem into well-conditioned subproblems.

Finally in section 5, we present some numerical experiments illustrating the performance of the new proposed approach compared to state-of-the-art algorithms.

2 Preliminary results

Following the strategy used in [2] to derive analytic solutions of the structured Procrustes problem, we reduce the original problem $(\mathcal{P})$ to a smaller problem whose size equals the rank of X. The following lemma [3, Theorem 1] that gives an equivalent characterization for a positive semidefinite matrix will be repeatedly used in doing so.

Lemma 2.1 *Let the integer s be such that $0 < s < n$, and $R = R^T \in \mathbb{R}^{n,n}$ be partitioned as $R = \begin{bmatrix} B & C^T \\ C & D \end{bmatrix}$ with $B \in \mathbb{R}^{s,s}$, $C \in \mathbb{R}^{n-s,s}$ and $D \in \mathbb{R}^{n-s,n-s}$. Then $R \succeq 0$ if and only if*

1) *$B \succeq 0$,*

2) *$\ker(B) \subseteq \ker(C)$, and*

3) *$D - CB^{\dagger}C^T \succeq 0$, where $B^{\dagger}$ denotes the Moore-Penrose pseudoinverse of B.*

If X is of full rank then the existence of a unique solution to the problem $(\mathcal{P})$ is guaranteed by [22, Theorem 2.2]. We restate the result with an alternative simpler proof in a form that will be used in this paper.

Lemma 2.2 *Let $X, B \in \mathbb{R}^{n,m}$. If X has rank n, then the infimum of $(\mathcal{P})$ is attained for a unique $A \in \mathcal{S}^n_{\succeq}$.*

Proof. Let us define the continuous function $f : \mathcal{S}^n_{\succeq} \to \mathbb{R}$ such that $A \to f(A) = \|AX - B\|_F^2$. Then f is a strongly convex function as its Hessian is positive definite: in fact, since X has rank n we have $\frac{d^2 f}{dA^2} = \left(I_n \otimes XX^T\right) \succ 0$, where I_n denotes the identity matrix of size n and $\otimes$ denotes the Kronecker product. Each diagonal block XX^T in the Hessian corresponds to a row of A since $\|AX - B\|_F^2 = \sum_{i=1}^n \|A(i,:)X - B(i,:)\|_F^2$. Note that $\mathcal{S}^n_{\succeq}$ is a closed convex cone in the set of real symmetric matrices of size n, but not bounded. However, the set

$$\mathcal{B}^n_{\succeq} := \left\{A \in \mathcal{S}^n_{\succeq} \mid \|AX - B\|_F \leq \|0 - B\|_F = \|B\|_F\right\}$$

is closed and bounded in $\mathcal{S}^n_{\succeq}$. Since the zero matrix, that is, the matrix with all its entries equal to zero, is an element of $\mathcal{S}^n_{\succeq}$, replacing the feasible set in $(\mathcal{P})$ by $\mathcal{B}^n_{\succeq}$ does not change the infimum. Therefore, since f is continuous and bounded below ($f \geq 0$), its infimum is attained at some point $\hat{A} \in \mathcal{B}^n_{\succeq}$. Further the uniqueness of such a $\hat{A}$ follows from the strong convexity of f. □

In the next section we use the fact that the trace of product of two positive semidefinite matrices is nonnegative. The following elementary lemma gives more than that.

Lemma 2.3 *Let $P, Q \in \mathcal{S}^n_{\succeq}$. Then all eigenvalues of PQ are nonnegative.*

Proof. Let $L \in \mathcal{S}^n_{\succeq}$ be such that $P = L^{\frac{1}{2}}L^{\frac{1}{2}}$ then $PQ = L^{\frac{1}{2}}(L^{\frac{1}{2}}Q)$ and $L^{\frac{1}{2}}QL^{\frac{1}{2}}$ have the same nonzero eigenvalues. But then $L^{\frac{1}{2}}QL^{\frac{1}{2}} \in \mathcal{S}^n_{\succeq}$ as $Q \in \mathcal{S}^n_{\succeq}$ implies that all eigenvalues of $L^{\frac{1}{2}}(L^{\frac{1}{2}}Q) = PQ$ are nonnegative. □

We close the section with a result that will be used in obtaining solutions of the PSDP problem $(\mathcal{P})$ with the extremum properties of minimal rank, minimal Frobenius norm, or minimal spectral norm, and investigate their uniqueness.

Lemma 2.4 *Let the integer s be such that $0 < s < n$. Let $B \in \mathcal{S}^s_\succeq$ and $C \in \mathbb{R}^{n-s,s}$ be such that $\ker(B) \subseteq \ker(C)$. Define $\mathcal{D} := \{K \in \mathbb{R}^{n-s,n-s} : K - CB^\dagger C^T \succeq 0\}$ and define $f : \mathcal{D} \to \mathbb{R}^{n,n}$ by*

$$f(K) := \begin{bmatrix} B & C^T \\ C & K \end{bmatrix}.$$

Then the matrix $\hat{K} = CB^\dagger C^T \in \mathcal{D}$ is a solution of the minimal Frobenius norm problem $\min_{K\in\mathcal{D}} \|f(K)\|_F$, the minimal rank problem $\min_{K\in\mathcal{D}} \operatorname{rank}(f(K))$, and the minimal spectral norm problem $\min_{K\in\mathcal{D}} \|f(K)\|_2$. Moreover, for the minimal Frobenius norm and minimal rank problems, it is the unique solution.

Proof. Let $K \in \mathcal{D}$. Using Schur compliment of B in $f(K)$, we have

$$\operatorname{rank}(f(K)) = \operatorname{rank}\left(\begin{bmatrix} B & 0 \\ 0 & K - CB^\dagger C^T \end{bmatrix}\right) \geq \operatorname{rank}(B).$$

This implies that

$$\min_{K\in\mathcal{D}} \operatorname{rank}(f(K)) \geq \operatorname{rank}(B),$$

and the minimum is uniquely attained when $K = CB^\dagger C^T$. For the minimal norm problems, observe that

$$\inf_{K\in\mathcal{D}} \|f(K)\|_G = \inf_{K - CB^\dagger C^T \succeq 0} \left\| \begin{bmatrix} B & C^T \\ C & K \end{bmatrix} \right\|_G = \inf_{\Delta\succeq 0} \left\| \begin{bmatrix} B & C^T \\ C & \Delta + CB^\dagger C^T \end{bmatrix} \right\|_G, \qquad G = F \text{ or } 2.$$

For the Frobenius norm, we have

$$\begin{aligned} \inf_{K\in\mathcal{D}} \|f(K)\|_F^2 = \inf_{\Delta\succeq 0} \left\| \begin{bmatrix} B & C^T \\ C & \Delta + CB^\dagger C^T \end{bmatrix} \right\|_F^2 &= \|B\|_F^2 + 2\|C\|_F^2 + \inf_{\Delta\succeq 0} \|\Delta + CB^\dagger C^T\|_F^2 \\ &= \|B\|_F^2 + 2\|C\|_F^2 + \|CB^\dagger C^T\|_F^2, \end{aligned}$$

where the last equality holds as $\inf_{\Delta\succeq 0} \|\Delta + CB^\dagger C^T\|_F$ is attained uniquely at $\Delta = 0$ because $CB^\dagger C^T \succeq 0$. Similarly, for the 2-norm,

$$\begin{aligned} \inf_{K\in\mathcal{D}} \|f(K)\|_2 = \inf_{\Delta\succeq 0} \left\| \begin{bmatrix} B & C^T \\ C & \Delta + CB^\dagger C^T \end{bmatrix} \right\|_2 &= \inf_{\Delta\succeq 0} \left\| \underbrace{\begin{bmatrix} B & C^T \\ C & CB^\dagger C^T \end{bmatrix}}_{R} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \Delta \end{bmatrix}}_{\Delta_R} \right\|_2 \\ &= \inf_{\Delta\succeq 0} \sup_{x\in\mathbb{C}^n\setminus\{0\}} \frac{x^*(R+\Delta_R)x}{x^*x} \geq \sup_{x\in\mathbb{C}^n\setminus\{0\}} \frac{x^*Rx}{x^*x}, \end{aligned}$$

where the last inequality is due to the fact that $x^*\Delta_R x \geq 0$ for all $x \in \mathbb{C}^n$ as $\Delta_R \succeq 0$. Therefore infimum is attained when $\Delta = 0$. □

3 Semi-analytical solutions for the PSDP problem

In this section, we present a semi-analytic solution for problem $(\mathcal{P})$. It is semi analytic in the sense that we reduce the original problem $(\mathcal{P})$ into a smaller problem that always has a unique solution. Then assuming the solution for the subproblem is known, we characterize a family of positive semidefinite matrices that either solve the problem $(\mathcal{P})$ analytically or give arbitrary accurate approximations to the infimum in $(\mathcal{P})$.

