Numerische Mathematik (2020) 144:729–749
https://doi.org/10.1007/s00211-020-01100-8

Numerische Mathematik

Nonlinear power-like iteration by polar decomposition and its application to tensor approximation

Bo Dong[1] · Nan Jiang[2] · Moody T. Chu[3]

Received: 22 May 2018 / Revised: 21 June 2019 / Published online: 8 February 2020

Abstract

Low rank tensor approximation is an important subject with a wide range of applications. Most prevailing techniques for computing the low rank approximation in the Tucker format often first assemble relevant factors into matrices and then update by turns one factor matrix at a time. In order to improve two factor matrices simultaneously, a special system of nonlinear matrix equations over a certain product Stiefel manifold must be resolved at every update. The solution to the system consists of orbit varieties invariant under the orthogonal group action, which thus imposes challenges on its analysis. This paper proposes a scheme similar to the power method for subspace iterations except that the polar decomposition is used as the normalization process and that the iteration can be applied to both the orbits and the cross-sections. The notion of quotient manifold is employed to factor out the effect of orbital solutions. The dynamics of the iteration is completely characterized. An isometric isomorphism between the tangent spaces of two properly identified Riemannian manifolds is established to lend a hand to the proof of convergence.

Bo Dong: This work was supported in part by the National Natural Science Foundation of China under Grant 11871136, the Chinese Scholarship Council and the Fundamental Research Funds for the Central Universities.
Nan Jiang: This work was supported in part by the Chinese Scholarship Council.
Moody T. Chu: This work was supported in part by the National Science Foundation under Grants DMS-1316779 and DMS-1912816.

✉ Nan Jiang
jngrace@hrbeu.edu.cn

Bo Dong
dongbo@dlut.edu.cn

Moody T. Chu
chu@math.ncsu.edu

1 School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, Liaoning, China

2 College of Automation, Harbin Engineering University, Harbin 150001, Heilongjiang, China

3 Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

Mathematics Subject Classification 65F10 · 15A24 · 65H10 · 15A72 · 58D19

1 Introduction

Let $\mathscr{S}(p, q)$, $p \geq q$, denote the Stiefel manifold

$$\mathscr{S}(p, q) := \{Q \in \mathbb{R}^{p\times q} | Q^\top Q = \mathrm{I}_q\}. \tag{1}$$

Suppose that $\mathcal{A} : \mathscr{S}(p_2, q_2) \times \mathscr{S}(p_2, q_2) \to \mathbb{R}^{p_1\times p_1}$ and $\mathcal{B} : \mathscr{S}(p_1, q_1) \times \mathscr{S}(p_1, q_1) \to \mathbb{R}^{p_2\times p_2}$ are two given bilinear maps with certain symmetric properties which will be specified in the subsequent discussion. This paper concerns about solving the nonlinear matrix equations

$$\begin{cases} \mathcal{A}(Q_2, Q_2)Q_1 = Q_1P_1, \\ \mathcal{B}(Q_1, Q_1)Q_2 = Q_2P_2, \end{cases} \tag{2}$$

where the products on the right-hand side of (2) are the polar decompositions of the matrices on the left-hand side, respectively. Such a problem arises as a basic computational mechanism when solving the so called Tucker nearest problem. We can get rid of the reference to P_1 and P_2 by the substitutions of

$$\begin{cases} P_1 = Q_1^\top \mathcal{A}(Q_2, Q_2)Q_1, \\ P_2 = Q_2^\top \mathcal{B}(Q_1, Q_1)Q_2, \end{cases}$$

provided that the expressions on the right side above are known to be symmetric and positive semi-definite. The system (2) is thus cast as an under-determined polynomial system

$$\begin{cases} \mathcal{A}(Q_2, Q_2)Q_1 = Q_1{Q_1}^\top \mathcal{A}(Q_2, Q_2)Q_1, \\ \mathcal{B}(Q_1, Q_1)Q_2 = Q_2{Q_2}^\top \mathcal{B}(Q_1, Q_1)Q_2, \\ {Q_1}^\top Q_1 = \mathrm{I}_{q_1}, \\ {Q_2}^\top Q_2 = \mathrm{I}_{q_2}, \end{cases} \tag{3}$$

in the unknowns $Q_1 \in \mathbb{R}^{p_1\times q_1}$ and $Q_2 \in \mathbb{R}^{p_2\times q_2}$, whose solutions will be shown to have rich algebraic properties known as orbital varieties [4].

To motivate where the system of nonlinear matrix equations arises and why it is useful, we briefly outline the background of the Tucker nearest problem. The low rank approximation is one principal tool of great power and interest when dealing with entangled and large-scale data. In recent years, scientists and practitioners turn to higher-dimensional arrays, i.e., tensors, for the advantage of greater descriptive flexibility and more fine-grained data collection. Inevitably, the sizes grow rapidly and content analytics becomes a much more challenging task. Low rank tensor approximations, if done properly, are appealing for at least the benefits of storage saving and feature finding. Among a variety of structured or unstructured low rank tensor approximations, the Tucker nearest problem is one of the most fundamental and important formulations with versatile applications.

