Linear Algebra and its Applications 438 (2013) 1863–1882

Contents lists available at SciVerse ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Iterative methods for low rank approximation of graph similarity matrices ☆

T.P. Cason, P.-A. Absil *, P. Van Dooren

Department of Mathematical Engineering, ICTEAM Institute, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 1 December 2010
Accepted 23 November 2011
Available online 9 January 2012

Submitted by V. Mehrmann

Keywords:
Graph theory
Node to node similarity
Trace maximization
Low-rank approximation
Algorithm
Set of low-rank matrices

ABSTRACT

In this paper, we analyze an algorithm to compute a low-rank approximation of the similarity matrix **S** introduced by Blondel et al. in [1]. This problem can be reformulated as an optimization problem of a continuous function $\Phi(S) = \mathrm{tr}\left(S^T \mathcal{M}^2(S)\right)$ where S is constrained to have unit Frobenius norm, and $\mathcal{M}^2$ is a non-negative linear map. We restrict the feasible set to the set of matrices of unit Frobenius norm with either k nonzero identical singular values or at most k nonzero (not necessarily identical) singular values. We first characterize the stationary points of the associated optimization problems and further consider iterative algorithms to find one of them. We analyze the convergence properties of our algorithm and prove that accumulation points are stationary points of $\Phi(S)$. We finally compare our method in terms of speed and accuracy to the full rank algorithm proposed in [1].

1. Introduction

Node-to-node similarity measures compare the nodes of a graph G_A with the nodes of an other graph G_B according to some similarity criterion, and have been applied to many practical problems such as comparing chemical structures [2], navigating in complex networks like the World Wide Web [3], and analyzing different kinds of biological data [4].

☆ This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

* Corresponding author.
E-mail address: absil@inma.ucl.ac.be (P.-A. Absil).

0024-3795/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.laa.2011.12.004

These node-to-node similarity measures are conveniently stored in the so-called similarity matrix, **S**, whose (i, j) entry tells how the node i is similar to the node j. In [1], Blondel et al. define a node-to-node similarity measure as a fixed point of an iterative process, and prove that their measure is equivalent to the solution of an eigenvalue problem of a dimension that is the product of the number of nodes in both graphs. For large graphs, computing this similarity measure can hence be quite expensive. In [5], Fraikin et al. approach the similarity matrix defined by Blondel et al. by a rank-k matrix with k identical singular values (note that this approximation is exact with $k = 1$ when one of the two graphs to compare is undirected) and propose to reduce the computational cost of the Blondel et al. similarity by using a low-rank iterative scheme that experimentally converges towards their approximation.

In this paper, we propose two low-rank iterative schemes that converge towards two approximations of the Blondel et al. similarity matrix with respectively either k nonzero identical singular values or at most k nonzero (not necessarily identical) singular values, and further analyze the convergence properties of our algorithms.

This paper is organized as follows. Section 2 introduces the notations further used in the article. Section 3 recalls the similarity matrix defined by Blondel et al. Section 4 shows that the similarity matrix defined by Blondel et al. is the solution of an optimization problem. Sections 5 and 6 analyze different low-rank approximations of this optimization problem. Section 7 analyzes the complexity of our algorithms and Section 8 presents experimental results. And finally, Section 9 gives our conclusions.

2. Notations

Throughout, G_A and G_B stand for graphs with respectively m and n nodes. These graphs are conveniently represented by $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$, their respective adjacency matrices, *i.e.* $A_{ij} = 1$ (resp. $B_{ij} = 1$) if there is an edge from node i to node j in G_A (resp. G_B), otherwise $A_{ij} = 0$ (resp. $B_{ij} = 0$). And $C_A(i)$ and $P_A(i)$ denote respectively the set of children and parents of node i in G_A.

In this paper, we use the following matrix functions:

- The Frobenius inner product defined as

$$\langle \cdot, \cdot \rangle_F : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R} : S_1, S_2 \mapsto \langle S_1, S_2 \rangle_F := \operatorname{tr}\left(S_1^T S_2\right),$$

  along with the associated Frobenius norm, $\|S\|_F := \sqrt{\langle S, S \rangle_F}$.
- The vectorization of a matrix is defined as

$$\operatorname{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn} : S = \begin{bmatrix} S(\cdot, 1) & \cdots & S(\cdot, n) \end{bmatrix} \mapsto \operatorname{vec}(S) = \begin{bmatrix} S(\cdot, 1) \\ \vdots \\ S(\cdot, n) \end{bmatrix}.$$

We further consider the following sets of matrices of Frobenius norm 1:

$$\mathcal{S}(m, n) := \operatorname{Norm}(1, m, n) = \left\{ S \in \mathbb{R}^{m\times n} : \|S\|_F = 1 \right\},$$

$$\mathcal{S}_k(m, n) := \left\{ U\hat{I}_k V^T \in \mathbb{R}^{m\times n} : \begin{array}{c} U \in \operatorname{St}(k, m),\ V \in \operatorname{St}(k, n), \\ \hat{I}_k = I_k / \|I_k\|_F = I_k/\sqrt{k} \end{array} \right\}, \quad \text{and}$$

$$\mathcal{S}_{\leqslant k}(m, n) := \left\{ UDV^T \in \mathbb{R}^{m\times n} : \begin{array}{c} U \in \operatorname{St}(k, m),\ V \in \operatorname{St}(k, n), \\ D \text{ diagonal}, \|D\|_F = 1 \end{array} \right\},$$

where $\operatorname{St}(k, m)$ denotes the Stiefel manifold, *i.e.*

$$\operatorname{St}(k, m) := \left\{ U \in \mathbb{R}^{m\times k} : U^T U = I_k \right\}.$$

That is, $\mathcal{S}_k(m, n)$ is the set of all $m \times n$ matrices with unit Frobenius norm and k nonzero equal singular values, and $\mathcal{S}_{\leqslant k}(m, n)$ is the set of all $m \times n$ matrices with unit Frobenius norm and rank less than or equal to k.

$\mathrm{O}(m)$ denotes the set of orthogonal matrices of order m, *i.e.*

$$\mathrm{O}(m) := \left\{ Q \in \mathbb{R}^{m\times m} : \ Q^T Q = I_m = QQ^T \right\} .$$

$\mathrm{Diag}(k, m, n)$ denotes a set of diagonal matrices defined as follows

$$\mathrm{Diag}(k, m, n) := \left\{ D \in \mathbb{R}^{m\times n} : \ D \text{ diagonal, and } D_{ii} = 0 \text{ for all } i > k \right\} .$$

$\mathcal{S}_{\mathrm{Skew}}(k)$ denotes the set of skew-symmetric matrices of order k. $\mathcal{S}_{\mathrm{Sym}}(k)$ denotes the set of symmetric matrices of order k. $\mathbf{1}$ is the matrix whose entries are all equal to 1.

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{S}$, a non empty algebraic subset of $\mathcal{H}$. A vector $\xi \in \mathcal{H}$ is an *analytic admissible direction* for $\mathcal{S}$ at $S \in \mathcal{S}$ if there exists an analytic curve $\gamma(t) : \mathbb{R} \mapsto \mathcal{H}$ with $\gamma(0) = S$, and $\gamma(t) \in \mathcal{S}$, for all $t \geqslant 0$, such that

$$\lim_{t\to 0} \frac{\gamma(t) - \gamma(0)}{t} = \xi.$$

According to [6, Proposition 2], the *contingent cone* (see, e.g., [7]) to $\mathcal{S}$ at S, denoted $C_S\mathcal{S}$, is equal to the set of all analytic admissible directions for $\mathcal{S}$ at S, *i.e.*

$$C_S\mathcal{S} = \left\{ \begin{array}{c} \dot\gamma(0) : \ \gamma \text{ is an analytic curve with } \gamma(0) = S, \\ \text{and } \gamma(t) \in \mathcal{S}, \text{ for all } t \geqslant 0. \end{array} \right\} .$$

The normal cone to $\mathcal{S}$ at S, denoted $N_S\mathcal{S}$, is defined as

$$N_S\mathcal{S} := \left\{ \zeta \in \mathcal{H} : \langle \zeta, \xi \rangle_F \leqslant 0, \forall \xi \in C_S\mathcal{S} \right\} .$$

3. The similarity matrix

Node-to-node similarity measures compare the nodes of a graph G_A with the nodes of an other graph G_B according to some similarity criterion. In [1], Blondel et al. introduce a recursive requirement which states that the similarity between node i and node j should be large if the similarity between the neighbors of node i and the neighbors of node j is large. More specifically, they define a similarity measure by means of the following algorithm.

Algorithm A0

Given: graphs G_A and G_B respectively of order m and n.
$S^0 \leftarrow \mathbf{1}/\left\|\mathbf{1}\right\|_F \in \mathbb{R}^{m\times n}$
for $t = 1, 2, \ldots, t_{max}$ **do**

$$S^t \leftarrow \frac{\mathcal{M}(S^{t-1})}{\left\|\mathcal{M}(S^{t-1})\right\|_F}, \ \text{with } [\mathcal{M}(S)]_{ij} := \sum_{\substack{k\in C_A(i)\\ l\in C_B(j)}} S_{kl} + \sum_{\substack{k\in P_A(i)\\ l\in P_B(j)}} S_{kl}$$

endfor
$\mathbf{S} \leftarrow S^t$

where t_{max} is an even number that is "sufficiently large".

In this algorithm, they first initialize all similarity scores to the same value, and further update them in the *reinforcement loop*, that can be justified as follows:

- $[\mathcal{M}(S^{t-1})]_{ij} : S^t_{ij}$, the similarity score between node i and node j at step t, is the sum of all (k, l) entries of S^{t-1} such that node k is a child of node i in G_A and node l is a child of node j in G_B, plus the

sum of all (k, l) entries of S^{t-1} such that node k is a parent of node i in G_A and node l is a parent of node j in G_B. Doing so, the similarity score between two nodes increases if they have many highly similar children or parents.

- $\mathcal{M}(S^{t-1})/\left\|\mathcal{M}(S^{t-1})\right\|_F$: since they are not interested in the absolute value of S^t_{ij} but only in the relative score of two different pairs, they normalize the whole similarity matrix S^t to avoid over- or under-flow.