Theorem 3.1 *Let $X, B \in \mathbb{R}^{n,m}$, and let $r = \operatorname{rank}(X)$. Let also $X = U\Sigma V^T$ be a singular value decomposition of X, where $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{n,n}$ with $U_1 \in \mathbb{R}^{n,r}$, $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{m,m}$ with $V_1 \in \mathbb{R}^{m,r}$, and $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n,m}$ with $\Sigma_1 \in \mathbb{R}^{r,r}$. Then*

$$\inf_{A \in \mathcal{S}^n_\succeq} \|AX - B\|_F^2 = \min_{A_{11} \in \mathcal{S}^r_\succeq} \|A_{11}\Sigma_1 - U_1^T B V_1\|_F^2 + \|BV_2\|_F^2. \tag{3.1}$$

Further, let $\hat{A}_{11} \in \mathcal{S}^r_\succeq$ be such that

$$\hat{A}_{11} = \operatorname{argmin}_{A_{11} \in \mathcal{S}^r_\succeq} \|A_{11}\Sigma_1 - U_1^T B V_1\|_F^2. \tag{3.2}$$

The following holds.

1) *If $\ker(\hat{A}_{11}) \subseteq \ker(U_2^T B V_1 \Sigma_1^{-1})$, then A_{opt} attains the infimum in (3.1) if and only if*

$$A_{opt} := U_1 \hat{A}_{11} U_1^T + U_2 (U_2^T B V_1 \Sigma_1^{-1}) U_1^T + U_1 (U_2^T B V_1 \Sigma_1^{-1})^T U_2^T + U_2 K U_2^T, \tag{3.3}$$

where $K \in \mathbb{R}^{n-r,n-r}$ is such that $K - (U_2^T B V_1 \Sigma_1^{-1}) \hat{A}_{11}^\dagger (U_2^T B V_1 \Sigma_1^{-1})^T \succeq 0$.

2) *Otherwise, the infimum in (3.1) is not attained. Let $\operatorname{rank}(\hat{A}_{11}) = s < r$ and let $\epsilon > 0$ be sufficiently small. Let $\hat{A}_{11} = \begin{bmatrix} \hat{U}_1 & \hat{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{U}_1^T \\ \hat{U}_2^T \end{bmatrix}$ be a singular value decomposition of $\hat{A}_{11}$, where $\hat{U}_1 \in \mathbb{R}^{r,s}$ and $\hat{\Sigma}_1 \in \mathbb{R}^{s,s}$. Define*

$$\hat{A}_{11}^\epsilon := \begin{bmatrix} \hat{U}_1 & \hat{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \Upsilon \end{bmatrix} \begin{bmatrix} \hat{U}_1^T \\ \hat{U}_2^T \end{bmatrix}, \tag{3.4}$$

where $\Upsilon \in \mathbb{R}^{r-s,r-s}$ is a diagonal matrix with diagonal entries each equal to $\frac{\epsilon}{\beta}$, where

$$\beta = 4\sqrt{(r-s)} \|\Sigma_1\|_F \|\hat{A}_{11}\Sigma_1 - U_1^T B V_1\|_F.$$

Define

$$A_\epsilon := U_1 \hat{A}_{11}^\epsilon U_1^T + U_2 (U_2^T B V_1 \Sigma_1^{-1}) U_1^T + U_1 (U_2^T B V_1 \Sigma_1^{-1})^T U_2^T + U_2 K_\epsilon U_2^T, \tag{3.5}$$

where $K_\epsilon \in \mathbb{R}^{n-r,n-r}$ is such that $K_\epsilon - (U_2^T B V_1 \Sigma_1^{-1}) (\hat{A}_{11}^\epsilon)^{-1} (U_2^T B V_1 \Sigma_1^{-1})^T \succeq 0$. Then $A_\epsilon \in \mathcal{S}^n_\succeq$ and

$$\|A_\epsilon X - B\|_F^2 < \inf_{A \in \mathcal{S}^n_\succeq} \|AX - B\|_F^2 + \epsilon.$$

Proof. Let $A \in \mathbb{R}^{n,n}$ and set

$$\hat{A} := U^T A U = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}, \tag{3.6}$$

where $A_{11} \in \mathbb{R}^{r,r}$, $A_{21} \in \mathbb{R}^{n-r,r}$ and $A_{22} \in \mathbb{R}^{n-r,n-r}$. Then $A \succeq 0$ if and only if $\hat{A} \succeq 0$. By Lemma 2.1,

$\hat{A} \succeq 0$ if and only if $A_{11} \succeq 0$, $\ker(A_{11}) \subseteq \ker(A_{21})$ and $A_{22} - A_{21}A_{11}^{\dagger}A_{21}^T \succeq 0$. Thus we have

$$\begin{aligned}
\|AX - B\|_F^2 &= \|U^T AUU^T X - U^T B\|_F^2 = \|\hat{A}U^T X - U^T B\|_F^2 \\
&= \left\| \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} X - \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} B \right\|_F^2 \\
&= \left\| \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} U_1^T X \\ 0 \end{bmatrix} - \begin{bmatrix} U_1^T B \\ U_2^T B \end{bmatrix} \right\|_F^2 \\
&= \|A_{11}U_1^T X - U_1^T B\|_F^2 + \|A_{21}U_1^T X - U_2^T B\|_F^2 \\
&= \left\| A_{11}U_1^T \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} - U_1^T B \begin{bmatrix} V_1 & V_2 \end{bmatrix} \right\|_F^2 + \\
&\quad \left\| A_{21}U_1^T \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} - U_2^T B \begin{bmatrix} V_1 & V_2 \end{bmatrix} \right\|_F^2 \\
&= \left\| \begin{bmatrix} A_{11}\Sigma_1 - U_1^T BV_1 & -U_1^T BV_2 \end{bmatrix} \right\|_F^2 + \left\| \begin{bmatrix} A_{21}\Sigma_1 - U_2^T BV_1 & -U_2^T BV_2 \end{bmatrix} \right\|_F^2 \\
&= \|A_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \|U_1^T BV_2\|_F^2 + \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 + \|U_2^T BV_2\|_F^2 \\
&= \|A_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 + \|BV_2\|_F^2, \qquad (3.7)
\end{aligned}$$

where in the last equality we used $\|U_1^T BV_2\|_F^2 + \|U_2^T BV_2\|_F^2 = \|BV_2\|_F^2$ since the Frobenius norm is unitarily invariant. Thus taking the infimum in (3.7) over $\mathcal{S}_{\succeq}^n$, we obtain

$$\begin{aligned}
&\inf_{A \in \mathcal{S}_{\succeq}^n} \|AX - B\|_F^2 \\
&\quad = \inf_{A_{11} \succeq 0,\, \ker(A_{11}) \subseteq \ker(A_{21}),\, A_{22} - A_{21}A_{11}^{\dagger}A_{21}^T \succeq 0} \|U\hat{A}U^T X - B\|_F^2 \\
&\quad = \inf_{A_{11} \succeq 0,\, \ker(A_{11}) \subseteq \ker(A_{21})} \|A_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 + \|BV_2\|_F^2, \qquad (3.8)
\end{aligned}$$

which implies that the infimum does not depend on the A_{22} block of matrix $\hat{A}$ in (3.6). Again from (3.8), we have

$$\begin{aligned}
\inf_{A \in \mathcal{S}_{\succeq}^n} \|AX - B\|_F^2 &= \inf_{A_{11} \succeq 0,\, \ker(A_{11}) \subseteq \ker(A_{21})} \|A_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 + \|BV_2\|_F^2 \\
&\geq \inf_{A_{11} \in \mathcal{S}_{\succeq}^r} \|A_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \inf_{A_{21} \in \mathbb{R}^{n-r,r}} \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 + \|BV_2\|_F^2 \qquad (3.9) \\
&= \|\hat{A}_{11}\Sigma_1 - U_1^T BV_1\|_F^2 + \|BV_2\|_F^2, \qquad (3.10)
\end{aligned}$$

where the last equality follows by Lemma 2.2 since (i) the first infimum in the right hand side of (3.9) is attained at a unique $\hat{A}_{11} \in \mathcal{S}_{\succeq}^r$ (Lemma 2.2), and (ii) the value of the second infimum is zero, that is, $\inf_{A_{21} \in \mathbb{R}^{n-r,r}} \|A_{21}\Sigma_1 - U_2^T BV_1\|_F^2 = 0$, which is attained at $A_{21} = U_2^T BV_1\Sigma_1^{-1}$. In order to show that equality in (3.9) holds instead of inequality, we consider two cases.