For convenience, let $[\![k]\!]$ denote the set $\{1, \ldots, k\}$. Let $\mathbf{a} \circ \mathbf{b}$ denote the tensor product enumerated in such a way that, if $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$, then its vectorization is given by

$$\mathbf{vec}(\mathbf{a} \circ \mathbf{b}) = \mathbf{vec}(\mathbf{a}\mathbf{b}^\top) = [a_1 b_1, a_2 b_1, \ldots, a_m b_1, a_1 b_2, \ldots, a_m b_n]^\top.$$

Given an order-k tensor $T \in \mathbb{R}^{I_1 \times \ldots \times I_k}$ and a fixed rank parameter $\mathbf{r} = (r_1, \ldots, r_k)$ of positive integers, the Tucker nearest problem is to find scalars $c_{j_1,\ldots,j_k}$ and vectors $\mathbf{v}_{j_\ell}^{(\ell)} \in \mathbb{R}^{I_\ell}$, $j_\ell = 1, \ldots, r_\ell$, $\ell = 1, \ldots, k$, such that the objective function

$$h(C, V^{(1)}, \ldots, V^{(k)}) := \left\| \sum_{j_1=1}^{r_1} \ldots \sum_{j_k=1}^{r_k} c_{j_1,\ldots,j_k} \mathbf{v}_{j_1}^{(1)} \circ \ldots \circ \mathbf{v}_{j_k}^{(k)} - T \right\|_F,$$

subject to the condition that

$$V^{(\ell)} := [\mathbf{v}_1^{(\ell)}, \ldots, \mathbf{v}_{r_\ell}^{(\ell)}] \in \mathscr{S}(I_\ell, r_\ell), \quad \ell \in [\![k]\!],$$

is minimized. The collections

$$C := [c_{j_1,\ldots,j_k}] \in \mathbb{R}^{r_1 \times \ldots \times r_k}$$

and $V^{(\ell)}$ are referred to respectively as the core tensor and the factor matrix in the literature. It can be argued that, given $V^{(\ell)} \in \mathscr{S}(I_\ell, r_\ell)$, $\ell \in [\![k]\!]$, the optimal C is given by [1, Formula (12)]

$$\mathbf{vec}(C) = ({V^{(k)}}^\top \otimes \ldots \otimes {V^{(1)}}^\top)\mathbf{vec}(T), \tag{4}$$

where $\otimes$ stands for the conventional Kronecker product. In this way, the Tucker nearest problem is equivalent to the problem of maximizing the Frobenius norm of the core tensor C defined in (4) subject to the constraint that $V^{(\ell)} \in \mathscr{S}(I_\ell, r_\ell)$, $\ell \in [\![k]\!]$.

A conventional approach to maximizing $\|C\|_F$ is to matricize (4) in terms of the mode-d folding. Specifically, it can be shown that [1, Formula (11)]

$$C_{(d)} = {V^{(d)}}^\top \underbrace{T_{(d)}(V^{(k)} \otimes \ldots \otimes V^{(d+1)} \otimes V^{(d-1)} \otimes \ldots \otimes V^{(1)})}_{\Upsilon^{[d]}},$$

where d can be any integer in the set $[\![k]\!]$ and $T_{(d)} \in \mathbb{R}^{I_d \times \prod_{\ell \neq d} I_\ell}$ is a rearrangement of T such that the element $\tau_{i_1,\ldots,i_{d-1},i,i_{d+1},\ldots,i_k}$ of the tensor T is placed at the (i, j) entry of the matrix $T_{(d)}$ with

$$j := i_1 + \sum_{\zeta \neq d} (i_\zeta - 1) \prod_{\eta=1}^{\zeta-1} I_\eta.$$

Such a reformulation motivates the notion of alternating directions that updates one factor matrix $V^{(d)}$ at a time for each $d \in [\![k]\!]$. More specifically, for a fixed $\Upsilon^{[d]} \in \mathbb{R}^{I_d \times \prod_{\ell \neq d} r_\ell}$, we solve the optimization problem

$$\max_{V^{(d)}\in\mathscr{S}(I_d,r_d)} f(V^{(d)}) := \frac{1}{2}\left\|V^{(d)T}\Upsilon^{[d]}\right\|_F^2 \tag{5}$$

for $V^{(d)}$ and repeat this procedure by varying d in turn. It can easily be derived that the optimal $V^{(d)} \in \mathscr{S}(I_d, r_d)$ is given by the first r_d left singular vectors of the matrix $\Upsilon^{[d]}$.