One can rewrite $\mathcal{M}(S)$ in terms of matrix operations over S, *i.e.*

$$[\mathcal{M}(S)]_{ij} = \sum_{k,l} A_{ik} S_{kl} B_{jl} + \sum_{k,l} A_{ki} S_{kl} B_{lj} = \left[ASB^T + A^T SB\right]_{ij} . \tag{1}$$

Since $\mathrm{vec}\left(ASB^T\right) = (B \otimes A)\,\mathrm{vec}(S)$, Eq. (1) can be rewritten under its so-called vector form:

$$\mathrm{vec}(\mathcal{M}(S)) = M\,\mathrm{vec}(S) := \left(B \otimes A + B^T \otimes A^T\right)\,\mathrm{vec}(S).$$

In [1], Blondel et al. show that Algorithm A0 is in fact the power method applied to the matrix M. This matrix is non-negative and hence, according to Perron–Frobenius Theorem, there exists a real positive eigenvalue ρ, called the Perron root, such that any other eigenvalue λ satisfies $|\lambda| \leqslant \rho$. Since M is symmetric, its eigenvalues are real and hence M can have at most two extremal eigenvalues (*i.e.* of maximum modulus), ρ and possibly $-\rho$. As a direct consequence, M^2 has only one extremal eigenvalue, namely ρ^2 (but possibly of multiplicity higher than one), and the even iterates of the reinforcement loop in Algorithm A0 converge towards $S^{2\infty}$, the normalized orthogonal projection of S^0 onto

$$E_{\rho^2}(\mathcal{M}^2) := \left\{S \;\text{ s.t. }\; \rho^2 S = \mathcal{M}^2(S)\right\} ,$$

the eigenspace of $\mathcal{M}^2$ associated to ρ^2, with respect to the Frobenius inner product (see [1] for details about the proof of convergence). Notice that, since $S^{2\infty}$ is a fixed point of the even iterates of the reinforcement loop in Algorithm A0, one can write

$$\rho^2 S^{2\infty} = \mathcal{M}^2(S^{2\infty}) .$$

4. From similarity to optimization

In this section, we show that the similarity matrix defined by Blondel et al. is the solution of an optimization problem.

One can first observe that the iteration in Algorithm A0 is such that

$$S^t \in \underset{\|S\|_F=1}{\mathrm{argmax}} \left\langle S, \mathcal{M}(S^{t-1})\right\rangle_F .$$

This result is easy to prove using to the Cauchy–Schwarz inequality

$$\begin{aligned}\left\langle S^t, \mathcal{M}(S^{t-1})\right\rangle_F &= \left\langle \frac{\mathcal{M}(S^{t-1})}{\left\|\mathcal{M}(S^{t-1})\right\|_F}, \mathcal{M}(S^{t-1})\right\rangle_F = \left\|\mathcal{M}(S^{t-1})\right\|_F \\ &= \|S\|_F \left\|\mathcal{M}(S^{t-1})\right\|_F \overset{CS}{\geqslant} \left\langle S, \mathcal{M}(S^{t-1})\right\rangle_F , \qquad \text{since } \|S\|_F = 1.\end{aligned}$$

Moreover, one can prove that $S^{2\infty}$ is a solution of

$$\underset{S\in\mathcal{S}(m,n)}{\mathrm{argmax}}\ \Phi(S) , \quad \text{where} \quad \Phi(S) := \left\langle S, \mathcal{M}^2(S)\right\rangle_F = \mathrm{tr}\left(S^T\,\mathcal{M}^2(S)\right) , \tag{2}$$

and $\mathcal{M}^2(S) = \mathcal{M}(\mathcal{M}(S))$ is defined in Eq. (1). Indeed, since the Perron root is equal to the spectral radius, *i.e.*

$$\rho = \max_{\|S\|_F=1} \|\mathcal{M}(S)\|_F , \tag{3}$$

and the map $\mathcal{M}(\cdot)$ is self-adjoint (*i.e.* $\langle S_1, \mathcal{M}(S_2)\rangle_F = \langle \mathcal{M}(S_1), S_2\rangle_F$), we have

$$\max_{\|S\|_F=1} \left\langle S, \mathcal{M}^2(S)\right\rangle_F = \max_{\|S\|_F=1} \|\mathcal{M}(S)\|_F^2 = \rho^2 = \left\langle S^{2\infty}, \mathcal{M}^2\left(S^{2\infty}\right)\right\rangle_F . \tag{4}$$

The problem (2) maximizes a continuous function Φ on a compact domain. Hence, according to first order optimality condition, if $S^{2\infty}$ is a maximizer of (2) then $S^{2\infty}$ is a stationary point of (2). The concept of stationary point in the context of (2) is recalled in Definition 1 below.

The *gradient* of a differentiable function Φ at a point S, denoted grad $\Phi(S)$, is defined as the unique vector that satisfies

$$\langle \xi, \text{grad } \Phi(S)\rangle_F = \mathrm{D}\Phi(S)[\xi] , \tag{5}$$

for all admissible directions ξ.

Definition 1. A point $S \in \mathcal{S}(m, n)$ is a *stationary point* of (2) if grad $\Phi(S)$ belongs to $N_S\mathcal{S}(m, n)$, the normal cone to $\mathcal{S}(m, n)$ at S, *i.e.*

$$\text{grad } \Phi(S) \in N_S\mathcal{S}(m, n) := \left\{\zeta : \langle \zeta, \xi\rangle_F \leqslant 0, \forall \xi \in C_S\mathcal{S}(m, n)\right\} . \tag{6}$$

For example, the tangent cone to $\mathcal{S}(m, n) = \{S : \|S\|_F = 1\}$ at a point S is given by

$$C_S\mathcal{S}(m, n) = \left\{\xi : \langle \xi, S\rangle_F = 0\right\} ,$$

and the normal cone to $\mathcal{S}(m, n)$ is

$$N_S\mathcal{S}(m, n) := \left\{\zeta : \langle \zeta, \xi\rangle_F \leqslant 0, \forall \xi \in C_S\mathcal{S}(m, n)\right\} = \{\alpha S : \alpha \in \mathbb{R}\} .$$

Since the linear map $\mathcal{M}$ is self-adjoint (*i.e.* $\langle S_1, \mathcal{M}(S_2)\rangle_F = \langle \mathcal{M}(S_1), S_2\rangle_F$) one can write

$$\mathrm{D}\Phi(S)[\xi] = \left\langle \xi, \mathcal{M}^2(S)\right\rangle_F + \left\langle S, \mathcal{M}^2(\xi)\right\rangle_F = \left\langle \xi, 2\mathcal{M}^2(S)\right\rangle_F , \tag{7}$$

and the gradient of Φ at a point S is then $2\mathcal{M}^2(S)$. And clearly, one can observe that

$$\text{grad } \Phi(S^{2\infty}) = 2\mathcal{M}^2(S^{2\infty}) = 2\rho^2 S^{2\infty} \in N_{S^{2\infty}}\mathcal{S}(m, n) = \left\{\alpha S^{2\infty} : \alpha \in \mathbb{R}\right\} , \tag{8}$$

and it follows directly that $S^{2\infty}$ is a stationary point of (2).

When S is large, Algorithm A0 becomes relatively expensive in terms of computational cost. Hence one can think of modifying the problem in order to find an approximation of S at lower cost. This paper considers two kinds of low-rank approximations of the similarity matrix S, either by matrices of norm 1 with k nonzero identical singular values or by matrices of norm 1 with at most k nonzero (not necessarily identical) singular values. We first characterize the stationary points of the associated optimization problems and further consider iterative algorithms to find one of them.

5. Approximation with identical singular values

We first consider the following approximations for the feasible set of (2).

Problem 1. *Solve* (2) *with* $\mathcal{S}(m, n)$ *replaced by* $\mathcal{S}_k(m, n)$*, the set of rank-k matrices of norm* 1 *with k identical singular values, i.e.*

$$\mathcal{S}_k(m, n) := \left\{ U\hat{I}_k V^T \in \mathbb{R}^{m\times n} : \begin{array}{c} U \in \mathrm{St}(k, m),\ V \in \mathrm{St}(k, n), \\ \hat{I}_k = I_k/\|I_k\|_F = I_k/\sqrt{k} \end{array} \right\} . \tag{9}$$

Note that if the extremal eigenvalues (*i.e.* eigenvalues of maximum modulus) of $\mathcal{M}$ are positive, then Problem 1 is equivalent to the problem considered in [5].

5.1. The feasible set of Problem 1 and its stationary points

Problem 1 is defined on a feasible set that has a *manifold* structure [8]. In this case, the tangent cone is called the tangent space since it admits a structure of vector space (see [9, Section 3.5] for details). The tangent space to the Stiefel manifold $\mathrm{St}(k, m)$ at a point U is given by

$$C_U\mathrm{St}(k, m) := \left\{\xi : \xi^T U + U^T\xi = 0_k\right\} ;$$

see, e.g., [9, Section 3.5.7]. Equivalently,

$$C_U\mathrm{St}(k, m) = \left\{U\Omega + U_\perp K : \Omega \in \mathcal{S}_{\mathrm{Skew}}(k),\ K \in \mathbb{R}^{m-k\times k}\right\} . \tag{10}$$

where $U_\perp$ denotes any matrix such that $\begin{bmatrix}U & U_\perp\end{bmatrix}^T \begin{bmatrix}U & U_\perp\end{bmatrix} = I_m$. For Problem 1, it follows that the tangent space to $\mathcal{S}_k(m, n)$ at a point $S = U\hat{I}_k V^T \in \mathcal{S}_k(m, n)$ is given by

$$\begin{aligned} C_S\mathcal{S}_k(m, n) &:= \{\dot\gamma(0) : \gamma \text{ curve on } \mathcal{S}_k(m, n) \text{ with } \gamma(0) = S\} \\ &= \left\{\begin{array}{l} U\Omega V^T + UK_V^T V_\perp^T + U_\perp K_U V^T \ s.t. \\ \quad \Omega \in \mathcal{S}_{\mathrm{Skew}}(k),\ K_U \in \mathbb{R}^{m-k\times k},\ K_V \in \mathbb{R}^{n-k\times k} \end{array}\right\} . \end{aligned}$$

And subsequently, the normal space to $\mathcal{S}_k(m, n)$ at a point $S = U\hat{I}_k V^T \in \mathcal{S}_k(m, n)$ is given by

$$\begin{aligned} N_S\mathcal{S}_k(m, n) &:= \{\zeta : \langle\zeta, \xi\rangle_F \leqslant 0, \forall\xi \in C_S\mathcal{S}_k(m, n)\} \\ &= \left\{UHV^T + U_\perp KV_\perp^T \ s.t.\ H \in \mathcal{S}_{\mathrm{Sym}}(k),\ K \in \mathbb{R}^{m-k\times n-k}\right\} . \end{aligned}$$

As mentioned in Definition 1, a matrix S is defined to be a stationary point of (2) if $\mathrm{grad}\,\Phi(S)$ $(= 2\mathcal{M}^2(S))$ belongs to the normal cone to the feasible set at S. Hence, $S = U\hat{I}_k V^T \in \mathcal{S}_k(m, n)$ is a stationary point of Problem 1 if and only if

$$\mathcal{M}^2(S) \in N_S\mathcal{S}_k(m, n) = \left\{\begin{array}{l} UHV^T + U_\perp KV_\perp^T \ s.t. \\ \quad H \in \mathcal{S}_{\mathrm{Sym}}(k),\ K \in \mathbb{R}^{m-k\times n-k} \end{array}\right\} . \tag{11}$$

5.2. Algorithm for Problem 1 and its convergence analysis

We now propose an algorithm and further prove that it converges towards stationary points of Problem 1 (see Theorem 5.5).