Case-1: When $\ker(\hat{A}_{11}) \subseteq \ker(U_2^T BV_1\Sigma_1^{-1})$. In this case, taking $A_{11} = \hat{A}_{11}$ and $A_{21} = U_2^T BV_1\Sigma_1^{-1}$ in (3.6), we define

$$A_{opt} := U \begin{bmatrix} \hat{A}_{11} & (U_2^T BV_1\Sigma_1^{-1})^T \\ U_2^T BV_1\Sigma_1^{-1} & K \end{bmatrix} U^T, \qquad (3.11)$$

for some K such that $K-(U_2^TBV_1\Sigma_1^{-1})\hat{A}_{11}^{\dagger}(U_2^TBV_1\Sigma_1^{-1})^T \succeq 0$, which upon simplification yields (3.3). By Lemma 2.1, we have $A_{opt} \in \mathcal{S}_{\succeq}^n$ and, in view of (3.7), it satisfies

$$\|A_{opt}X-B\|_F^2 = \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \|BV_2\|_F^2. \tag{3.12}$$

This implies equality in (3.9) in the case when $\ker(\hat{A}_{11}) \subseteq \ker(U_2^TBV_1\Sigma_1^{-1})$. This completes the proof of 1).

Case-2: If suppose $\ker(\hat{A}_{11}) \nsubseteq \ker(U_2^TBV_1\Sigma_1^{-1})$, then let $0<\epsilon<\min\{1, \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2\}$, and let $\hat{A}_{11}^{\epsilon}$ be as defined in (3.4). Then we have

$$\begin{aligned}\|\hat{A}_{11}^{\epsilon}\Sigma_1 - U_1^TBV_1\|_F^2 &= \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1 + \hat{U}_2\Upsilon\hat{U}_2^T\Sigma_1\|_F^2 \\ &\leq \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \|\Upsilon\|_F^2\|\Sigma_1\|_F^2 + 2\|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F\|\Upsilon\|_F\|\Sigma_1\|_F \\ &< \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \epsilon,\end{aligned}$$

where the last inequality follows by using the fact that $\epsilon < \min\{1, \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2\}$ and using $\|\Upsilon\|_F = \frac{\epsilon}{4\|\Sigma_1\|_F\|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F}$. Thus we have

$$\|\hat{A}_{11}^{\epsilon}\Sigma_1 - U_1^TBV_1\|_F^2 < \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \epsilon. \tag{3.13}$$

Note that $\hat{A}_{11}^{\epsilon}$ is nonsingular. Thus again using $\hat{A}_{11}^{\epsilon}$ and by setting $A_{21} = U_2^TBV_1\Sigma_1^{-1}$ in (3.6), we define

$$A_{\epsilon} := U\begin{bmatrix} \hat{A}_{11}^{\epsilon} & (U_2^TBV_1\Sigma_1^{-1})^T \\ U_2^TBV_1\Sigma_1^{-1} & K_{\epsilon}\end{bmatrix}U^T, \tag{3.14}$$

for some K_{ϵ} such that $K_{\epsilon} - (U_2^TBV_1\Sigma_1^{-1})(\hat{A}_{11}^{\epsilon})^{-1}(U_2^TBV_1\Sigma_1^{-1})^T \succeq 0$. This upon simplification yields (3.5), and, by construction, $A_{\epsilon} \in \mathcal{S}_{\succeq}^n$. Thus in view of (3.7) and (3.13), we have

$$\begin{aligned}\|A_{\epsilon}X-B\|_F^2 &= \|\hat{A}_{11}^{\epsilon}\Sigma_1 - U_1^TBV_1\|_F^2 + \|BV_2\|_F^2 \\ &< \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \|BV_2\|_F^2 + \epsilon.\end{aligned} \tag{3.15}$$

Thus as ϵ tends to zero, from (3.10) and (3.15), we get the equality in (3.9). Hence

$$\inf_{A\in\mathcal{S}_{\succeq}^n}\|AX-B\|_F^2 = \|\hat{A}_{11}\Sigma_1 - U_1^TBV_1\|_F^2 + \|BV_2\|_F^2.$$

This infimum is attained when $\epsilon = 0$, but then by Lemma 2.1 $A_{\epsilon} \notin \mathcal{S}_{\succeq}^n$ as $\ker(\hat{A}_{11}) \nsubseteq \ker(U_2^TBV_1\Sigma_1^{-1})$. Thus the uniqueness of $\hat{A}_{11}$ and the fact that $\inf_{A_{21}\in\mathbb{R}^{n-r,r}}\|A_{21}\Sigma_1 - U_2^TBV_1\|_F^2 = 0$ implies that the infimum is not attained. This completes the proof of 2). □

Remark 3.2 In view of Theorem 3.1, a necessary and sufficient condition for the infimum to be attained in $(\mathcal{P})$ is that $\ker(\hat{A}_{11}) \subseteq \ker(U_2^TBV_1\Sigma_1^{-1})$. This condition coincides with the result [22, Theorem 2.2]. However, we provide a completely different and relatively simpler proof, and describe the analytic solutions of $(\mathcal{P})$ depending on $\hat{A}_{11}$.

Note that when $\mathrm{rank}(X) = n$, U_2 is an n-by-0 empty matrix hence $\ker(U_2^TBV_1\Sigma_1^{-1})$ is the full space so that the condition $\ker(\hat{A}_{11}) \subseteq \ker(U_2^TBV_1\Sigma_1^{-1})$ for the infimum to be attained is always met. Note also that when $\hat{A}_{11} \succ 0$ we have $\ker(\hat{A}_{11}) = \{0\}$ hence the condition $\ker(\hat{A}_{11}) \subseteq \ker(U_2^TBV_1\Sigma_1^{-1})$ is always met.

Using Lemma 2.4 in (3.11) and in (3.14) for matrices A_{opt} and A_ϵ, we can characterize the minimum norm solutions of $(\mathcal{P})$.

Corollary 3.3 *Let $X, B \in \mathbb{R}^{n,m}$, and let $r = \text{rank}(X)$. Let also U_1, U_2, V_1, V_2, Σ_1 and $\hat{A}_{11}$ be as in Theorem 3.1.*

1) *If $\ker(\hat{A}_{11}) \subseteq \ker(U_2^T B V_1 \Sigma_1^{-1})$, then A_{opt} in (3.3) with $K = (U_2^T B V_1 \Sigma_1^{-1})\hat{A}_{11}^{\dagger}(U_2^T B V_1 \Sigma_1^{-1})^T$ is the unique solution of the problem (3.1) with minimal rank, minimal Frobenius norm and minimal spectral norm.*

2) *Otherwise, for a given $\epsilon > 0$, the matrix A_ϵ in (3.5) with $K_\epsilon = (U_2^T B V_1 \Sigma_1^{-1})(\hat{A}_{11}^{\epsilon})^{-1}(U_2^T B V_1 \Sigma_1^{-1})^T$ is the unique matrix in $\mathcal{S}_{\preceq}^n$ with minimal rank, minimal Frobenius norm and minimal spectral norm, such that*

$$\|A_\epsilon X - B\|_F^2 < \inf_{A \in \mathcal{S}_{\preceq}^n} \|AX - B\|_F^2 + \epsilon.$$

In the following theorem, we show that if $(BX^T + XB^T) \preceq 0$ then computing the exact value of the infimum in $(\mathcal{P})$ does not require the solution $\hat{A}_{11}$ of the subproblem (3.2) as in Theorem 3.1.