In a similar vein, it is possible to reorganize the elements in such a way that we can deal with the update of two factor matrices concurrently. The idea is that if $[\![k]\!] = \boldsymbol{\alpha}\cup\boldsymbol{\beta}$ is a partition of $[\![k]\!]$ with $\boldsymbol{\alpha} := \{\alpha_1, \alpha_2\}$ and $\boldsymbol{\beta} := \{\beta_1, \ldots, \beta_{k-2}\}$, then we can rearrange C into such a matrix that

$$C_{(\boldsymbol{\alpha},\boldsymbol{\beta})} := (V^{(\alpha_2)} \otimes V^{(\alpha_1)})^\top \underbrace{T_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(V^{(\beta_{k-2})} \otimes \ldots \otimes V^{(\beta_1)})}_{\Upsilon^{(\boldsymbol{\alpha},\boldsymbol{\beta})}},$$

where $T_{(\boldsymbol{\alpha},\boldsymbol{\beta})} \in \mathbb{R}^{\prod_{i=1}^{2} I_{\alpha_i}\times\prod_{j=1}^{k-2} I_{\beta_j}}$ is a multi-mode folding of T whose precise definition will not be described here but can be found in [1, Formula (10)]. In the same spirit of (5), to maximize $\|C_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\|_F$ we need to update the two factor matrices $(V^{(\alpha_1)}, V^{(\alpha_2)})$ per fixed $\Upsilon^{(\boldsymbol{\alpha},\boldsymbol{\beta})}$ and alternate the partition $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ throughout $[\![k]\!]$. The recurring scheme in the sequence of alternating updates is to solve the constrained optimization problems

$$\max_{Q_1\in\mathscr{S}(p_1,q_1), Q_2\in\mathscr{S}(p_2,q_2)} g(Q_1, Q_2) := \frac{1}{2}\|\Omega(Q_2 \otimes Q_1)\|_F^2, \tag{6}$$

where $\Omega \in \mathbb{R}^{m\times p_1 p_2}$ with $m \geq p_1 p_2$ is a fixed matrix. With this background in mind, we now characterize more specifically the first order optimality condition of (6) that leads to the nonlinear matrix equation (2).

Let $\Theta := \Omega^\top\Omega(Q_2 \otimes Q_1) \in \mathbb{R}^{p_1 p_2\times q_1 q_2}$ be partitioned as a $p_2 \times q_2$ block matrix with blocks of size $p_1 \times q_1$, which is then regarded as an order-4 tensor $\Theta = [\theta_{i_1 i_2 j_1 j_2}]$ in $\mathbb{R}^{p_1\times q_1\times p_2\times q_2}$. For $A = [a_{i_1 i_2}] \in \mathbb{R}^{p_1\times q_1}$ and $B = [b_{j_1 j_2}] \in \mathbb{R}^{p_2\times q_2}$, define the multiplications

$$\Theta \circledast_{\{1,2\}} B := \left[\sum_{j_1=1}^{p_2}\sum_{j_2=1}^{q_2} \theta_{i_1 i_2 j_1 j_2} b_{j_1 j_2}\right] \in \mathbb{R}^{p_1\times q_1} \tag{7}$$

and

$$\Theta \circledast_{\{3,4\}} A := \left[\sum_{i_1=1}^{p_1}\sum_{i_2=1}^{q_1} \theta_{i_1 i_2 j_1 j_2} a_{i_1 i_2}\right] \in \mathbb{R}^{p_2\times q_2}, \tag{8}$$

respectively.

Lemma 1 *The critical point $(Q_1, Q_2) \in \mathscr{S}(p_1, q_1)\times\mathscr{S}(p_2, q_2)$ for problem (6) must satisfy the equations*

$$\begin{cases} \Theta \circledast_{\{1,2\}} Q_2 = Q_1 P_1, \\ \Theta \circledast_{\{3,4\}} Q_1 = Q_2 P_2, \end{cases} \tag{9}$$

where the products on the right-hand side of (9) *stand for the polar decompositions of the matrices on the left-hand side, respectively.*

Proof First, we calculate the gradient of g over the product topology $\mathbb{R}^{p_1\times q_1}\times\mathbb{R}^{p_2\times q_2}$ with no constraints. This can be quickly achieved in the operator form via the action of the Fréchet derivative followed by the Riesz representation theorem. To express the gradient in the algebraic form requires some tedius but straightforward algebraic manipulations. Ultimately, the gradient of g can be expressed as

$$\frac{\partial g}{\partial Q_1} = \Theta \circledast_{\{1,2\}} Q_2,$$
$$\frac{\partial g}{\partial Q_2} = \Theta \circledast_{\{3,4\}} Q_1,$$

where $\Theta\circledast_{\{1,2\}}$ and $\Theta\circledast_{\{3,4\}}$ are defined as in (7) and (8).