Algorithm A1

1: $S^0 \leftarrow \mathbf{1}/\,\|\mathbf{1}\|_F$
2: **for** $t = 1, 2, \ldots$
3: Compute $S^t \in \mathcal{S}_k(m, n)$ according to

$$S^t\ (= U^t\hat{I}_k[V^t]^T)\ \leftarrow f(S^{t-1}) := \operatorname*{argmax}_{\tilde S\in\mathcal{S}_k(m,n)} \left\langle\tilde S, \mathcal{M}^2(S^{t-1})\right\rangle_F \tag{12}$$

4: **end**

where $\mathcal{M}^2(S) = \mathcal{M}(\mathcal{M}(S))$ is defined in Eq. (1).

Let $\mathcal{M}^2(S^{t-1})$ have an ordered singular value decomposition (SVD)

$$\mathcal{M}^2(S^{t-1}) = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}, \tag{13}$$

with $P_1 \in \mathbb{R}^{m\times k}$, $P_2 \in \mathbb{R}^{m\times(m-k)}$, $Q_1 \in \mathbb{R}^{n\times k}$, $Q_2 \in \mathbb{R}^{n\times(n-k)}$, $\Sigma_1 \in \mathbb{R}^{k\times k}$ and $\Sigma_2 \in \mathbb{R}^{(m-k)\times(n-k)}$. Let $\nu(S^{t-1})$ denote the singular value gap,

$$\nu(S^{t-1}) := \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2) = \sigma_k(\mathcal{M}^2(S^{t-1})) - \sigma_{k+1}(\mathcal{M}^2(S^{t-1})). \tag{14}$$

When $\nu(S^{t-1}) \neq 0$ (i.e., $\nu(S^{t-1}) > 0$), the next iterate S^t is uniquely defined by (12) and is equal to $P_1\hat{I}_kQ_1^T$, as we will show in Lemma 5.1 below. When $\nu(S^{t-1}) = 0$, however, S^t is no longer uniquely defined by (12); in this case, S^t is chosen arbitrarily in $\underset{\tilde{S}\in\mathcal{S}_k(m,n)}{\operatorname{argmax}} \left\langle \tilde{S}, \mathcal{M}^2(S^{t-1}) \right\rangle_F$. In practice, in our numerical experiments, we systematically choose $S^t := P_1\hat{I}_kQ_1^T$, where P_1 and Q_1 are returned by the SVD function.

We first state a few intermediate results in order to prove convergence of Algorithm A1 to the stationary points of Problem 1.

Lemma 5.1. *Let $M \in \mathbb{R}^{m\times n}$ and its ordered singular value decomposition*

$$M = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = P\Sigma Q^T \tag{15}$$

with $P_1 \in \mathbb{R}^{m\times k}$, $P_2 \in \mathbb{R}^{m\times(m-k)}$, $Q_1 \in \mathbb{R}^{n\times k}$, $Q_2 \in \mathbb{R}^{n\times(n-k)}$, $\Sigma_1 \in \mathbb{R}^{k\times k}$ and $\Sigma_2 \in \mathbb{R}^{(m-k)\times(n-k)}$. Then

$$\max_{S=U\hat{I}_kV^T\in\mathcal{S}_k(m,n)} \langle S, M\rangle_F = \operatorname{tr}\left(\hat{I}_k\Sigma_1\right). \tag{16}$$

Moreover, if $\nu := \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2) > 0$, then the maximizing solution S is unique and equals $P_1\hat{I}_kQ_1^T$.

Proof. We have

$$\begin{aligned} \operatorname{tr}\left(\hat{I}_kU^TMV\right) &\leqslant \sum_{i=1}^{k} \sigma_i(\hat{I}_kU^TMV) \\ &\leqslant \sum_{i=1}^{k} \sigma_i(\hat{I}_k)\,\sigma_i(\Sigma_1) \leqslant \operatorname{tr}\left(\hat{I}_k\Sigma_1\right) \end{aligned} \tag{17}$$

according to [10, Formulas 3.1.10b, and Lemma 3.3.1]. The upper bound is reached for $U = P_1$, and $V = Q_1$. The uniqueness of $P_1\hat{I}_kQ_1^T$ is a well known result discussed, *e.g.*, in [10, Theorem 3.1.1 and 3.1.1']. □

Theorem 5.2. *Let $\mathcal{M}^2(S)$ have an ordered singular value decomposition*

$$\mathcal{M}^2(S) = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = P\Sigma Q^T, \tag{18}$$

with $P_1 \in \mathbb{R}^{m\times k}$, $P_2 \in \mathbb{R}^{m\times(m-k)}$, $Q_1 \in \mathbb{R}^{n\times k}$, $Q_2 \in \mathbb{R}^{n\times(n-k)}$, $\Sigma_1 \in \mathbb{R}^{k\times k}$ and $\Sigma_2 \in \mathbb{R}^{(m-k)\times(n-k)}$, and $S^+ := f(S)$, with f the function defined in Algorithm A1. Then

$$\Phi(S^+) - \Phi(S) \geqslant \left\|\mathcal{M}(S^+ - S)\right\|_F^2 + \nu\sqrt{k}\,\left\|S^+ - S\right\|_F^2 \tag{19}$$

with Φ the function defined in Eq. (2), and $\nu = \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2)$. In particular, Algorithm A1 is an ascent iteration for Φ.

Proof. Adding and subtracting S to S^+ and using the self-adjointness of $\mathcal{M}$ yield

$$\Phi(S^+ - S + S) = \left\|\mathcal{M}(S^+ - S)\right\|_F^2 + 2\left\langle S^+ - S, \mathcal{M}^2(S)\right\rangle_F + \Phi(S) \tag{20}$$

According to Lemma 5.1, $S^+ = P_1\hat{I}_k Q_1^T$. Hence, using the ordered singular value decomposition of $\mathcal{M}^2(S)$, the second term of the right-hand side of Eq. (20) becomes

$$\left\langle S^+ - S, \mathcal{M}^2(S)\right\rangle_F = \mathrm{tr}\left(\left(\hat{I}_k - Q_1^T S^T P_1\right)\Sigma_1\right) - \mathrm{tr}\left(Q_2^T S^T P_2 \Sigma_2\right) \tag{21}$$

Moreover, one can observe that

$$\left(\hat{I}_k - Q_1^T(U\hat{I}_k V^T)^T P_1\right)_{ii} \geqslant \frac{1}{\sqrt{k}}\left(1 - \left\|V^T(Q_1)_{\cdot i}\right\|_F \left\|U^T(P_1)_{\cdot i}\right\|_F\right) \geqslant 0\,. \tag{22}$$

As a consequence, the first term of the right hand side of (21) is bounded below by

$$\mathrm{tr}\left(\left(\hat{I}_k - Q_1^T S^T P_1\right)\Sigma_1\right) \geqslant \mathrm{tr}\left(\hat{I}_k - Q_1^T S^T P_1\right)\sigma_{min}(\Sigma_1)\,, \tag{23}$$

and the second term of the right hand side of (21) is bounded below by

$$-\mathrm{tr}\left(Q_2^T S^T P_2 \Sigma_2\right) \geqslant -\mathrm{tr}\left(\left|Q_2^T S^T P_2 I_{m-k,n-k}\right|\right)\sigma_{max}(\Sigma_2)\,. \tag{24}$$

Using (23) and (24) with (21) yields

$$\begin{aligned}\left\langle S^+ - S, \mathcal{M}^2(S)\right\rangle_F \geqslant\ & \mathrm{tr}\left(\hat{I}_k - Q_1^T S^T P_1\right)\sigma_{min}(\Sigma_1) \\ & -\mathrm{tr}\left(\left|Q_2^T S^T P_2 I_{m-k,n-k}\right|\right)\sigma_{max}(\Sigma_2)\,.\end{aligned} \tag{25}$$

Adding and subtracting $\sigma_{max}(\Sigma_2)$ to $\sigma_{min}(\Sigma_1)$ along with $Q_1^T S^T P_1 \leqslant |Q_1^T S^T P_1|$ yields

$$\begin{aligned}\left\langle S^+ - S, \mathcal{M}^2(S)\right\rangle_F \geqslant\ & \mathrm{tr}\left(\hat{I}_k - Q_1^T S^T P_1\right)\left(\sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2)\right) \\ & + \left[\mathrm{tr}\left(\hat{I}_k\right) - \mathrm{tr}\left(\left|Q^T S^T P I_{m,n}\right|\right)\right]\sigma_{max}(\Sigma_2)\,.\end{aligned} \tag{26}$$

One can observe that

$$\begin{aligned}\mathrm{tr}\left(\left|Q^T S^T P I_{m,n}\right|\right) &= \mathrm{tr}\left(Q^T V \hat{I}_k U^T P \tilde{I}_{m,n}\right) = \mathrm{tr}\left(\hat{I}_k U^T P \tilde{I}_{m,n} Q^T V\right) \\ &= \sum_{i=1}^{k}\frac{1}{\sqrt{k}}(U_{\cdot i})^T P \tilde{I}_{m,n} Q^T V_{\cdot i} \leqslant \sum_{i=1}^{k}\frac{1}{\sqrt{k}}\left\|P^T U_{\cdot i}\right\|_F \left\|Q^T V_{\cdot i}\right\|_F = \sqrt{k} = \mathrm{tr}\left(\hat{I}_k\right)\end{aligned}$$

where $(\tilde{I}_{m,n})_{ij} = (I_{m,n})_{ij}\ \mathrm{sign}\left((Q_{\cdot j})^T V \hat{I}_k U^T P_{\cdot i}\right)$. And Eq. (26) reduces to

$$\left\langle S^+ - S, \mathcal{M}^2(S)\right\rangle_F \geqslant \mathrm{tr}\left(\hat{I}_k - U^T P_1 \hat{I}_k Q_1^T V\right)\left(\sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2)\right)\,. \tag{27}$$

Finally, one can see that

$$\left\|S^+ - S\right\|_F^2 = 2\,\frac{1}{\sqrt{k}}\,\mathrm{tr}\left(\hat{I}_k - U^T P_1 \hat{I}_k Q_1^T V\right)\,. \tag{28}$$

Combining (28), (27), and (20) gives the desired result. □

Lemma 5.3. *If S is a fixed point of Algorithm A1, then S is a stationary point of Problem 1.*

Proof. Let $S = U\hat{I}_k V^T$ be a fixed point. Then, in view of Lemma 5.1, the ordered singular value decomposition of $\mathcal{M}^2(U\hat{I}_k V^T) = U^+\Sigma_1[V^+]^T + U_\perp\Sigma_2 V_\perp^T$ with $U\hat{I}_k V^T = U^+\hat{I}_k[V^+]^T$. Since $U\hat{I}_k V^T$ and $U^+\hat{I}_k[V^+]^T$ are two singular value decompositions of the same matrix, there must be a square orthogonal matrix Q such that $U^+ = UQ$ and $V^+ = VQ$.