Theorem 3.4 *Let $X, B \in \mathbb{R}^{n,m}$, and let $r = \text{rank}(X)$. Let also U_1, U_2, V_1, V_2 and Σ_1 be as defined in Theorem 3.1. If $U_1^T(BX^T + XB^T)U_1 \preceq 0$, then*

$$\inf_{A \in \mathcal{S}_{\preceq}^n} \|AX - B\|_F^2 = \|U_1^T B V_1\|_F^2 + \|BV_2\|_F^2, \tag{3.16}$$

and it is not attained for any $A \in \mathcal{S}_{\preceq}^n$. In this case, let $\epsilon > 0$ be sufficiently small and let $A_{11}^{\epsilon} \in \mathbb{R}^{r,r}$ be a diagonal matrix with diagonal entries each equal to $\frac{\epsilon}{\alpha}$, where $\alpha = 4\sqrt{n}\|\Sigma_1\|_F\|U_1^T B V_1\|_F$. Define

$$A_\epsilon := U_1 A_{11}^{\epsilon} U_1^T + U_2(U_2^T B V_1 \Sigma_1^{-1})U_1^T + U_1(U_2^T B V_1 \Sigma_1^{-1})^T U_2^T + U_2 K_\epsilon U_2^T, \tag{3.17}$$

where $K_\epsilon \in \mathbb{R}^{n-r,n-r}$ is such that $K_\epsilon - (U_2^T B V_1 \Sigma_1^{-1})(A_{11}^{\epsilon})^{-1}(U_2^T B V_1 \Sigma_1^{-1})^T \succeq 0$. Then $A_\epsilon \in \mathcal{S}_{\preceq}^n$ and

$$\|A_\epsilon X - B\|_F^2 < \inf_{A \in \mathcal{S}_{\preceq}^n} \|AX - B\|_F^2 + \epsilon. \tag{3.18}$$

Proof. We only give a proof of (3.16) and skip the proof of (3.18) as it is similar to the proof of 2) in Theorem 3.1. In Theorem 3.1 we proved that

$$\inf_{A \in \mathcal{S}_{\preceq}^n} \|AX - B\|_F^2 = \min_{A_{11} \in \mathcal{S}_{\preceq}^r} \|A_{11}\Sigma_1 - U_1^T B V_1\|_F^2 + \|BV_2\|_F^2. \tag{3.19}$$

Observe that for any $A_{11} \in \mathcal{S}_{\preceq}^r$, we have

$$\begin{aligned}
\|A_{11}\Sigma_1 - U_1^T B V_1\|_F^2 &= \text{trace}\left((A_{11}\Sigma_1 - U_1^T B V_1)^T(A_{11}\Sigma_1 - U_1^T B V_1)\right) \\
&= \|A_{11}\Sigma_1\|_F^2 + \|U_1^T B V_1\|_F^2 - \text{trace}\left(U_1^T B V_1 \Sigma_1 A_{11} + \Sigma_1 V_1^T B^T U_1 A_{11}\right) \\
&= \|A_{11}\Sigma_1\|_F^2 + \|U_1^T B V_1\|_F^2 - \text{trace}\left(U_1^T(BX^T + XB^T)U_1 A_{11}\right) \\
&\geq \|U_1^T B V_1\|_F^2,
\end{aligned} \tag{3.20}$$

where the third equality follows by using the fact that $X = U_1\Sigma_1 V_1^T$. The last inequality in (3.20) follows by using Lemma 2.3 since trace $\left(U_1^T(BX^T + XB^T)U_1 A_{11}\right) \leq 0$ as $-U_1^T(BX^T + BX^T)U_1 \succeq 0$ and $A_{11} \succeq 0$. Therefore from (3.20), we obtain

$$\min_{A_{11} \in \mathcal{S}^r_{\succeq}} \|A_{11}\Sigma_1 - U_1^T B V_1\|_F^2 \geq \|U_1^T B V_1\|_F^2,$$

and equality holds when $A_{11} = 0$. Plugging this in (3.19), we obtain

$$\inf_{A \in \mathcal{S}^n_{\succeq}} \|AX - B\|_F^2 = \|U_1^T B V_1\|_F^2 + \|BV_2\|_F^2. \tag{3.21}$$

Using the arguments similar to that of Case-2 in Theorem 3.1, it follows that infimum in (3.21) is not attained and for a sufficiently small $\epsilon > 0$, A_ϵ in (3.17) satisfies (3.18). This completes the proof. □

Remark 3.5 When $r = \text{rank}(X) < n$, Theorem 3.1 reduces the original problem $(\mathcal{P})$ to a smaller problem of the form

$$\min_{\widetilde{A} \succeq 0} \|\widetilde{A}\Sigma - \widetilde{B}\|_F^2,$$

where $\Sigma \in \mathcal{S}^r_{\succ}$ is a diagonal matrix with positive diagonal entries, and then a unique solution is guaranteed by Lemma 2.2. However, finding an analytic solution to the subproblem is a challenging task and still an open problem. We will discuss in Section 4.1 the use of an optimal first-order method to solve this problem.

3.1 Computational cost of the semi-analytical approach

In view of Theorem 3.1, we have that the semi-analytical approach completes in a three steps. The first step is to compute the singular value decomposition of X, which takes $O(\max(m,n)\min(m,n)^2)$ floating point operations [19]. The second step is to compute the solution $\hat{A}_{11}$ of the subproblem (3.2), and cost of this depends on the method used to solve it (see Section 4.1). The third step is to form A_{opt}. This involves matrix-matrix multiplication and costs $O(n^2 r)$.

3.2 Analytical solution for $\text{rank}(X) = 1$

We note that the case $\text{rank}(X) = 1$ differs from the case $\text{rank}(X) > 1$ because in that case the subproblem (3.2) has a closed-form solution:

$$\hat{A}_{11} = \max\left(0, U_1^T B V_1 \Sigma_1^{-1}\right) \in \mathbb{R}.$$

Therefore, we can provide a complete analytical characterization of the set of optimal solution of $(\mathcal{P})$ in this particular case. Although this follows directly from Theorem 3.1 and the observation above, we state the result here in the case $\text{rank}(X) = 1$ for the sake of completeness.

The rank-one case is also important in solving the particular case when X and B are vectors, that is, when one is looking for $A \succeq 0$ such that $\|Ax - b\|_2$ is minimized where x and b are vectors.

Theorem 3.6 *Let $X, B \in \mathbb{R}^{n,m}$ be such that $\text{rank}(X) = 1$. Let $X = U\Sigma V^T$ be a singular value decomposition of X, where $U = [u \;\; U_1] \in \mathbb{R}^{n,n}$ with $u \in \mathbb{R}^n$, $V = [v \;\; V_1] \in \mathbb{R}^{m,m}$ with $v \in \mathbb{R}^m$, and $\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n,m}$ with $\sigma > 0$. Then the following hold.*

1) *If* $u^T Bv > 0$, *then*

$$\inf_{A \in \mathcal{S}^n_+} \|AX - B\|_F = \|BV_1\|_F,$$

and A_{opt} *attains the infimum if and only if*

$$A_{opt} = \sigma^{-1}\left((u^T Bv)uu^T + U_1 U_1^T Bvu^T + uv^T B^T U_1 U_1^T\right) + U_1 K U_1^T, \tag{3.22}$$

for some matrix K *such that* $K - \frac{1}{\sigma u^T Bv} U_1^T (Bv)(Bv)^T U_1 \succeq 0$. *In particular,* A_{opt} *can be chosen to be of rank one by choosing* $K = \frac{1}{\sigma u^T Bv} U_1^T (Bv)(Bv)^T U_1$.

2) *If* $u^T Bv \leq 0$, *then*

$$\inf_{A \in \mathcal{S}^n_+} \|AX - B\|_F^2 = \|u^T Bv\|_F^2 + \|BV_1\|_F^2. \tag{3.23}$$

Further, if $U_1^T Bv = 0$, *then the infimum in* (3.23) *is attained by a matrix* A_{opt} *of the form* (3.22). *If* $U_1^T Bv \neq 0$, *then the infimum in* (3.23) *is not attained. In the later case for any arbitrary small* $\epsilon > 0$, *choose* $n_0 \in \mathbb{N}$ *such that* $\frac{\sigma^2}{n_0^2} - 2\sigma \frac{u^T Bv}{n_0} < \epsilon$ *and define*

$$A_{n_0} = \sigma^{-1}\left(\frac{1}{n_0} uu^T + U_1 U_1^T Bvu^T + uv^T B^T U_1 U_1^T\right) + U_1 K_{n_0} U_1^T,$$

for some K_{n_0} *with* $K_{n_0} - \frac{n_0}{\sigma^2} U_1^T (Bv)(Bv)^T U_1 \succeq 0$. *Then* $A_{n_0} \succeq 0$ *and*

$$\|A_{n_0} X - B\|_F^2 < \inf_{A \succeq 0} \|AX - B\|_F^2 + \epsilon.$$

Moreover, A_{n_0} *can be chosen to be of rank one by choosing* $K_{n_0} = \frac{n_0}{\sigma^2} U_1^T (Bv)(Bv)^T U_1$.