Next, the tangent space $\mathfrak{T}_Q\mathscr{S}(p,q)$ at $Q\in\mathscr{S}(p,q)$ is composed of matrices $H\in\mathbb{R}^{p\times q}$ of the form

$$H = QK + (\mathrm{I}_p - QQ^\top)W, \tag{10}$$

where $K\in\mathbb{R}^{q\times q}$ is skew-symmetric and $W\in\mathbb{R}^{p\times q}$ is arbitrary. The projection of an arbitrary matrix $Z\in\mathbb{R}^{p\times q}$ onto the tangent space $\mathfrak{T}_Q\mathscr{S}(p,q)$ of $\mathscr{S}(p,q)$ at Q is given by [2]

$$\mathrm{Proj}_{\mathfrak{T}_Q\mathscr{S}(p,q)}Z = Q\frac{Q^\top Z - Z^\top Q}{2} + (\mathrm{I}_p - QQ^\top)Z. \tag{11}$$

Since the constraints of (6) are not coupled, the optimality condition for (6) is that the projections of these partial gradients onto the respective constraints must vanish. Note that the two terms on the right-hand side of (11) are mutually orthogonal. We thus see that the relationships

$$Q_1^\top(\Theta\circledast_{\{1,2\}}Q_2) = (\Theta\circledast_{\{1,2\}}Q_2)^\top Q_1, \tag{12}$$
$$\Theta\circledast_{\{1,2\}}Q_2 = Q_1Q_1^\top(\Theta\circledast_{\{1,2\}}Q_2) \tag{13}$$

must hold for the first partial gradient. Similar conditions hold for the second partial gradient.

Finally, we explain where the polar decomposition comes into play. Denote the entries of Q_2 by $Q_2=[\pi_{ij}]$, where $\pi_{ij}\in\mathbb{R}$, and the blocks of Θ by $\Theta=[\theta_{ij}]$, where $\theta_{ij}\in\mathbb{R}^{p_1\times q_1}$, $i\in[\![p_2]\!]$, and $j\in[\![q_2]\!]$. Then we can write

$$\Theta\circledast_{\{1,2\}}Q_2 = \sum_{i=1}^{p_2}\sum_{j=1}^{q_2}\pi_{ij}\theta_{ij}. \tag{14}$$

Obviously, the matrix $(Q_2 \otimes Q_1)^\top \Omega^\top \Omega (Q_2 \otimes Q_1)$ is symmetric and positive semi-definite. Furthermore, it is easy to check that

$$
\begin{aligned}
&(Q_2 \otimes Q_1)^\top \Omega^\top \Omega (Q_2 \otimes Q_1) \\
&= \begin{bmatrix} Q_1^\top \pi_{11} & Q_1^\top \pi_{21} & \dots & Q_1^\top \pi_{p_2 1} \\ Q_1^\top \pi_{12} & & & \\ \vdots & & & \\ Q_1^\top \pi_{1 q_2} & & \dots & Q_1^\top \pi_{p_2 q_2} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1 q_2} \\ \theta_{21} & \theta_{22} & & \\ \vdots & & & \\ \theta_{p_2 1} & \theta_{p_2 2} & \dots & \theta_{p_2 q_1} \end{bmatrix}.
\end{aligned}
$$

By comparing with the expression (14), we observe that the $q_1 \times q_1$ matrix $Q_1^\top (\Theta \circledast_{\{1,2\}} Q_2)$ occuring in (12) and (13) is precisely the summation of the $q_1 \times q_1$ diagonal blocks along the principal diagonal of $(Q_2 \otimes Q_1)^\top \Omega^\top \Omega (Q_2 \otimes Q_1)$. Thus, not only that the equation (12) is automatically satisfied, but also that the matrix $P_1 := Q_1^\top (\Theta \circledast_{\{1,2\}} Q_2)$ is guaranteed to be positive semi-definite. The right-hand side of (13) is indeed the polar decomposition of $\Theta \circledast_{\{1,2\}} Q_2$. The assertion is therefore proved. □

We shall further explore the structure of (9) in the next section and cast it as a special case of the system(2). This work is about a numerical procedure and the associated convergence analysis for solving the system(9) in particular and the system(2) in general.

Even though the scope of this paper is limited to only a special type of nonlinear matrix equation, solving (9) is an indispensable part for solving the Tucker nearest problem as outlined above. To our knowledge, current updating techniques for the Tucker nearest problem have been limited to one factor matrix at a time. The lack of simultaneous factor matrix updating might be attributable to the nonlinearity such as that involved in(9) which seems too complicated to handle. Our contribution in this paper therefore is innovative. Additionally,the system (2) resembles a nonlinear eigenvalue problem where the role of eigenvalues is being replaced by positive semi-definite matrices. Our numerical procedure resembles the conventional simultaneous subspace iteration where the normalization is being carried out by positive semi-definite matrices. In all, there might be enough mathematics of interest in this study.