Hence,

$$\mathcal{M}^2(U\hat{I}_k V^T) = U\,Q\Sigma_1 Q^T\,V^T + U_\perp\Sigma_2 V_\perp^T\,, \tag{29}$$

and we conclude that $S = U\hat{I}_k V^T$ is a stationary point of Problem 1 since Eq. (29) satisfies the stationarity condition (11). □

Lemma 5.4. *Let S be a nonstationary point of Problem 1. There exists an $\epsilon > 0$ such that for all $\|S_\epsilon\|_F < \epsilon$, $S + S_\epsilon$ is not a stationary point.*

Proof. The critical points of an analytic (non-constant) function form a closed set with empty interior (actually an analytic set). □

The next theorem states the main convergence result for Algorithm A1. Note that the existence of an accumulation point is guaranteed by the fact that the iteration evolves on the compact set $\mathcal{S}_k(m, n)$. In practice, in our experiments, the sequences of iterates always had a single accumulation point S', with $\nu(S') \neq 0$; by virtue of the next theorem, it thus follows that S' is a stationary point of Problem 1. Moreover, since the iteration is an ascent iteration for Φ, convergence to stationary points that are not local maxima is not expected to occur in practice.

Theorem 5.5. *Let S', with $\nu(S') \neq 0$ (14), be an accumulation point of the sequence $\{S_i\}$ constructed by Algorithm A1. Then S' is a stationary point of Problem 1.*

Proof. This proof relies heavily on [11, Section 1.3 and Theorem 3]. Let S' be such an accumulation point, *i.e.* $S_i \to S'$ for $i \in K$ where $K \in \mathbb{N}$ is an infinite index set. Assume that $S' = U'\hat{I}_k V'^T$ is not a stationary point of the iteration. Let $\bar{B}_\epsilon(S')$ denote the closed ball $\{S \in \mathbb{R}^{m\times n} : \|S - S'\| \leqslant \epsilon\}$. One can choose an $\epsilon > 0$ using Lemma 5.4 such that all $S \in \bar{B}_\epsilon(S')$ are nonstationary points with nonzero gap. In view of Lemma 5.3, these points are not fixed points either, *i.e.*, $\|S^+ - S\|_F^2 > 0$. In view of Theorem 5.2, $\Phi(S^+) - \Phi(S) > 0$ for all $S \in \bar{B}_\epsilon(S')$.

In order to proceed, we need to show that $\Phi(S^+) - \Phi(S)$ is actually bounded away from zero on $\bar{B}_\epsilon(S')$, i.e., (30). To this end, it is sufficient to show that $S \mapsto \Phi(S^+) - \Phi(S)$ is continuous for all $S \in \bar{B}_\epsilon(S')$. The function $S \mapsto \Phi(S)$ is continuous in view of the definition of Φ in (2). To conclude the continuity argument, we show that the function $S \mapsto S^+$ is also continuous at all points where the gap is nonzero. To see this, observe that $S \mapsto S^+$ is the composition of the function $S \mapsto \mathcal{M}^2(S)$ and of the function $M \mapsto P_1\hat{I}_k Q_1^T = \frac{1}{k}P_1 Q_1^T$ defined from Lemma 5.1. The first function is continuous; it remains to show that the second one is. Consider $M_\star \in \mathcal{M}^2(\bar{B}_\epsilon(S'))$. Let L_M be an orthonormal basis of the dominant k-dimensional left singular subspace of M such that $M \mapsto L_M$ is continuous at $M_\star$; this is possible in view of [12, Theorem 6.4]. Let R_M be chosen likewise for the right singular subspace. We then have $P_1 = L_M\tilde{P}_1$ and $Q_1 = R_M\tilde{Q}_1$, where $\tilde{P}_1, \tilde{Q}_2 \in \mathrm{O}(k)$. Thus $M = L_M\tilde{P}_1\Sigma_1\tilde{Q}_1^T R_M^T + P_2\Sigma_2 Q_2^T$ and $L_M^T M R_M = \tilde{P}_1\Sigma_1\tilde{Q}_1^T = \tilde{P}_1\tilde{Q}_1^T\,\tilde{Q}_1\Sigma_1\tilde{Q}_1^T$. Hence, in view of the polar decomposition, $\tilde{P}_1\tilde{Q}_1^T = L_M^T M R_M((L_M^T M R_M)^T L_M^T M R_M)^{-1/2}$. Finally, $P_1 Q_1^T = L_M\tilde{P}_1\tilde{Q}_1^T R_M^T = L_M L_M^T M R_M((L_M^T M R_M)^T L_M^T M R_M)^{-1/2} R_M^T$. This function is continuous at $M = M_\star$. (Note that the argument of the square root is positive-definite locally in view of the nonzero gap assumption.) Since $M_\star$ is arbitrary, the claim follows. We have thus shown that

$$\delta := \min_{\|S-S'\|_F \leqslant \epsilon} \Phi(S^+) - \Phi(S) > 0\,. \tag{30}$$

Since $S_i \to S'$ for $i \in K$, there exists a $k \in K$ such that for all $i \geqslant k$, $i \in K$,

$$\left\|S_i - S'\right\|_F^2 \leqslant \epsilon \quad \text{and thus} \quad \Phi(S_{i+1}) - \Phi(S_i) \geqslant \delta\,. \tag{31}$$

Hence, for any two consecutive S_i, S_{i+j} points of the subsequence, with $i, i+j \geqslant k$, $i, i+j \in K$, we must have

$$\Phi(S_{i+j}) - \Phi(S_i) \geqslant \Phi(S_{i+1}) - \Phi(S_i) \geqslant \delta > 0\,. \tag{32}$$

But, since Φ is continuous, the sequence $\{\Phi(S_i)\}_{i\in K}$ must converge. This is contradicted by (32), which implies S' has to be a stationary point of Problem 1. □

6. Approximation of rank at most *k*

We now consider the following approximations for the feasible set of (2).

Problem 2. *Solve* (2) *with* $\mathcal{S}(m,n)$ *replaced by* $\mathcal{S}_{\leqslant k}(m,n)$*, the set of matrices of norm* 1 *with rank at most k, i.e.*

$$\mathcal{S}_{\leqslant k}(m,n) := \left\{ \begin{array}{c} UDV^T \in \mathbb{R}^{m\times n} :\ U \in \mathrm{St}(k,m),\ V \in \mathrm{St}(k,n), \\ \\ D \text{ diagonal}, \|D\|_F = 1 \end{array} \right\}. \tag{33}$$

Note that $\mathcal{S}_{\leqslant k}(m,n)$ is an algebraic set since $\mathrm{rank}(S) \leqslant k$ is equivalent to saying that all minors of S of order $k+1$ are equal to zero.

6.1. The feasible set of Problem 2 and its tangent cone

Problem 2 is defined on a feasible set that does not have a *manifold* structure. Indeed, as we will further see, the tangent cone $C_S\mathcal{S}_{\leqslant k}(m,n)$ is no longer a tangent space when $\mathrm{rank}(S) < k$.

Theorem 6.1. *Let* $S \in \mathcal{S}_{\leqslant k}(m,n)$ *be of rank* $r \leqslant k$ *and let*

$$S = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix}$$

be and ordered singular value decomposition, with $U_r \in \mathbb{R}^{m\times r}$, $U_{r\perp} \in \mathbb{R}^{m\times(m-r)}$, $V_r \in \mathbb{R}^{n\times r}$, $V_{r\perp} \in \mathbb{R}^{n\times(n-k)}$, $D_r \in \mathrm{Diag}(k,k,k)$.

The tangent cone to $\mathcal{S}_{\leqslant k}(m,n)$ *at S is given by*

$$C_S\mathcal{S}_{\leqslant k}(m,n) = \left\{ \begin{array}{l} \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} A & B \\ C & R_{k-r} \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix} : \\ \quad B, C \text{ arbitrary},\ \mathrm{tr}(AD_r) = 0, \\ \\ \quad \mathrm{rank}(R_{k-r}) \leqslant k-r \end{array} \right\} \tag{34}$$

Proof. Let us first define the surjection

$$\psi : \hat{M} \to \mathcal{S}_{\leqslant k}(m,n) : (U,D,V) \mapsto UDV^T \tag{35}$$

where $\hat{M} := \mathrm{O}(m) \times (\mathrm{Diag}(k,m,n) \cap \mathrm{Norm}(1,m,n)) \times \mathrm{O}(n)$. Let $\hat{\gamma}(t) : \mathbb{R} \mapsto \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times n}$ be an analytic curve with $\hat{\gamma}(0) = (U,D,V)$, and $\hat{\gamma}(t) \in \hat{M}$, for all $t \geqslant 0$, and γ be a curve defined as

$$\gamma(t) : \mathbb{R} \mapsto \mathbb{R}^{m\times n} : t \to \gamma(t) := \psi(\hat{\gamma}(t))\,.$$

Clearly, $\gamma(0) = UDV^T$, and $\gamma(t) \in \mathcal{S}_{\leqslant k}(m,n)$, for all $t \geqslant 0$. Moreover, since ψ is analytic, γ is also analytic, and we have

$$\dot{\gamma}(0) = \mathrm{D}\psi(U,D,V)\left[\dot{\hat{\gamma}}(0)\right].$$

This holds for any analytic curve $\hat{\gamma}$, so one can write

$$C_S\mathcal{S}_{\leqslant k}(m,n) \supseteq \bigcup_{(U,D,V)\in\psi^{-1}(S)} \mathrm{D}\psi(U,D,V)\left[C_{(U,D,V)}\hat{M}\right]. \tag{36}$$

The next step, which will be fulfilled in (48), is to show that the right-hand side of (36) is equal to the right-hand side of (34).