4 An algorithmic solution to the PSDP problem

In this section, we first describe the fast gradient method to solve $(\mathcal{P})$ (Section 4.1). We then propose a new very efficient initialization strategy for $(\mathcal{P})$ when X is diagonal and ill-conditioned (Section 4.2). Finally, we explain the advantages of combining the semi-analytical approach, FGM and our new initialization strategy. In particular, this allows us to guarantee linear convergence for solving $(\mathcal{P})$ (Section 4.3).

4.1 Fast Gradient Method

In order to be able to solve large-scale problem $(\mathcal{P})$, say for n up to a 1000, it makes sense to use first-order methods. In this section, we describe the fast gradient method (FGM) applied to $(\mathcal{P})$; see Algorithm FGM. We choose FGM because it is simple to implement and it is an optimal first-order method for smooth convex optimization, that is, no first-order method can have a faster convergence rate. In fact, FGM is guaranteed to decrease the objective function value at sublinear rate $O(1/t^2)$ where t is the iteration number. Moreover, in the strongly convex case, when $\kappa = \frac{\sigma_1(X)}{\sigma_n(X)} > 0$, the decrease is guaranteed to be at linear rate $O((1 - 1/\kappa)^t)$. This is much faster than the standard gradient descent method, with respective rate of $O(1/t)$ and $O((1 - 1/\kappa^2)^t)$. We refer the reader to [14, 15] for more details on FGM.

The computational cost of FGM is $O(n^3 + n^2 m)$ operations per iteration. The most expensive steps are

Algorithm FGM Fast Gradient Method for $(\mathcal{P})$ [15, p.90]

Require: An initial guess $A \in \mathcal{S}^n_+$, number of iterations T (other stopping criteria can be used).

Ensure: An approximate solution $A \approx \operatorname{argmin}_{\tilde{A} \succeq 0} \|\tilde{A}X - B\|_F$.

1: $L = \sigma_1^2(X)$, $q = \frac{\sigma_n^2(X)}{L}$.
2: $\alpha_1 \in (0,1)$; $Y = A$.
3: **for** $k = 1 : T$ **do**
4: $\quad \hat{A} = A$. $\quad$ *% Keep the previous iterate in memory.*
5: $\quad G_Y = YXX^T - BX^T$. *% Compute the gradient, XX^T and BX^T can be pre-computed.*
6: $\quad A = \mathcal{P}_{\succeq 0}\left(Y - \frac{1}{L} G_Y\right)$. *% Projected gradient step.*
7: $\quad \alpha_{k+1} = \frac{1}{2}\left(q - \alpha_k^2 + \sqrt{(q-\alpha_k^2)^2 + 4\alpha_k^2}\right)$, $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$.
8: $\quad Y = A + \beta_k\left(A - \hat{A}\right)$. *% Linear combination of the current and previous iterates.*
9: **end for**

- the computation of the singular values of X (step 1) requiring $O\left(\min(m,n)^2 \max(m,n)\right)$ operations [19].
- the computation of the gradient in $O(mn^2 + n^3)$ operations (step 5). Note that the n-by-n matrices XX^T and BX^T should be computed only once in which case the remaining computational cost for the gradient computation per iteration is $O(n^3)$.
- The projection step in $O(n^3)$ operations (step 6), as it requires the eigenvalue decomposition of a symmetric n-by-n matrix; see (1.1).

Denoting T the total number of iterations (typically, $T \geq 100$), the total computational cost of FGM is $O(Tn^3 + mn^2)$. In most cases, $Tn \geq m$ hence the computational cost of FGM will be $O(Tn^3)$.

4.2 Initialization

In this section, we present three initialization strategies.

4.2.1 Projection of the optimal unconstrained solution

In [5], the authors propose to use as an initialization the projection of the optimal solution of the unconstrained problem, that is,

$$\mathcal{P}_{\succeq}\left(\operatorname{argmin}_{A \in \mathbb{R}^{n \times n}} \|AX - B\|_F\right).$$

This initialization can sometimes perform well. However, it comes with no guarantee and provides very bad initialization in several situations, in particular for ill-conditioned problems; see Section 4.2.4 below for some examples.

4.2.2 Diagonal matrix

It is rather straightforward to compute the optimal solution of $(\mathcal{P})$ assuming that the matrix A is diagonal. In fact, the problem reduces to n independent least squares problem in one variables with

a nonnegativity constraint. We have

$$\|\mathrm{diag}(a)X - B\|_F^2 = \sum_{i=1}^{n} \|a_i X(i,:) - B(i,:)\|_F^2,$$

where $\mathrm{diag}(a)$ is a diagonal matrix whose diagonal elements are given by $a = [a_1, a_2, \ldots, a_n]^T \in \mathbb{R}^n$. The optimal solution for each subproblem is given by

$$a_i^* = \mathrm{argmin}_{a_i \geq 0} \|a_i X(i,:) - B(i,:)\|_F^2 = \max\left(0, \frac{B(i,:)X(i,:)^T}{\|X(i,:)\|_2^2}\right), \quad i = 1, 2, \ldots n.$$

4.2.3 Recursive decomposition for ill-conditioned and diagonal X

As we have explained previously, the convergence of first-order algorithms for $(\mathcal{P})$ will depend on the conditioning of X. Using the semi-analytical approach from Section 3, $(\mathcal{P})$ can be reduced to a problem where X is diagonal with positive diagonal elements. If X is well-conditioned, then FGM will converge fast and the initialization strategy does not play a crucial role. However, when X is ill-conditioned, FGM will be more sensitive to initialization as it converges slower.

Let us generalize the idea from the previous section by assuming that A is block diagonal instead of diagonal. For simplicity, let us assume that the diagonal matrix X is partitioned into to blocks X_1 and X_2 (this generalizes easily to more than two blocks):

$$\min_{A_1 \succeq 0, A_2 \succeq 0} \left\| \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} - \begin{pmatrix} B_1 & B_{12} \\ B_{21} & B_2 \end{pmatrix} \right\|_F.$$

This problem can be decoupled into two independent subproblems: for $i = 1, 2$,

$$\min_{A_i \succeq 0} \|A_i X_i - B_i\|_F.$$

Let us denote $\kappa(X_i)$ the condition number of X_i. If X_1 and X_2 are well conditioned, that is, $\max_i \kappa(X_i)$ is small, good approximate solutions to these subproblems can be obtained much faster than for the ill-conditioned X. Moreover, since X is diagonal, partitioning X into two blocks in order to minimize $\max_i \kappa(X_i)$ can be done as follows:

1) Sort the diagonal entries of X such that $x_1 \leq x_2 \leq \cdots \leq x_n$ (in our reduction, the diagonal entries of X are already sorted in nonincreasing order, since we use the standard SVD). This can be done in $O(n \log(n))$ operations.

2) Pick the partition $[1, 2, \ldots, k] \cup [k+1, k+2, \ldots, n]$ such that $\max\left(\frac{x_k}{x_1}, \frac{x_n}{x_{k+1}}\right)$ is minimized. This can be done in $O(n)$ operations.

Finally, it is straightforward to use this idea recursively as follows: As long as a block X_i is not well-conditioned, that is, $\kappa(X_i) > \kappa_M$ for some parameter κ_M (we use $\kappa_M = 100$), the block X_i is partitioned into two blocks as explained above. Once X has been partitioned into well-conditioned subblocks, we combine the diagonal initialization along with 100 iterations of FGM to approximately solve the (well-conditioned) subproblems. As we will see, not only this initialization will provide a solution with low initial error, it will also allow the FGM to converge faster to the optimal solution.

Note that this initialization is only applicable when X is diagonal hence can be used only in combination with the semi-analytical reduction described in Section 3.