2 Basics

The tensor Θ involved in (9) is a function of two unknown matrices Q_1 and Q_2. So the polar decompositions on the right-hand side are implicitly defined. Though it is true that the symmetric and positive semi-definite matrices P_1 and P_2 are uniquely determined from

$$
\begin{cases} P_1 = ((\Theta \circledast_{\{1,2\}} Q_2)^\top (\Theta \circledast_{\{1,2\}} Q_2))^{\frac{1}{2}}, \\ P_2 = ((\Theta \circledast_{\{3,4\}} Q_1)^\top (\Theta \circledast_{\{3,4\}} Q_1))^{\frac{1}{2}}, \end{cases}
$$

any attempt of substituting them into (9) only makes the nonlinearity more tangled. We have to search for some indirect approach.

2.1 Bilinear formulation

The operator-like multiplications defined by (7) and (8) and the way the optimality condition (9) comes into sight are natural from the tensor point of view [10,11], but appear cumbersome to manipulate. The following reformulation sheds better insight into the symmetry innate to the system (9), which will help to motivate a way to solve the polar equations.

Lemma 2 *Suppose* $Q_2 = [\pi_{ij}]$, $\widetilde{Q}_2 = [\widetilde{\pi}_{ij}] \in \mathbb{R}^{p_2 \times q_2}$. *Partition the matrix* $W := \Omega^\top \Omega \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$ *into blocks* $W = [W_{ij}]$ *with* $W_{ij} \in \mathbb{R}^{p_1 \times p_1}$ *and* $i, j \in [\![p_2]\!]$. *Then the multiplication* (7) *can be expressed as*

$$\Theta \circledast_{\{1,2\}} \widetilde{Q}_2 = \mathcal{A}(Q_2, \widetilde{Q}_2) Q_1, \tag{15}$$

where $\mathcal{A}(Q_2, \widetilde{Q}_2)$ *is a matrix in* $\mathbb{R}^{p_1 \times p_1}$ *defined by*

$$\mathcal{A}(Q_2, \widetilde{Q}_2) := \sum_{i=1}^{p_2} \sum_{j=1}^{q_2} \widetilde{\pi}_{ij} \left(\sum_{k=1}^{p_2} W_{ik} \pi_{kj} \right). \tag{16}$$

Proof Let the colon ":" denote an unspecified array of indices. The (i, j)-th block of Θ is a $p_1 \times q_1$ matrix given by

$$\theta_{::ij} = \left(\sum_{k=1}^{p_2} W_{ik} \pi_{kj} \right) Q_1.$$

By Definition (7),

$$\Theta \circledast_{\{1,2\}} \widetilde{Q}_2 = \sum_{i=1}^{p_2} \sum_{j=1}^{q_2} \theta_{::ij} \widetilde{\pi}_{ij}.$$

The relationship (15) follows by factoring Q_1 out of the summation. □

A similar relationship holds for the multiplication(8).

Lemma 3 *Suppose* $Q_1, \widetilde{Q}_1 \in \mathbb{R}^{p_1 \times q_1}$. *Let* $\Omega^\top \Omega = [W_{ij}]$ *be partitioned in the same way as that in Lemma 2. Then*

$$\Theta \circledast_{\{3,4\}} \widetilde{Q}_1 = \mathcal{B}(Q_1, \widetilde{Q}_1) Q_2, \tag{17}$$

where $\mathcal{B}(Q_1, \widetilde{Q}_1) \in \mathbb{R}^{p_2 \times p_2}$ *with entries defined by*

$$\mathcal{B}(Q_1, \widetilde{Q}_1) := \left[\langle \widetilde{Q}_1, W_{ij} Q_1 \rangle \right]. \tag{18}$$

Note that $\mathcal{A}(Q_2, \widetilde{Q}_2) \in \mathbb{R}^{p_1 \times p_1}$ is bilinear in Q_2 and $\widetilde{Q}_2$ and $\mathcal{B}(Q_1, \widetilde{Q}_1) \in \mathbb{R}^{p_2 \times p_2}$ is bilinear in Q_1 and $\widetilde{Q}_1$. In particular, the first optimality condition (9) is now expressed

by finding $Q_1 \in \mathscr{S}(p_1, q_1)$ and $Q_2 \in \mathscr{S}(p_2, q_2)$ such that the system (2) with the specially defined $\mathcal{A}$ and $\mathcal{B}$ is satisfied. We think that (2) is in a much more manageable form than (9). The remaining tasks will be on developing a numerical method for solving (2) and proving its convergence.

2.2 Symmetry and invariance

The operators $\mathcal{A}$ and $\mathcal{B}$ defined by (16) and (18) enjoy two important properties[1]—a sense of symmetry as well as the invariance under the right group action by orthogonal matrices—which we will characterize below. These properties are important tools for our convergence analysis.