One can show that $\hat{M}$ is a manifold, and its tangent space at a point $\hat{S} = (U,D,V) \in \psi^{-1}(S)$ is given by

$$C_{\hat{S}}\hat{M} = \left\{ \begin{array}{c} (\Omega_U U, \xi_D, \Omega_V V) :\ \Omega_U^T + \Omega_U = 0_m\,,\ \Omega_V^T + \Omega_V = 0_n, \\ \xi_D \in \mathrm{Diag}(k,m,n)\,,\ \mathrm{tr}\left(\xi_D^T D\right) = 0 \end{array} \right\}. \tag{37}$$

Let us choose $\xi := (\xi_U, \xi_D, \xi_V) = (\Omega_U U, \xi_D, \Omega_V V) \in C_{\hat{S}}\hat{M}$, and write the differential of ψ at $\hat{S}$ in that direction

$$\mathrm{D}\psi(\hat{S})[\xi] = \xi_U DV^T + U\xi_D V^T + UD\xi_V^T = \Omega_U S + U\xi_D V^T + S\Omega_V^T\,. \tag{38}$$

Let us first change the variables and rewrite Ω_U and Ω_V as follows

$$\Omega_U = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \tilde{\Omega}_U \begin{bmatrix} U_r^T \\ U_{r\perp}^T \end{bmatrix}, \quad \text{and} \quad \Omega_V = \begin{bmatrix} V_r & V_{r\perp} \end{bmatrix} \tilde{\Omega}_V \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix}. \tag{39}$$

The conditions on Ω_U and Ω_V can directly be translated in terms of $\tilde{\Omega}_U$ and $\tilde{\Omega}_V$, *i.e.*

$$\tilde{\Omega}_U^T + \tilde{\Omega}_U = 0_m, \quad \text{and} \quad \tilde{\Omega}_V^T + \tilde{\Omega}_V = 0_n\,.$$

Moreover, since (U,D,V) is in $\psi^{-1}(S)$, there must be $P \in \mathrm{O}(k) \cap \{-1,0,1\}^{k\times k}$ and $Q \in \mathrm{O}(k) \cap \{-1,0,1\}^{k\times k}$ such that $\begin{bmatrix} P^T & 0 \\ 0 & I_{m-k} \end{bmatrix} D \begin{bmatrix} Q & 0 \\ 0 & I_{n-k} \end{bmatrix} = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}$. Let us change the variable and rewrite ξ_D as follows

$$\xi_D = \begin{bmatrix} P & 0 \\ 0 & I_{m-k} \end{bmatrix} \tilde{\xi}_D \begin{bmatrix} Q^T & 0 \\ 0 & I_{m-k} \end{bmatrix}. \tag{40}$$

The conditions on ξ_D translate to

$$\tilde{\xi}_D \in \mathrm{Diag}(k,m,n), \qquad \text{and} \qquad \mathrm{tr}\left(\tilde{\xi}_D^T \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}\right) = 0.$$

Finally, let us define $U_P := U \begin{bmatrix} P & 0 \\ 0 & I_{m-k} \end{bmatrix}$, $V_Q := V \begin{bmatrix} Q & 0 \\ 0 & I_{m-k} \end{bmatrix}$, and $d_1, \ldots, d_s, r_1, \ldots, r_s$ such that

$$D_r = \begin{bmatrix} d_1 I_{r_1} & & \\ & \ddots & \\ & & d_s I_{r_s} \end{bmatrix}$$

with $d_1 > d_2 > \cdots > d_s > 0$ and $r_1 + r_2 + \cdots + r_s = r$. Since

$$S = U_P \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} V_Q^T = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix},$$

one can show that

$$\begin{bmatrix} U_r & U_{r\perp} \end{bmatrix}^T U_P = R_U := \begin{bmatrix} R_1 & & & \\ & \ddots & & \\ & & R_s & \\ & & & R_{\perp U} \end{bmatrix} \tag{41}$$

and

$$\begin{bmatrix} V_r & V_{r\perp} \end{bmatrix}^T V_Q = R_V := \begin{bmatrix} R_1 & & & \\ & \ddots & & \\ & & R_s & \\ & & & R_{\perp V} \end{bmatrix} \tag{42}$$

where $R_1, \ldots, R_s, R_{\perp U}$, and $R_{\perp V}$ are square orthogonal matrices respectively of order $r_1, \ldots, r_s, m-r$ and $n-r$.

By using the Eqs. (39), (40), (41), and (42), one can rewrite (38) as

$$\mathrm{D}\psi(\hat{S})[\xi] = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \left(\tilde{\Omega}_U \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} + R_U \tilde{\xi}_D R_V^T + \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \tilde{\Omega}_V^T \right) \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix} . \tag{43}$$

Let us write $\tilde{\Omega}_U$, $\tilde{\xi}_D$, and $\tilde{\Omega}_V$ as follows

$$\tilde{\Omega}_U = \begin{bmatrix} \omega_U & -K_U^T \\ K_U & \omega_{\perp U} \end{bmatrix} , \quad \tilde{\xi}_D = \begin{bmatrix} \tilde{\xi}_{r_1} & & & \\ & \ddots & & \\ & & \tilde{\xi}_{r_s} & \\ & & & \tilde{\xi}_{\perp} \end{bmatrix} , \quad \text{and} \quad \tilde{\Omega}_V = \begin{bmatrix} \omega_V & -K_V^T \\ K_V & \omega_{\perp V} \end{bmatrix} .$$

The conditions on $\tilde{\Omega}_U, \tilde{\xi}_D$, and $\tilde{\Omega}_V$ translate to $\omega_U^T + \omega_U = \omega_V^T + \omega_V = 0_r$, K_U, K_V, arbitrary matrices, $\tilde{\xi}_\perp \in \mathrm{Diag}(k-r, m-r, n-r)$, and $\tilde{\xi}_{r_i} \in \mathrm{Diag}(r_i, r_i, r_i)$ such that

$$\mathrm{tr}\left(\begin{bmatrix} \tilde{\xi}_{r_1} & & \\ & \ddots & \\ & & \tilde{\xi}_{r_s} \end{bmatrix}^T D_r \right) = 0 . \tag{44}$$

One can further change the variables and rewrite ω_U and ω_V as follows

$$\omega_U = \omega_1 + \omega_2 , \quad \text{and} \quad \omega_V = \omega_1 - \omega_2 . \tag{45}$$

The previous conditions translate to $\omega_1^T + \omega_1 = \omega_2^T + \omega_2 = 0_r$. Eq. (43) then becomes

$$\mathrm{D}\psi(\hat{S})[\xi] = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} A & D_r K_V^T \\ K_U D_r & R_{\perp U}\, \tilde{\xi}_\perp\, R_{\perp V}{}^T \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix} , \tag{46}$$

with

$$A = \omega_1 D_r - D_r \omega_1 + \begin{bmatrix} R_1 \tilde{\xi}_{r_1} R_1^T & & \\ & \ddots & \\ & & R_s \tilde{\xi}_{r_s} R_s^T \end{bmatrix} + \omega_2 D_r + D_r \omega_2 . \tag{47}$$

One can first see that $\mathrm{Skew}(A) = \omega_2 D_r + D_r \omega_2$. Moreover, the skew-symmetric part of A can be made equal to any skew-symmetric matrix Ω by choosing ω_2 such that

$$[\omega_2]_{ij} = \frac{\Omega_{ij}}{[D_r]_{ii} + [D_r]_{jj}} .$$

One can further see that

$$\mathrm{Sym}(A) = \omega_1 D_r - D_r \omega_1 + \begin{bmatrix} R_1 \tilde{\xi}_{r_1} R_1^T & & \\ & \ddots & \\ & & R_s \tilde{\xi}_{r_s} R_s^T \end{bmatrix}.$$

Moreover, the symmetric part of A can be made equal to any symmetric matrix H with $\mathrm{tr}(HD_r) = 0$ by choosing ω_1 and $R_1, \ldots, R_s, \tilde{\xi}_{r_1}, \ldots, \tilde{\xi}_{r_s}$ according to the constraints: block-partition H as

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1s} \\ \vdots & & \vdots \\ H_{s1} & \cdots & H_{ss} \end{bmatrix},$$

with $H_{ij} \in \mathbb{R}^{r_i \times r_j}$ and choose

$$\omega_1 = \begin{bmatrix} 0 & \frac{H_{12}}{d_2 - d_1} & \cdots & \frac{H_{1s}}{d_s - d_1} \\ \frac{H_{21}}{d_1 - d_2} & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & \frac{H_{s-1,s}}{d_s - d_{s-1}} \\ \frac{H_{s1}}{d_1 - d_s} & \cdots & \frac{H_{s,s-1}}{d_{s-1} - d_s} & 0 \end{bmatrix} \quad \text{and} \quad R_i \tilde{\xi}_{r_i} R_i^T = H_{ii},$$

to get $\mathrm{Sym}(A) = H$. The condition $\mathrm{tr}(HD_r) = 0$ comes from (44).

These observations yield that A can be made equal to any arbitrary matrix as long as

$$\mathrm{tr}(AD_r) = 0.$$

Returning to (46), let us define $B := D_r K_V^T$, $C := K_U D_r$, and $R_{k-r} := R_{\perp U}\, \tilde{\xi}_\perp\, R_{\perp V}{}^T$. Clearly, B and C can be set to any matrix by choosing $K_V^T = D_r^{-1} B$ and $K_U = C D_r^{-1}$, and, since $\tilde{\xi}_\perp \in \mathrm{Diag}(k - r, m - r, n - r)$, R_{k-r} can be any arbitrary matrix of rank less or equal to $k - r$ by choosing $R_{\perp U}$, $\tilde{\xi}_\perp$, $R_{\perp V}{}^T$ equal to its ordered singular value decomposition. In view of (36), we conclude that

$$C_S \mathcal{S}_{\leqslant k}(m, n) \supseteq \left\{ \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} A & B \\ C & R_{k-r} \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix} : \begin{array}{l} B, C \text{ arbitrary, } \mathrm{tr}(AD_r) = 0, \\ \\ \mathrm{rank}(R_{k-r}) \leqslant k - r \end{array} \right\} \tag{48}$$

The "$\subseteq$" part follows directly from [13, Theorem 1], that is, if $t \mapsto S(t)$ is a matrix-valued curve, then there exists a decomposition $S(t) = U(t)D(t)V(t)^T$, where $U(\cdot)$ and $V(\cdot)$ are orthonormal and analytic and $D(\cdot)$ is diagonal and analytic. □

6.2. Characterization of the stationary points of Problem 2

Let now $S \in \mathcal{S}_{\leqslant k}(m, n)$ be of rank $r \leqslant k$, with an ordered singular value decomposition given by

$$S = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix}.$$

Since grad $\Phi(S) = 2\mathcal{M}^2(S)$, S is a stationary point of Problem 2 if and only if

$$\mathcal{M}^2(S) \in N_S\mathcal{S}_{\leqslant k}(m,n) = \left\{ \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} \alpha D_r & 0 \\ 0 & R_\perp \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix} : \right.$$
$$\left. \alpha \in \mathbb{R}, R_\perp = 0 \text{ if } r < k, R_\perp \text{ arbitrary if } r = k \right\}$$

$$= \left\{ \alpha S + U_{r\perp} R_\perp V_{r\perp}^T : \alpha \in \mathbb{R}, R_\perp = 0 \text{ if } r < k, R_\perp \text{ arbitrary if } r = k \right\} \tag{49}$$

6.3. Algorithm for Problem 2 and its convergence analysis

We now propose the following algorithm to find a stationary point of Problem 2.