4.2.4 Preliminary numerical experiments

Let us compare the different initialization strategies in ill-conditioned cases (well-conditioned cases are not so interesting since most algorithms will converge fast, being less sensitive to initialization), and let us consider

$$X = \text{diag}(1, 2, \ldots, 10, 20, \ldots 100, 200 \ldots, 1000, 2000, \ldots, 10000), \text{ with } \kappa(X) = 10^4.$$

Recall that the optimal solution of $(\mathcal{P})$ will be unique in this case since X has rank $n = 37$ (Lemma 2.2). We generate B in two different ways:

- Gaussian. Each entry is randomly generated following a normal distribution with mean 0 and standard deviation 1 (`randn(n)` in Matlab).
- Uniform. Each entry is randomly generated following a uniform distribution in the interval [0,1] (`rand(n)` in Matlab).

In each case, we generate 100 such matrices. We will compare four initializations:

- Zero. This is the trivial initialization $A = 0$.
- Unconstrained. This is the projection onto $\mathcal{S}^n_+$ of the optimal solution of the unconstrained version of $(\mathcal{P})$; see Section 4.2.1.
- Diagonal. This is the optimal diagonal solution of $(\mathcal{P})$; see Section 4.2.2.
- Recursive. This is the recursive decomposition approach described in Section 4.2.3.

We will only compare the initializations combined with Algorithm FGM because, as we will see in Section 5, it consistently performs well.

Table 4.1 gives the average initial error for the four initialization strategies, along with the standard deviation, for the 100 randomly generated matrices B of the two types. The best result is highlighted in bold. Figure 4.1 displays the evolution of error (average over 100 runs) for the different initializations using FGM.

Table 4.1: Average initial error $\|AX - B\|_F$ and standard deviation (in brackets) obtained by the different initialization approaches.

	Zero	Unconstrained	Diagonal	Recursive
Gaussian	36.97 (0.81)	8764.30 (1673.09)	36.73 (0.81)	**33.72** (0.78)
Uniform	21.37 (0.27)	5799.61 (726.11)	21.08 (0.26)	**17.45** (0.29)

We observe the following

- The initialization based on the projection of the optimal solution of the unconstrained problem performs very badly. In fact, the error is significantly larger than with the trivial zero initialization.
- The diagonal initialization performs slightly better than the zero initialization.
- Our recursive initialization performs best, both in term of initial error and for enabling FGM to converge faster to the unique optimal solution of $(\mathcal{P})$. In particular, the error obtained with the diagonal initializations using 1000 iterations of FGM is higher than the error obtained with the recursive initialization using around 50 iterations of FGM.

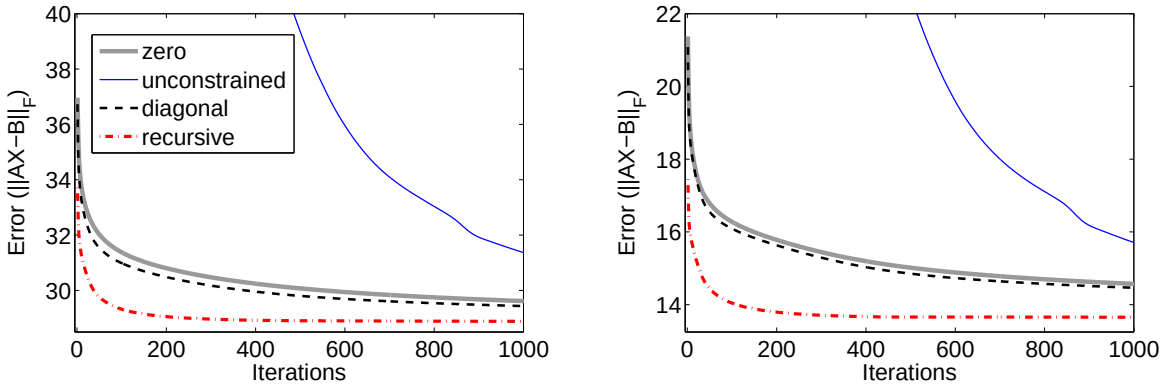


Figure 4.1: Evolution of the average error $\|AX - B\|_F$ over 100 randomly generated matrices B for the different initializations using FGM: Gaussian (left) and Uniform (right).

4.3 Combination of the semi-analytical approach, the recursive initialization and FGM

Combining the semi-analytical approach, the recursive initialization and FGM allow us to obtain a very efficient algorithm for $(\mathcal{P})$. The semi-analytical approach reduces the problem $(\mathcal{P})$ to a problem (i) involving only (possibly smaller) square matrices (where $m = n = \text{rank}(X)$) and (ii) that is strongly convex (since the 'new' X, denoted Σ_1 in our derivations, is diagonal with positive diagonal elements). This requires $O(\min(m,n)^2 \max(m,n) + n^2 r)$ operations; see Section 3.1.

The first advantage of the semi-analytical approach is that it reduces the computational cost of FGM in the following cases:

- If $m > n$, since the reduced problem only involves n-by-n matrices, this reduces computational cost per iteration of FGM from $\mathcal{O}(n^2 m)$ to $\mathcal{O}(n^3)$.
- If $r = \text{rank}(X) < n$, it reduces the problem in n^2 variables into a problem in r^2 variables. In particular, if $m < n$, this will always be the case since $r \leq m$.

The second advantage guarantees FGM to decrease the objective function value at linear rate $(1-1/\kappa_r)$ where $\kappa_r = \frac{\sigma_1(X)}{\sigma_r(X)}$ and $r = \text{rank}(X)$. This is, to the best of our knowledge, the first time an algorithm is proposed for $(\mathcal{P})$ with guaranteed linear convergence. (Moreover, recall this is an optimal rate for first-order methods.)

If X is ill-conditioned, that is, κ_r is large, the convergence could be slow. However, this is mitigated by our recursive decomposition strategy that solves well-conditioned subproblems to initialize the ill-conditioned one.

5 Numerical Experiments

In this section, we compare the following algorithms:

- Gradient. The projected gradient method applied on $(\mathcal{P})$ –this is FGM using $\beta_k = 0$ at each step (that is, $A = Y$ in Algorithm FGM). This method will serve as a baseline.

- FGM. The fast projected gradient method applied on $(\mathcal{P})$; see Algorithm FGM.
- ParTan. This is the method proposed in [5] and referred to as 'Parallel tangents'. This algorithm is rather similar to FGM but does not guarantee the optimal convergence rate. It can be seen as a heuristic variant of FGM where β_k is chosen as to minimize $\|YX - B\|_F$ without the PSD constraint, where $Y = A + \beta_k \hat{A}$ with A the current iterate and $\hat{A}$ the previous iterate; see Algorithm FGM. If this step does not decrease the objective function, then $\beta_k = 0$ is chosen (that is, a standard gradient step is used). Note that the computation of the $\beta_k's$ makes ParTan computationally slightly more expensive than FGM and Gradient.
- AN-FGM. This is the combination of the semi-analytical approach, reducing the problem to the case where X is diagonal with positive diagonal elements, and then using FGM on this reduced problem. We use the recursive initialization described in Section 4.2.3.

Note that we could combine the analytical approach and the recursive initialization with any other method. We choose FGM because it guarantees linear convergence, although it performs similarly as ParTan (see the numerical experiments below). For the first three algorithms, we use the diagonal initialization.

In these numerical experiments, we try as much as possible to cover all the different scenarios: we test for $m = n$, $m < n = 2m$ and $n < m = 2n$. In all cases, the matrix B is generated in the same way: each entry is randomly generated following a normal distribution with mean 0 and standard deviation 1 (`randn(m,n)` in Matlab). For the matrix X, we consider three cases

1) Well-conditioned. Each entry is randomly generated following a normal distribution with mean 0 and standard deviation 1 (`randn(m,n)` in Matlab).
2) Ill-conditioned. Let (U, Σ, V) be the compact SVD of a matrix generated as in the well-conditioned case. Then we generate $X = U\Lambda V$ where Λ is a diagonal matrix such that $\Lambda(i,i) = \alpha^{i-1}$ and $\alpha^{\min(m,n)-1} = 10^6 = \kappa(X)$.
3) Rank deficient. We perform the SVD (U, Σ, V) of a matrix generated as in the well-conditioned case, set the $r = \min(m,n)/2$ smallest singular values of Σ to zero to obtain Σ', and then compute $X = U\Sigma' V^T$ so that $\text{rank}(X) = \min(m,n)/2$.