Lemma 4 *For any $Q_1, \widetilde{Q}_1 \in \mathbb{R}^{p_1 \times q_1}$ and $Q_2, \widetilde{Q}_2 \in \mathbb{R}^{p_2 \times q_2}$, there is a symmetry within the operators $\mathcal{A}$ and $\mathcal{B}$ in the sense that*

$$\begin{cases} \mathcal{A}(Q_2, \widetilde{Q}_2) = \mathcal{A}(\widetilde{Q}_2, Q_2)^\top, \\ \mathcal{B}(Q_1, \widetilde{Q}_1) = \mathcal{B}(\widetilde{Q}_1, Q_1)^\top. \end{cases} \tag{19}$$

It also holds a transmutation relationship between $\mathcal{A}$ and $\mathcal{B}$ in the sense that

$$\langle \mathcal{B}(Q_1, \widetilde{Q}_1)Q_2, \widetilde{Q}_2 \rangle = \langle \mathcal{A}(Q_2, \widetilde{Q}_2)Q_1, \widetilde{Q}_1 \rangle. \tag{20}$$

Proof By definition, W is symmetric. It follows that $W_{ik}^\top = W_{ki}$ for all $i, k \in [\![p_2]\!]$. The symmetry in (19) is obvious from the Definitions (16)and (18). We also see that

$$\begin{aligned} \langle \mathcal{B}(Q_1, \widetilde{Q}_1)Q_2, \widetilde{Q}_2 \rangle &= \sum_{i=1}^{p_2} \sum_{j=1}^{q_2} \widetilde{\pi}_{ij} \left(\sum_{k=1}^{p_2} \langle \widetilde{Q}_1, W_{ik} Q_1 \rangle \pi_{kj} \right) \\ &= \left\langle \widetilde{Q}_1, \sum_{i=1}^{p_2} \sum_{j=1}^{q_2} \widetilde{\pi}_{ij} \left(\sum_{k=1}^{p_2} W_{ik} \pi_{kj} \right) Q_1 \right\rangle = \langle \widetilde{Q}_1, \mathcal{A}(Q_2, \widetilde{Q}_2)Q_1 \rangle, \end{aligned}$$

which proves the transmutation (20) between $\mathcal{A}$ and $\mathcal{B}$. □

Corollary 1 *If the bilinear maps $\mathcal{A}$ and $\mathcal{B}$ are defined via (16) and (18), respectively, then both $\mathcal{A}(Q_2, Q_2) \in \mathbb{R}^{p_1 \times p_1}$ and $\mathcal{B}(Q_1, Q_1) \in \mathbb{R}^{p_2 \times p_2}$ are symmetric and positive semi-definite matrices.*

Let the special case $\mathscr{S}(q, q)$ of orthogonal matrices be denoted as $\mathscr{O}(q)$. It should be straightforward to see from the definitions of (16) and (18) that the bilinear maps $\mathcal{A}$ and $\mathcal{B}$ are invariant under the right group action by orthogonal matrices in the following sense.

[1] These two characteristics mentioned are innate to our operators $\mathcal{A}$ and $\mathcal{B}$ specifically derived for tensor applications. For general (2), if we assume that the bilinear operators $\mathcal{A}$ and $\mathcal{B}$ satisfy the symmetry (19), (20), and the invariance (21), then the remaining discussion can be equally applied.

Lemma 5 *If $U_1 \in \mathcal{O}(q_1)$ and $U_2 \in \mathcal{O}(q_2)$ are orthogonal, then*

$$\begin{cases} \mathcal{A}(Q_2U_2, \widetilde{Q}_2U_2) = \mathcal{A}(Q_2, \widetilde{Q}_2), \\ \mathcal{B}(Q_1U_1, \widetilde{Q}_1U_1) = \mathcal{B}(Q_1, \widetilde{Q}_1). \end{cases} \tag{21}$$

Lemma 6 *If (Q_1, Q_2) is a solution to (2), then the entire orbit*

$$\mathcal{X}(Q_1, Q_2) := \{(Q_1U_1, Q_2U_2) | U_1 \in \mathcal{O}(q_1), U_2 \in \mathcal{O}(q_2)\} \tag{22}$$

by the right group action is also a solution. The solutions to the system (2) therefore are not isolated.

Proof If (Q_1, Q_2) is a solution, then by (21) we can write

$$\begin{cases} \mathcal{A}(Q_2U_2, Q_2U_2)Q_1U_1 = Q_1U_1(U_1^\top P_1U_1), \\ \mathcal{B}(Q_1U_1, Q_1U_1)Q_2U_2 = Q_2U_2(U_2^\top P_2U_2), \end{cases} \tag{23}$$

implying that the pair (Q_1U_1, Q_2U_2) for any $U_1 \in \mathcal{O}(q_1)$ and $U_2 \in \mathcal{O}(q_2)$ also satisfies (2). □

For our application to problem (6), note that every such a right actioned pair (Q_1U_1, Q_2U_2) results in the same objective value as $g(Q_1, Q_2)$ because $(Q_2U_2) \otimes (Q_1U_1) = (Q_2 \otimes Q_1)(U_2 \otimes U_1)$, whereas $U_2 \otimes U_1$ is itself orthogonal [7].