Algorithm A2

1: $S^0 \leftarrow \mathbf{1}/\|\mathbf{1}\|_F$
2: **for** $t = 1, 2, \ldots$
3: Compute $S^t \in \mathcal{S}_k(m,n)$ according to

$$S^t \ (= U^t D^t [V^t]^T) \ \leftarrow f(S^{t-1}) := \underset{\tilde{S} \in \mathcal{S}_{\leqslant k}(m,n)}{\operatorname{argmax}} \left\langle \tilde{S}, \mathcal{M}^2(S^{t-1}) \right\rangle_F . \tag{50}$$

4: **end**

where $\mathcal{M}^2(S) = \mathcal{M}(\mathcal{M}(S))$ is defined in Eq. (1).

When $\nu(S^{t-1}) \neq 0$ (14) or $\sigma_k(\mathcal{M}^2(S^{t-1})) = 0$, the next iterate S^t is uniquely defined by (50) and is equal to $\frac{1}{\|\Sigma_1\|_F} P_1 \Sigma_1 Q_1^T$, as we will show in Lemma 6.2 below. Otherwise, (i.e., when $\sigma_k(\mathcal{M}^2(S^{t-1})) = \sigma_{k+1}(\mathcal{M}^2(S^{t-1})) > 0$), S^t is no longer uniquely defined by (50); in this case, S^t is chosen arbitrarily in $\underset{\tilde{S} \in \mathcal{S}_{\leqslant k}(m,n)}{\operatorname{argmax}} \left\langle \tilde{S}, \mathcal{M}^2(S^{t-1}) \right\rangle_F$. This case was never observed in our numerical experiments, but if it did, a possible choice would have been $S^t := \frac{1}{\|\Sigma_1\|_F} P_1 \Sigma_1 Q_1^T$, where P_1, Σ_1 and Q_1 are returned by the SVD function.

We first state a few intermediate results in order to prove convergence of Algorithm A2 to the stationary points of Problem 2 (see Theorem 6.6).

Lemma 6.2. *Let $M \in \mathbb{R}^{m\times n}$ and its ordered singular value decomposition*

$$M = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = P \Sigma Q^T \tag{51}$$

with $P_1 \in \mathbb{R}^{m\times k}$, $P_2 \in \mathbb{R}^{m\times(m-k)}$, $Q_1 \in \mathbb{R}^{n\times k}$, $Q_2 \in \mathbb{R}^{n\times(n-k)}$, $\Sigma_1 \in \mathbb{R}^{k\times k}$ and $\Sigma_2 \in \mathbb{R}^{(m-k)\times(n-k)}$. Then

$$\max_{S = UDV^T \in \mathcal{S}_{\leqslant k}(m,n)} \langle S, M \rangle_F = \operatorname{tr}\left(\hat{\Sigma}_1 \Sigma_1\right). \tag{52}$$

where $\hat{\Sigma}_1 := \Sigma_1/\|\Sigma_1\|_F$.

Moreover, if $\nu := \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2) > 0$ or if $\|\Sigma_2\|_F = 0$, then the maximizing solution S is unique and equals $P_1 \hat{\Sigma}_1 Q_1^T$.

Proof. We have

$$\operatorname{tr}\left(DV^T MU\right) \leqslant \sum_{i=1}^{k} \sigma_i(DV^T MU) \leqslant \sum_{i=1}^{k} \sigma_i(D)\,\sigma_i(V^T MU) \leqslant \sum_{i=1}^{k} \sigma_i(D)\,\sigma_i(\Sigma_1) \leqslant \operatorname{tr}\left(\hat{\Sigma}_1 \Sigma_1\right) \quad (53)$$

according to [10, Formulas 3.1.10b, and Lemma 3.3.1]. The upper bound is reached for $U = P_1, D = \hat{\Sigma}_1$ and $V = Q_1$. The uniqueness of $P_1 \hat{\Sigma}_1 Q_1^T$ is a well known result discussed, *e.g.*, in [10, Theorem 3.1.1 and 3.1.1']. □

Theorem 6.3. *Let $S \in \mathcal{S}_{\leqslant k}(m, n)$ and $\mathcal{M}^2(S)$ have an ordered singular value decomposition*

$$\mathcal{M}^2(S) = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} = P \Sigma Q^T , \quad (54)$$

with $P_1 \in \mathbb{R}^{m\times k}$, $P_2 \in \mathbb{R}^{m\times(m-k)}$, $Q_1 \in \mathbb{R}^{n\times k}$, $Q_2 \in \mathbb{R}^{n\times(n-k)}$, $\Sigma_1 \in \mathbb{R}^{k\times k}$ and $\Sigma_2 \in \mathbb{R}^{(m-k)\times(n-k)}$, and $S^+ := f(S)$, with f the function defined in Algorithm A2. Then

$$\Phi(S^+) - \Phi(S) \geqslant \left\| \mathcal{M}(S^+ - S) \right\|_F^2 , \quad (55)$$

with Φ the function defined in Eq. (2), and $\nu = \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2)$. Moreover, if $\nu = \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2) > 0$, then the inequality becomes an equality iff S is a fixed point of the iteration, i.e. $S^+ = S$.

Proof. Adding and subtracting S to S^+ and using the self-adjointness of $\mathcal{M}$ yield

$$\Phi(S^+ - S + S) = \left\| \mathcal{M}(S^+ - S) \right\|_F^2 + 2 \left\langle S^+ - S, \mathcal{M}^2(S) \right\rangle_F + \Phi(S) . \quad (56)$$

One can observe that $\left\langle S^+ - S, \mathcal{M}^2(S) \right\rangle_F$ is positive since S^+ maximizes the scalar product with $\mathcal{M}^2(S)$, and hence

$$\left\langle S^+, \mathcal{M}^2(S) \right\rangle_F \geqslant \left\langle S, \mathcal{M}^2(S) \right\rangle_F . \quad (57)$$

The Eq. (56) combined with (57) gives the desired result.

Moreover, if $\nu = \sigma_{min}(\Sigma_1) - \sigma_{max}(\Sigma_2) > 0$, Lemma 6.2 ensures that the maximizing solution is unique. Hence, unless $S^+ = S$, the last term of the right hand side is strictly positive. □

Lemma 6.4. *If S is a fixed point of Algorithm A2 then S is a stationary point of Problem 2.*

Proof. Let $S = UDV^T$ be a fixed point, *i.e.* $S^+ = U^+ D^+ [V^+]^T = UDV^T$ and the ordered singular value decomposition of $\mathcal{M}^2(UDV^T)$ is

$$\mathcal{M}^2(UDV^T) = \begin{bmatrix} U^+ & U_\perp \end{bmatrix} \begin{bmatrix} \alpha D^+ & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} [V^+]^T \\ V_\perp^T \end{bmatrix} , \quad (58)$$

with $\sigma_{min}(\alpha D^+) \geqslant \sigma_{max}(\Sigma_2)$. Remark that if the rank of $U^+ D^+ [V^+]^T$ is lower than k, then $\sigma_{min}(\alpha D^+) = 0$ and $\Sigma_2 = 0_{m-k,n-k}$.

Let us now remind that the gradient of Φ at a point S is $2\mathcal{M}^2(S)$. One can verify that this expression is in the normal cone given by Eq. (49) and is hence a stationary point. □

Notice that all stationary points are not fixed points since Σ_2 has to be such that $\sigma_{min}(\alpha D^+) \geqslant \sigma_{max}(\Sigma_2)$.

Lemma 6.5. *Let S be a nonstationary point of Problem 2. There exists an $\epsilon > 0$ such that for all $\|S_\epsilon\|_F < \epsilon$, $S + S_\epsilon$ is not a stationary point.*

Proof. Since $S = U_r D_r V_r^T$ is not a stationary point, we have

$$2\,\mathcal{M}^2(U_r D_r V_r^T) = \begin{bmatrix} U_r & U_{r\perp} \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{r\perp}^T \end{bmatrix}, \tag{59}$$

with either, if $r = k$,

$$\delta_{max} := \max\left(\left\| M_{11} - \frac{\|M_{11}\|_F}{\|D_r\|_F} D_r \right\|_F^2, \|M_{12}\|_F^2, \|M_{21}\|_F^2 \right) > 0, \tag{60}$$

or, if $r < k$,

$$\delta_{max} := \max\left(\left\| M_{11} - \frac{\|M_{11}\|_F}{\|D_r\|_F} D_r \right\|_F^2, \|M_{12}\|_F^2, \|M_{21}\|_F^2, \|M_{22}\|_F^2 \right) > 0. \tag{61}$$

Since $\mathcal{M}^2(\cdot)$ is a continuous mapping, for all $\delta > 0$ there always exists $\epsilon(\delta) > 0$ such that for all $\|S_\epsilon\|_F < \epsilon(\delta)$, we have $\left\| \mathcal{M}^2(S + S_\epsilon) - \mathcal{M}^2(S) \right\|_F < \delta$. A reasoning similar to the one held for Lemma 5.4, when one chooses $\delta \leqslant \delta_{max}$, yields the desired result. □

The next theorem states the main convergence result for Algorithm A2. Note that the existence of an accumulation point is guaranteed by the fact that the iteration evolves on the compact set $\mathcal{S}_{\leqslant k}(m, n)$. In practice, in our experiments, the sequences of iterates always had a single accumulation point S', with $\nu(S') \neq 0$ or $\sigma_k(\mathcal{M}^2(S')) = 0$; by virtue of the next theorem, it thus follows that S' is a stationary point of Problem 2. Moreover, since the iteration is an ascent iteration for Φ, convergence to stationary points that are not local maxima is not expected to occur in practice.