The Matlab code is available from `https://sites.google.com/site/nicolasgillis/`. All tests are preformed using Matlab on a laptop Intel CORE i5-3210M CPU @2.5GHz 2.5GHz 6Go RAM.

For each experiment, we generate 10 such matrices and display the average results. For each algorithm, we perform 1000 iterations.

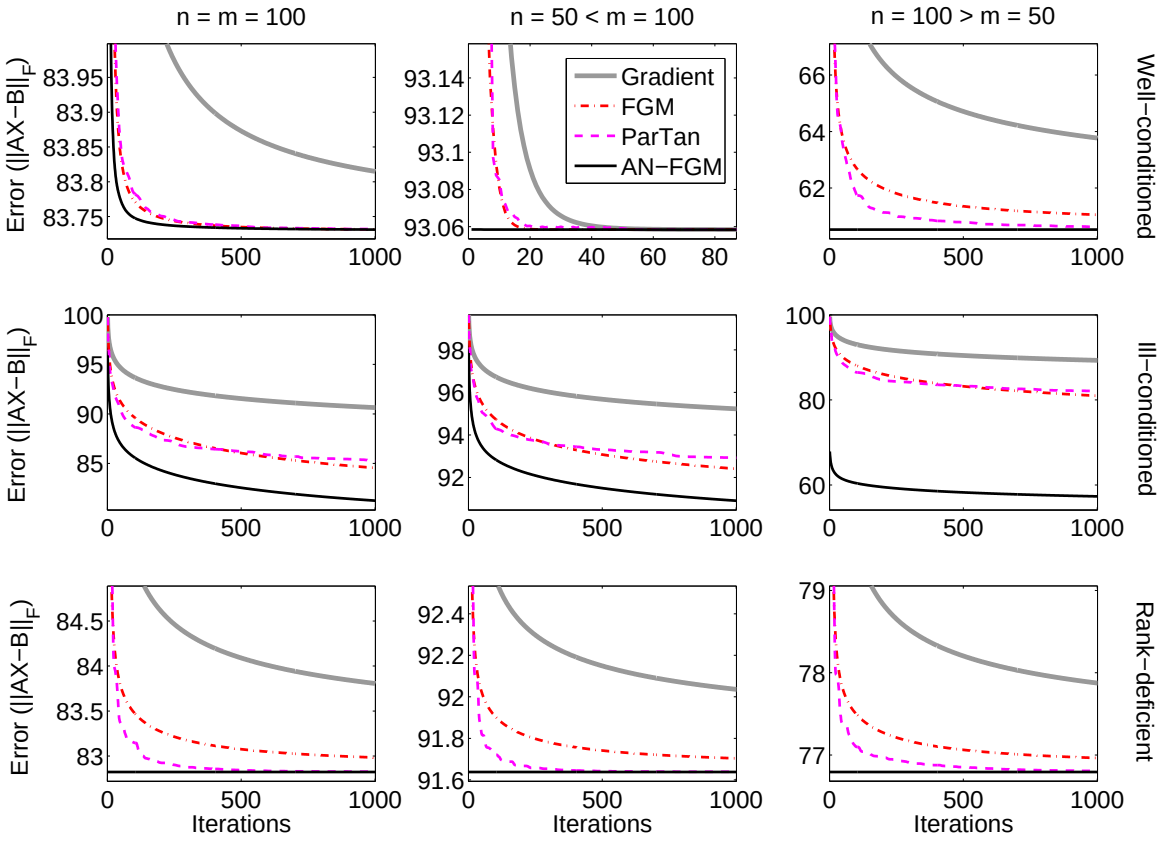
Gradient, FGM and ParTan require the same number of operations per iteration, that is, $O(n^3)$, assuming XX^T and BX^T are precomputed. After the preprocessing performed by our semi-analytical approach, the number of operations per iteration of AN-FGM is $O(r^3)$ where $r = \text{rank}(X) \leq n$. Therefore, in the cases $r \ll n$, the iterations cost of AN-FGM will be significantly smaller.

Figure 5.1 displays the evolution of the relative error in percent, that is,

$$\text{relative error } (\%) = 100 \frac{\|AX - B\|_F}{\|B\|_F},$$

for each algorithm in each of the nine cases described above. Table 5.1 reports the computational time required by each algorithm to perform the 1000 iterations. In bold we indicate the cases when

Figure 5.1: Evolution of the average relative error $100\frac{\|AX-B\|_F}{\|B\|_F}$ for the different algorithms.



AN-FGM is significantly faster than the other approaches, because of the dimension reduction of the problem.

Table 5.1: Computational times for the different algorithms to perform 1000 iterations.

		Gradient	FGM	ParTan	AN-FGM
Well-conditioned	$m = n = 100$	6.71	6.57	8.16	7.01
	$m = 2n = 100$	1.48	1.49	2.13	1.52
	$n = 2m = 100$	5.58	5.59	6.31	**1.45**
Ill-conditioned	$m = n = 100$	5.87	5.67	7.49	5.64
	$m = 2n = 100$	1.37	1.32	2.00	1.33
	$n = 2m = 100$	5.49	5.20	6.75	**1.28**
Rank-deficient	$m = n = 100$	6.22	6.56	7.27	**1.58**
	$m = 2n = 100$	1.67	1.42	2.22	**0.34**
	$n = 2m = 100$	5.33	5.33	6.55	**0.36**

We observe the following:

- In terms of computational time, as mentioned above, AN-FGM will be faster when $\text{rank}(X) \ll n$: this happens when X is rank-deficient, and when $m = n/2$ –these are the bold results in Table 5.1. In all other cases, all algorithms have roughly the same computational cost, namely $O(Tn^3)$ where T is the number of iterations (here $T = 1000$); see the discussion in Section 4.1.

- In all cases, the gradient method performs the worse. This is not surprising since it only uses the gradient information of the current iterate as opposed to FGM and ParTan (and it is not an optimal first-order method).

- For well-conditioned X:

  For $m = n = 100$, FGM, ParTan and AN-FGM perform similarly. The reason is that the semi-analytical approach cannot reduce the problem. The slight advantage of AN-FGM compared to FGM and ParTan comes from the fact the the recursive initialization already computes 100 iterations of FGM.

  For $n = 50 < m = 100$, all algorithms converge very fast, within 100 iterations: the reason is that the condition number of X is much smaller than in the case $n = m = 100$ (the average condition number of a 50-by-100 Gaussian matrix is below 10, while for a 100-by-100Gaussian matrix it is above 1000).

  For $n = 100 > m = 50$, AN-FGM performs best because it reduces the dimension of the problem from 100 to 50 (hence also reducing the computational cost per iteration; see Table 5.1).

  To summarize, for well-conditioned problems, AN-FGM brings a significant speed up when $n > m$ since it allows to (i) reduce the number of variables, and (ii) reduce the computational cost per iteration.

- For ill-conditioned X, AN-FGM outperforms the other approaches, especially for $n = 2m = 100$ (for the same reasons as in the well-conditioned case). For $m = n = 100$ and $n = 50 < m = 100$, the better performance of AN-FGM is explained by the recursive initialization; see Section 4.2.4. FGM and ParTan perform similarly, with a slight advantage for FGM.

- For rank-deficient X, it is easy to analyze: AN-FGM outperforms the other approaches because it reduces the problem size (from an n-by-n variable problem to an r-by-r where $r \ll n$) and the reduced problem is well-conditioned (since the non-zero singular values of X come from a randomly generated matrix). It is interesting and surprising to note that ParTan performs better than FGM in this case. However, we prefer not to use ParTan because it is a heuristic to combine several iterates and comes with no guarantee (authors only prove that there is at least one subsequence of the iterates converging to the optimal solution [5, Lemma 4.1]).

Summary. AN-FGM outperforms all other first-order methods, except when $\text{rank}(X) = n$ and X is well-conditioned. Moreover, when $r = \text{rank}(X) \ll n$, its per-iteration cost is much lower, namely $O(r^3)$ vs. $O(n^3)$.