2.3 Quotient manifold

By Corollary 1, the constrained system (2) is equivalent to the free system (3). The latter is independent of P_1 and P_2. A careful check reveals that the first subsystem in (3) involves only $(p_1 - q_1)q_1$ independent equations while the third subsystem involves $\frac{q_1(q_1+1)}{2}$ equations. Similar counts hold for the second and the fourth subsystems. In total, the system (3) is under-determined and has $\frac{q_1(q_1-1)}{2} + \frac{q_2(q_2-1)}{2}$ degrees of freedom which corresponds precisely to the dimensionality of $\mathcal{O}(q_1) \times \mathcal{O}(q_2)$. Each orbit $\mathcal{X}$ is isomorphic to $\mathcal{O}(q_1) \times \mathcal{O}(q_2)$, but there might be disjoint orbits. An example will be given in Sect. 4.

Given that the system (3) has orbital varieties and every element in the orbit results in the same objective value, it is sufficient to reconsider the system as over the quotient manifold[2]

$$\mathcal{S}(p_1, q_1)/\mathcal{O}(q_1) \times \mathcal{S}(p_2, q_2)/\mathcal{O}(q_2).$$

In this way, we "shrink" an orbit of infinitely many points in $\mathcal{S}(p_1, q_1) \times \mathcal{S}(p_2, q_2)$ to a single point in the quotient manifold over which the polynomial system (3) has the same numbers of equations and unknowns.

[2] Though it is not directly relevant to our discussion, in topology it can be proved that since $\mathcal{O}(q)$ is compact, the quotient space $\mathcal{S}(p, q)/\mathcal{O}(q)$ is Hausdorff. Furthermore, since the right group action is free, the quotient space is indeed a manifold.

Without causing ambiguity, we shall use the same notation $Q_i, i = 1, 2$, to represent interchangeably between the orbit

$$[Q_i] := \{Q_i U | U \in \mathcal{O}(q_i)\} \in \mathcal{S}(p_i, q_i)/\mathcal{O}(q_i)$$

and the element $Q_i \in \mathcal{S}(p_i, q_i)$. In what follows, the calculation is applied to points in $\mathcal{S}(p_1, q_1) \times \mathcal{S}(p_2, q_2)$ as usual, but we shall use the induced metric embedded in the quotient manifold $\mathcal{S}(p_1, q_1)/\mathcal{O}(q_1) \times \mathcal{S}(p_2, q_2)/\mathcal{O}(q_2)$ to argue the convergence.

3 Numerical method

At first glance, the system (2) resembles a nonlinear eigenvalue problem [8], except that in place of eigenvalues are the symmetric and positive semi-definite matrices P_1 and P_2. It is natural to formulate an iterative scheme analogously to the conventional power method for eigenvalue computation [5]. To deal with the nonlinearity, the simplest approach might be to alternate directions in the iteration. We thus propose the scheme[3]

$$\begin{cases} [Q_{1,[s+1]}, P_{1,[s+1]}] = \mathsf{poldec}(\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]}), \\ [Q_{2,[s+1]}, P_{2,[s+1]}] = \mathsf{poldec}(\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s]}), \end{cases} \tag{24}$$

where poldec denotes any algorithm for computing the polar decomposition, such as that in [6]. We remark again that the expression (24) is invariant under the right group action by orthogonal matrices (see (23)), so it can be interpreted as an iteration on orbits. Since there is a one-to-one correspondence of elements among orbits, it suffices to consider the evolution of just one "cross-section" as indicated by the scheme. The notion of cross-sections will be explained in Sect. 3.3 in the context of Riemannian geometry. The primary concern is what the iterative dynamics will lead to. Our goal is to prove that the sequence $\{(Q_{1,[s]}, Q_{2,[s]})\}$ will converge to a limit point in $\mathcal{S}(p_1, q_1) \times \mathcal{S}(p_2, q_2)$ that solves (2).

Recall that if $Z = QP$ is the polar decomposition of $Z \in \mathbb{R}^{p\times q}$, then

$$Q = \underset{U\in\mathcal{S}(p,q)}{\arg\min} \|Z - U\|_F = \underset{U\in\mathcal{S}(p,q)}{\arg\max} \langle Z, U\rangle, \tag{25}$$

whereas

$$\langle Z, Q\rangle = \langle P, I\rangle = \mathrm{trace}(P).$$

Therefore, the definitions of $Q_{1,[s+1]}$ and $Q_{2,[s+1]}$ in (24) enjoy the variational properties of maximizing the traces of $P_{1,[s+1]}$ and $P_{2,[s+1]}$, respectively. These properties will be exploited to prove the convergence. We shall argue the convergence in two aspects. We first prove the convergence of the traces in general, which then will be employed to argue the convergence of the iterates under mild assumptions.