Theorem 6.6. *Let S', with $\nu(S') > 0$ (14) or $\sigma_k(\mathcal{M}^2(S')) = 0$, be an accumulation point of the sequence $\{S_i\}$ constructed by A2. Then S' is a stationary point of Problem 2.*

Proof. A reasoning similar to the one held for Theorem 5.5, along with Theorem 6.3 and Lemmas 6.5, 6.4, yields the desired result. In this parallelism Theorem 5.2 is replaced by Theorem 6.3, which lacks the term $\nu\sqrt{k}\,\|S^+ - S\|_F^2$. However, Theorem 6.3 still ensures that $\Phi(S^+) - \Phi(S) > 0$ holds whenever S is not a fixed point of the algorithm. The continuity argument for S_+ still holds since we now simply have $S_+ = L_M L_M^T M R_M R_M^T$. Thus the reasoning in the proof of Theorem 5.5 still holds for Problem 2, and the result follows. □

7. Complexity analysis

Let A and B respectively contain $m\alpha$ and $n\beta$ nonzero elements.

Let us first consider the complexity of one step of Algorithm A0, *i.e.*

$$S^t \leftarrow \frac{AS^{t-1}B^T + A^T S^{t-1} B}{\left\| AS^{t-1}B^T + A^T S^{t-1} B \right\|_F}$$

Assuming that S^{t-1} is a dense matrix, the products AS^{t-1} and $A^T S^{t-1}$ require less than $2mn\alpha$ flops each, while the subsequent products $(AS^{t-1})B^T$ and $(A^T S^{t-1})B$ require less than $2mn\beta$ flops each. The sum and the calculation of the Frobenius norm requires $2mn$ flops, while the scaling requires one division and nm multiplications. Then, the total complexity per iteration step is of the order of $4(\alpha + \beta)mn$ flops.

Let us now consider the complexity of one step of Algorithms A1 and A2. In these algorithms, the rank of S is at most k. Hence, in practice, we do not really work with $S \in \mathbb{R}^{m\times n}$ itself but with its singular value factorization $(U, D, V) \in \mathbb{R}^{m\times k} \times \mathrm{Diag}(k, k, k) \times \mathbb{R}^{n\times k}$. When k is small, the space required to

store the factors of S (i.e. $mk + k + nk$ elements) is smaller than the one required to store S itself (i.e. mn elements). Similarly, in practice, we do not really compute $\mathcal{M}^2(S) \in \mathbb{R}^{m\times n}$ itself but its singular value factorization. We now show how we compute the factors of the singular value decomposition of $\mathcal{M}^2(S)$. Notice first that $\mathcal{M}^2(UDV^T)$ can be written as $U_A\,(I_4 \otimes D)\,V_B^T$ with

$$U_A := \begin{bmatrix} U_A^T A \\ U_A^T A A^T \\ U_A^T A A^T A \\ U_A^T A A^T A A^T \end{bmatrix}^T, \quad (I_4 \otimes D) := \begin{bmatrix} D & & & \\ & D & & \\ & & D & \\ & & & D \end{bmatrix}, \text{ and } V_B := \begin{bmatrix} V_B^T B \\ V_B^T B B^T \\ V_B^T B B^T B \\ V_B^T B B^T B B^T \end{bmatrix}^T. \tag{62}$$

One can further compute $Q_A \in \mathbb{R}^{m\times 4k}$ and $R_A \in \mathbb{R}^{4k\times 4k}$ (resp. $Q_B \in \mathbb{R}^{n\times 4k}$ and $R_B \in \mathbb{R}^{4k\times 4k}$), the factors of the QR decomposition of U_A (resp. V_B), *i.e.* $Q_A R_A = U_A$, with $Q_A^T Q_A = I_{4k}$ and $[R_A]_{ij} = 0$ for all $i > j$, and rewrite Eq. (62) as $\mathcal{M}^2(UDV^T) = Q_A R_A\,(I_4 \otimes D)\,R_B^T Q_B^T$. We further compute $(\bar{U}, \bar{D}, \bar{V}) \in \mathbb{R}^{4k\times 4k} \times \mathrm{Diag}(4k, 4k, 4k) \times \mathbb{R}^{n\times 4k}$, the factors of the singular value decomposition of $R_A\,(I_4 \otimes D)\,R_B^T$. Finally, one can write $\mathcal{M}^2(UDV^T) = Q_A \bar{U}\,\bar{D}\,\bar{V}^T Q_B^T$, and the factors of its singular value decomposition are given by $(Q_A\bar{U}, \bar{D}, Q_B\bar{V})$. And, eventually, Algorithm A2 can be rewritten as follows.

$(U^0, D^0, V^0) \leftarrow SVD_k(\mathbf{1}/\,\|\mathbf{1}\|_F)$;
for $t = 1, 2, \ldots$
  $U' \leftarrow [AU^{t-1}D^{t-1}, A^T U^{t-1} D^{t-1}] \in \mathbb{R}^{m\times 2k}$;
  $U'' \leftarrow [AU', A^T U'] \in \mathbb{R}^{m\times 4k}$;
  $V' \leftarrow [BV^{t-1}, B^T V^{t-1}] \in \mathbb{R}^{n\times 2k}$;
  $V'' \leftarrow [BV', B^T V'] \in \mathbb{R}^{n\times 4k}$;
  $(Q_U, R_U) \leftarrow QR(U'') \in \mathbb{R}^{m\times 4k} \times \mathbb{R}^{4k\times 4k}$;
  $(Q_V, R_V) \leftarrow QR(V'') \in \mathbb{R}^{n\times 4k} \times \mathbb{R}^{4k\times 4k}$;
  $(U''', D''', V''') \leftarrow SVD_k(R_U R_V^T) \in \mathbb{R}^{m\times k} \times \mathbb{R}^{k\times k} \times \mathbb{R}^{n\times k}$;
  $(U^t, D^t, V^t) \leftarrow (Q_U U''', \frac{D'''}{\|D'''\|}, Q_V V''')$;
end

We now consider the complexity of computing the singular value decomposition of $\mathcal{M}^2(UDV^T)$ with UDV^T a matrix of rank k. The products AU, A^TU, AU', and A^TU' require less than $2mk\alpha$ flops each, whereas the products BV, B^TV, BV', and B^TV' require less than $2nk\beta$ flops each. The QR factorization of a matrix $M \in \mathbb{R}^{m\times k}$ requires $4mk^2 - \frac{4}{3}k^3$ flops [14, p. 337] and subsequently computing Q_AR_A and Q_BR_B require receptively less than $4m(4k)^2 - \frac{4}{3}(4k)^3$ and $4n(4k)^2 - \frac{4}{3}(4k)^3$ flops. The product $R_A\,(I_4 \otimes D)\,R_B^T$ require less than $2(4k)^3$ flops. The complexity of computing the singular value decomposition of $\bar{U}\bar{D}\bar{V} \in \mathbb{R}^{4k\times 4k}$ up to a given precision is of the order of k^3 flops. Finally, the products $Q_A\bar{U}$ and $Q_B\bar{V}$ require less than $2m(4k)^2$ and $2n(4k)^2$ flops each. Hence, in total, computing the singular value decomposition of $\mathcal{M}^2(UDV^T)$ requires less than $8mk\alpha + 8nk\beta + 96mk^2 + 96nk^2 + O(k^3)$ flops. Let now $\mathcal{M}^2(S^{t-1})$ admit the following ordered singular value decomposition

$$\mathcal{M}^2(S^{t-1}) = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$$

with $P_1 \in \mathbb{R}^{m\times k}$, $Q_1 \in \mathbb{R}^{n\times k}$, and $\Sigma_1 \in \mathbb{R}^{k\times k}$. According to Lemmas 5.1 and 6.2, one step of Algorithms A1 and A2 consists in choosing S^t respectively equal to $P_1\hat{I}_kQ_1^T$ and $P_1\hat{\Sigma}_1Q_1^T$. The number of operations required to compute one step of Algorithms A1 and A2 is then equal to the one required to compute the singular value decomposition of $\mathcal{M}^2(UDV^T)$ which costs

$$8mk\alpha + 8nk\beta + 96mk^2 + 96nk^2 + O(k^3) \text{ flops.}$$

Let us remind that one step of Algorithms A1 and A2 are low-rank approximations of two steps of A0 which costs $8(\alpha + \beta)mn$ flops.

8. Experiments

We look at the performances of our method to compute self-similarity matrices. This means that A and B are equal. In other words, the self-similarity matrix expresses how a node of a graph is similar to other nodes of the same graph. We ran several experiments to compute low-rank approximations of self-similarity matrices on random graphs. Results about the average computational time, and

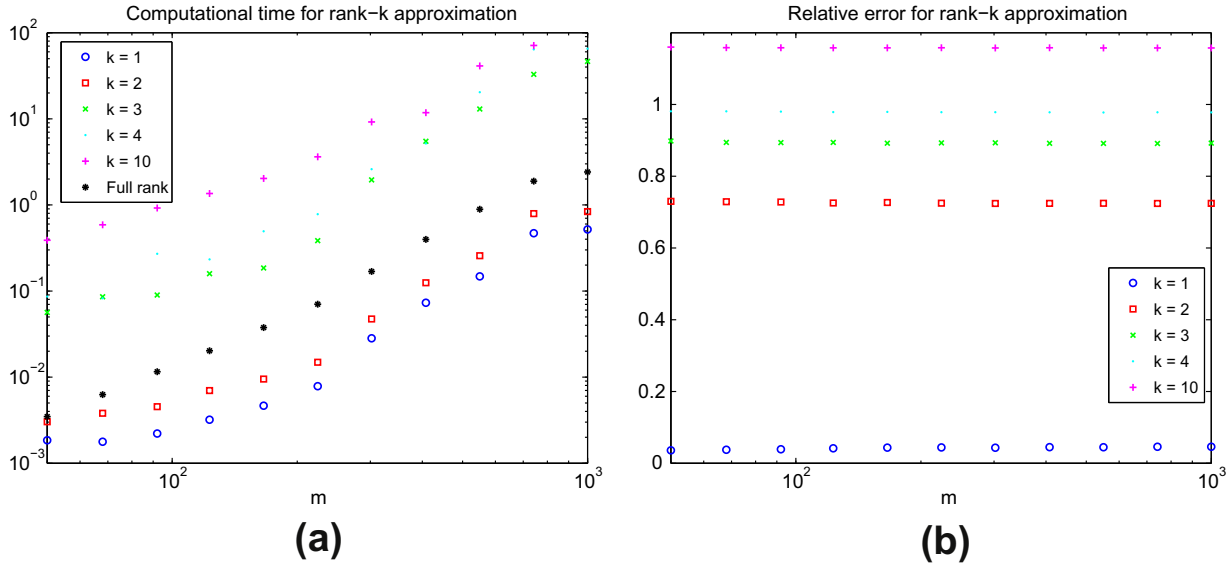

Fig. 1. We compute rank-k approximations with exactly k identical nonzero eigenvalues of the self-similarity matrix of a connected Erdós–Rényi graph with probability $10/m$, where m is the order of this graph. The graph is built such that the average number of outgoing edges of a node is 10. The algorithm stops when $\|\Delta_S\|_F \leqslant 10^{-6}\,\|S\|_F$. The full rank results are obtained using Algorithm A0 which was investigated in [1]. (a) shows the average computational time versus m, the order of this graph, (b) shows the average relative error of the rank-k approximations of the self-similarity matrix of a connected random graph versus m.