Remark 5.1 For all generated matrices in the rank-deficient cases and the cases $n = 2m$, we have $\text{rank}(X) < n$ implying that the infimum of $(\mathcal{P})$ is not necessarily attained; see Lemma 2.2. In fact, we have observed that the infimum is never attained. Although we do not have a rigorous explanation

for this fact, we believe that for randomly generated matrices X and B with $\text{rank}(X) < n$, it is very unlikely that the condition $\ker(\hat{A}_{11}) \subseteq \ker(U_2^T B V_1 \Sigma_1^{-1})$ for the infimum to be attained is met, here $\hat{A}_{11}$ is the solution of the PSD Procrustes subproblem (3.2), and U_2, V_1 and Σ_1^{-1} are factors in the SVD of X; see Theorem 3.1. We believe these conditions are not likely to be met when n is large because

- The solution $\hat{A}_{11}$ of the subproblem (3.2) is in general not positive definite. Take for example the simple case $\Sigma_1 = I$ for which $\hat{A}_{11}$ is the projection of $(C + C^T)/2$ on the cone of PSD matrices where $C = U_1^T B V_1$. Since U_1 and V_1 come from the SVD of X which is randomly generated, and B is randomly generated, the entries of C also follows a Gaussian-like distribution (for which the probability for an eigenvalue to be positive is $1/2$). Therefore, it is not likely for $(C+C^T)/2$ to be positive definite hence its projection is in most cases rank deficient; see (1.1).

- The kernel condition is not likely to be satisfied: the probability for one subspace to contain another subspace generated randomly is zero (of course, in our case the subspaces are not independent so a rigorous probabilistic analysis is non trivial).

5.1 Comparison with a second-order method

In this section, we compare AN-FGM with the interior point method SDPT3 (version 4.0) [18, 20], where we used CVX as a modeling system [7, 9]. This is a second-order method hence it is computationally more expensive but guarantees quadratic convergence. We perform in this section exactly the same experiment as in the previous section except that we use matrices of smaller size (for the sizes of the previous section, SDPT3 needs more than one minute to terminate). In order to have a fair comparison, we first run SDPT3 on $(\mathcal{P})$ and then run AN-FGM allowing the same computational time as for SDPT3. Table 5.2 gives the relative error in percent for the different types of matrices for SDPT3 (fourth column) and AN-FGM (third column) within the the same computational time (fifth column). The last column indicates the time for AN-FGM to obtain a solution with relative error up to 0.01% of the final solution generated by SDPT3 (/ indicates that AN-FGM was not able to achieve that accuracy within the allotted time).

We observe the following:

- In all the well-conditioned cases and rank-deficient cases, AN-FGM outperforms SDPT3, being order of magnitude faster (comparing the last two columns of Table 5.2). This is not surprising since AN-FGM has a much lower per-iteration cost while the convergence will be fast because the problems solved by FGM are well conditioned.

- For ill-conditioned cases, SDPT3 allows to obtain high accuracy solutions faster (except in the case $n = 2m = 60$). However, AN-FGM generates solution at most 2% from SDPT3 in the worst case.

  For $n = 2m = 60$, AN-FGM performs better because the subproblem solved by FGM has size only 30-by-30 hence can perform more iterations. Moreover, the infimum is not attained which explains why SDPT3 failed to return a solution with acceptable accuracy (`cvx_status = Failed` in the 10 cases); see also Remark 5.1.

Table 5.2: Comparison between AN-FGM and SDPT3.

		AN-FGM	SDPT3	Time (s.)	AN-FGM - 0.01% (s.)
Well-conditioned	$m = n = 60$	83.77	83.77	10.75	0.36
	$m = 2n = 60$	93.06	93.06	2.26	0.01
	$n = 2m = 60$	60.85	60.86	21.38	0.02
Ill-conditioned	$m = n = 60$	78.09	**76.19**	16.91	/
	$m = 2n = 60$	89.44	**88.70**	2.97	/
	$n = 2m = 60$	**54.08**	54.36	25.93	10.78
Rank-deficient	$m = n = 60$	82.96	82.97	20.21	0.01
	$m = 2n = 60$	91.85	91.85	2.96	0.00
	$n = 2m = 60$	77.10	77.10	19.60	0.00

Summary. Except for ill-conditioned problems of relative small size (m and n up to a hundred on our machine) where n is not significantly smaller than m, AN-FGM should be preferred to SDPT3. For large problems, SDPT3 quickly becomes impractical (for example, SDPT3 requires about 80 seconds for the well-conditioned case with $m = n = 100$, while Matlab crashed when we tried $m = n = 200$).

6 Conclusion

In this paper we have first completely described the set of optimal solution of the PSD Procrustes problem ($\mathcal{P}$) when the infimum is attained, and a sequence of solutions whose objective function value converges to the infimum when the infimum is not attained. This description relies on the solution of a smaller PSD Procrustes problem (where X and B are r-by-r matrices with $r = \text{rank}(X)$ and X is diagonal with positive diagonal elements) whose infimum is always attained. Then, we have applied an optimal first-order method (namely, the fast gradient method) on the subproblem that is guaranteed to converge linearly with rate $(1 - \kappa^{-1})$ where $\kappa = \frac{\sigma_1(X)}{\sigma_r(X)}$. Moreover, to mitigate the slow convergence in ill-conditioned cases, we proposed a new very effective recursive initialization scheme based on a hierarchical decomposition of the problem into well-conditioned subproblems. Finally, our new method, referred to as AN-FGM, was shown to outperform other first-order methods, and compete favorable with a second-order method (namely, SDPT3, a state-of-the-art interior-point method).

References

[1] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM J. on Optimization, 22 (2012), pp. 135–158.

[2] B. Adhikari and R. Alam, *Structured Procrustes problem*, Linear Algebra and its Applications, 490 (2016), pp. 145–161.

[3] A. Albert, *Conditions for positive and nonnegative definiteness in terms of pseudoinverses*, SIAM J. on Applied Mathematics, 17 (1969), pp. 434–440.

[4] J.C. Allwright, *Positive semidefinite matrices: Characterization via conical hulls and least-squares solution of a matrix equation*, SIAM Journal on Control and Optimization, 26 (1988), pp. 537–556.

[5] L.-E. Andersson and T. Elfving, *A constrained Procrustes problem*, SIAM J. on Matrix Analysis and Applications, 18 (1997), pp. 124–139.

[6] J. E. Brock, *Optimal matrices describing linear systems*, AIAA Journal, 6 (1968), pp. 1292–1296.

[7] Inc. CVX Research, *CVX: Matlab software for disciplined convex programming, version 2.0.* `http://cvxr.com/cvx`, Aug. 2012.

[8] J.C. Gower and G.B. Dijksterhuis, *Procrustes Problems*, Oxford University Press, 2004.

[9] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, V. Blondel, S. Boyd, and H. Kimura, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110.

[10] B.F. Green, *The orthogonal approximation of an oblique structure in factor analysis*, Psychometrika, 17 (1952), pp. 429–440.

[11] N.J. Higham, *The symmetric Procrustes problem*, BIT Numerical Mathematics, 28 (1988), pp. 133–143.

[12] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra and its Applications, 103 (1988), pp. 103–118.

[13] H.J. Larson, *Least squares estimation of the components of a symmetric matrix*, Technometrics, 8 (1966), pp. 360–362.

[14] Yu. Nesterov, *A method of solving a convex programming problem with convergence rate O(1/k^2)*, in Soviet Mathematics Doklady, vol. 27, 1983, pp. 372–376.

[15] ——, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2004.

[16] P.H. Schönemann, *A generalized solution of the orthogonal Procrustes problem*, Psychometrika, 31 (1966), pp. 1–10.

[17] T.J. Suffridge and T.L. Hayden, *Approximation by a Hermitian positive semidefinite Toeplitz matrix*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 721–734.

[18] K.-C. Toh, M.J. Todd, and R.H. Tütüncü, *SDPT3–a MATLAB software package for semidefinite programming, version 1.3*, Optimization methods and software, 11 (1999), pp. 545–581.

[19] Lloyd N. Trefethen and David Bau III, *Numerical Linear Algebra*, SIAM, 1997.

[20] R.H. Tütüncü, K.C. Toh, and M.J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Mathematical programming, 95 (2003), pp. 189–217.

[21] K.G. Woodgate, *Optimization over positive semidefinite matrices with application to quasi-Newton algorithms*, PhD thesis, University of London, London, U.K., 1987.

[22] K.G. Woodgate, *Least-squares solution of F=PG over positive semidefinite symmetric P*, Linear algebra and its applications, 245 (1996), pp. 171–190.