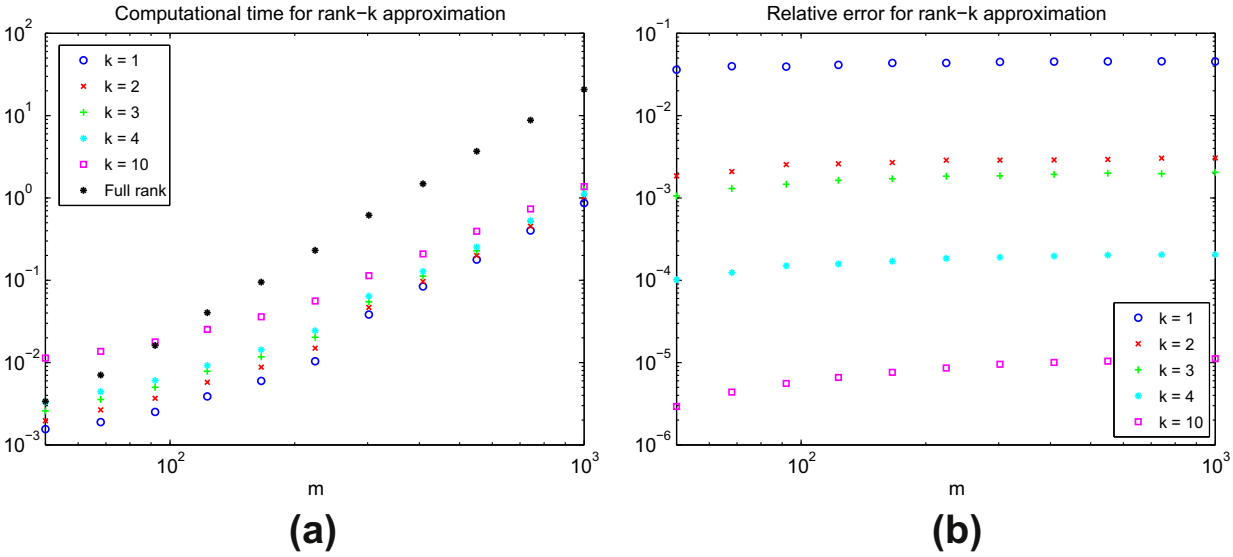

Fig. 2. We compute rank-k approximations with at most k nonzero eigenvalues of the self-similarity matrix of a connected random graph. The graph is built such that the average number of outgoing edges of a node is 10. The algorithm stops when $\|\Delta_S\|_F \leqslant 10^{-6}\,\|S\|_F$. The full rank results are obtained using Algorithm A0 which was investigated in [1]. (a) shows the average computational time versus m, the order of this graph, (b) shows the average relative error of the rank-k approximations of the self-similarity matrix of a connected random graph versus m.

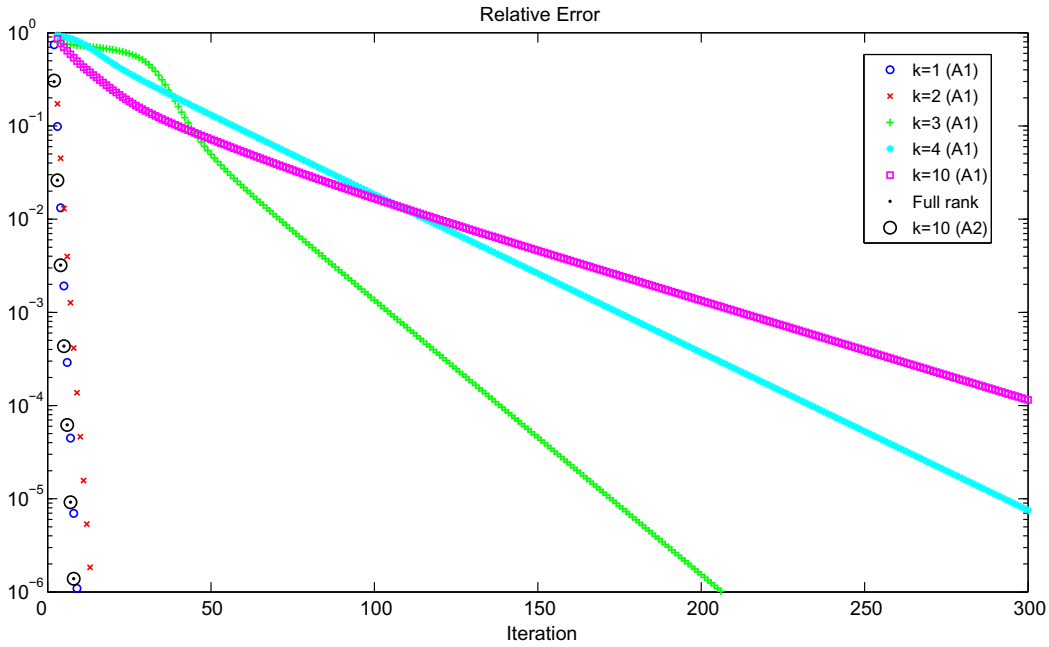

Fig. 3. The (A1) (resp. (A2)) lines refer to the experiment of Fig. 1 (resp. 2) which shows results for Algorithm A1 (resp. A2). The graph shows the relative distance between an iterate and the extremal point of the corresponding experiment $\left(\frac{\|S^t - S^\infty\|}{\|S^\infty\|}\right)$ versus t, the number of iterations.

the average relative error with respect to the full rank self-similarity matrices for exactly k identical nonzero eigenvalues and for at most k nonzero eigenvalues are shown respectively in Figs. 1 and 2. Results about the speed of convergence are shown in Fig. 3. As expected, we clearly notice that the smaller the rank of the approximation k, the smaller the computational time. We further notice that, when the order of the graph m increases, the algorithm for low rank approximation converges faster than the full rank algorithm. As far as the relative error is concerned, we observe that it does not vary much with m, the order of this graph. For exactly k identical nonzero eigenvalues, this error increases when the rank of the approximation increases. Indeed, rank 1 approximation have a relative error about 0.05 whereas higher rank approximation are about 0.7 up to 1.2 ! This reveals that the equal eigenvalues assumption is not adequate for this class of graphs. For at most k nonzero eigenvalues, the results are much more satisfactory since the error decreases when the rank of the approximation increases.

9. Conclusions

In this paper, we have considered two optimization problems (Problems 1 and 2) whose solutions are low-rank approximations of the similarity matrix **S** introduced by Blondel et al. in [1]. The cost functions of Problems 1 and 2 are the same as the one presented in Eq. (2) whereas their feasible sets are respectively set to $\mathcal{S}_k(m, n)$ and $\mathcal{S}_{\leqslant k}(m, n)$ instead of $\mathcal{S}(m, n)$. We have first characterized the stationary points of Problems 1 and 2. Then we have considered Algorithms A1 and A2 and proved that their accumulation points are stationary points of respectively Problems 1 and 2. Next, we have analyzed the complexity of one step of Algorithms A1 and A2 and compared them to the complexity of Algorithm A0 used to compute the original similarity matrix **S** introduced by Blondel et al. We have further performed numerical experiments and considered the performances of Algorithms A1 and A2. Finally, we have concluded that Problem 1 is not adequate to find a low rank approximation of the optimization problem presented in Eq. (2) since the relative error of the approximations of rank bigger than 2 is about 100%. On the other hand, the solution of Problem 2 appropriately approaches the solution of the optimization problem presented in Eq. (2). As expected, we have observed that the relative error of approximation decreases when the rank of the approximation increases, and the ratio between the time until convergence of Algorithm A2 and the time until convergence of Algorithm A0 decreases as m and n (the size of the problem) grow.

Acknowledgment

The authors are grateful to the anonymous referees for carefully checking the paper and for providing several helpful comments.

References

[1] V.D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P. Van Dooren, A measure of similarity between graph vertices: applications to synonym extraction and Web searching, SIAM Rev. 46 (4) (2004) 647–666.
[2] A.T. Balaban, Applications of graph theory in chemistry, J. Chem. Inform. Comput. Sci. 25 (3) (1985) 334–343.
[3] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632. doi:10.1145/324133.324140.
[4] M. Heymans, A.K. Singh, Deriving phylogenetic trees from the similarity analysis of metabolic pathways, Bioinformatics 19 (1) (2003) i138–i146.
[5] C. Fraikin, Y. Nesterov, P. Van Dooren, Optimizing the coupling between two isometric projections of matrices, SIAM J. Matrix Anal. Appl. 30 (1) (2008) 324–345. doi:10.1137/050643878.
[6] D.B. O'Shea, L.C. Wilson, Limits of tangent spaces to real surfaces, Amer. J. Math. 126 (5) (2004) 951–980.
[7] J.P. Aubin, Applied Functional Analysis, Pure and Applied Mathematics (New York), second ed., Wiley-Interscience, New York, 2000 (with exercises by Bernard Cornet and Jean-Michel Lasry, Translated from the French by Carole Labrousse).
[8] T.P. Cason, P.A. Absil, P. Van Dooren, Comparing two matrices by means of isometric projections, in: P. VanDooren, S. Bhattacharyya, R. Chan, V. Olshevsky, A. Routray (Eds.), Numerical Linear Algebra in Signals, Systems and Control, Lecture Notes in Electrical Engineering, vol. 80, Springer-Verlag, 2011, pp. 77–93. doi:10.1007/978-94-007-0602-6_4.
[9] P.A. Absil, R. Mahony, R. Sepulchre, Optimization Algorithms on Matrix Manifolds, Princeton University Press, New Jersey, 2008.
[10] R. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.
[11] E. Polak, Computational Methods in Optimization – A Unified Approach, Mathematics in Science and Engineering, vol. 77, Academic Press, New York, 1971.
[12] G.W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, SIAM Rev. 15 (1973) 727–764.
[13] A. Bunse-Gerstner, R. Byers, V. Mehrmann, N.K. Nichols, Numerical computation of an analytic singular value decomposition of a matrix valued function, Numer. Math. 60 (1) (1991) 1–39.
[14] N.J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.