[3] To demonstrate its analogy to the conventional power method and the subspace iteration method, we give two examples in Sect. 4.

3.1 Convergence of traces

To facilitate the argument and to convey the main point, we shall assume that the underlying problem is generic under the following three conditions. First, recall that low rank matrices of a specified size form a nowhere dense, zero measure, closed algebraic variety in the ambient space. For our application, $\Omega = \Upsilon^{(\boldsymbol{\alpha},\boldsymbol{\beta})^\top}$ is an ensemble of the original tensor T and a group of dynamically varying Stiefel matrices $V^{(\beta_1)}, \ldots, V^{(\beta_{k-2})}$. Therefore, it is reasonable to assume that the underlying Ω is of full column rank per given partition $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. It follows from Corollary 1 that both $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]}) \in \mathbb{R}^{p_1 \times p_1}$ and $\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]}) \in \mathbb{R}^{p_2 \times p_2}$ are symmetric and positive semi-definite. Second, observe from the Definitions (16) and (18) that for $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})$ and $\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})$ to be singular, the variables $Q_{1,[s+1]}$ and $Q_{2,[s]}$ must satisfy the special polynomials $\det(\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})) = 0$ and $\det(\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})) = 0$, respectively. Again, the algebraic varieties of these polynomials, parameterized upon Ω, are nowhere dense, zero measure, and closed. Therefore, it is reasonable to assume that generically $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})$ and $\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})$ are nonsingular and, hence, positive definite. Finally, since a countable union of these varieties is still nowhere dense, zero measure, and closed, the assumption that our iterates will do not hit these zero-measure sets should also be generic.

Under the above-mentioned generic assumptions, we may thus write $Q_{1,[s+1]}$ and $Q_{2,[s+1]}$ explicitly as

$$\begin{cases} Q_{1,[s+1]} = \mathcal{A}(Q_{2,[s]}, Q_{2,[s]}) Q_{1,[s]} (Q_{1,[s]}^\top \mathcal{A}^2(Q_{2,[s]}, Q_{2,[s]}) Q_{1,[s]})^{-\frac{1}{2}}, \\ Q_{2,[s+1]} = \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]}) Q_{2,[s]} (Q_{2,[s]}^\top \mathcal{B}^2(Q_{1,[s+1]}, Q_{1,[s+1]}) Q_{2,[s]})^{-\frac{1}{2}}. \end{cases} \tag{26}$$

We first establish a variational relationship in terms of traces.

Lemma 7 *The sequence* $\{(Q_{1,[s]}, Q_{2,[s]})\} \subset \mathscr{S}(p_1, q_1) \times \mathscr{S}(p_2, q_2)$ *generated by the scheme* (24) *satisfies the inequalities*

$$\begin{aligned} &\langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]}) Q_{1,[s]}, Q_{1,[s+1]} \rangle \\ &\quad \le \langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]}) Q_{2,[s]}, Q_{2,[s+1]} \rangle \\ &\quad \le \langle \mathcal{A}(Q_{2,[s+1]}, Q_{2,[s+1]}) Q_{1,[s+1]}, Q_{1,[s+2]} \rangle \le \langle \mathcal{B}(Q_{1,[s+2]}, Q_{1,[s+2]}) Q_{2,[s+1]}, Q_{2,[s+2]} \rangle, \end{aligned}$$

and equivalently

$$trace(P_{1,[s+1]}) \le trace(P_{2,[s+1]}) \le trace(P_{1,[s+2]}) \le trace(P_{2,[s+2]}). \tag{27}$$

Proof We claim first that

$$\begin{aligned} &\langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]}) Q_{1,[s]}, Q_{1,[s+1]} \rangle \\ &\quad \le \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]}) Q_{1,[s+1]}, Q_{1,[s+1]} \rangle, \end{aligned} \tag{28}$$

$$\begin{aligned}&\langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s]}, Q_{2,[s+1]}\rangle \\ &\quad \le \langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s+1]}, Q_{2,[s+1]}\rangle. \end{aligned} \tag{29}$$

Using (28) and the definition of $Q_{2,[s+1]}$, the first inequality in (27) follows from the fact that

$$\begin{aligned}\langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s+1]}, Q_{1,[s+1]}\rangle &= \langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s]}, Q_{2,[s]}\rangle \\ &\le \langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s]}, Q_{2,[s+1]}\rangle.\end{aligned}$$

Likewise, using (29), the second inequality in (27) follows from

$$\begin{aligned}&\langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})Q_{2,[s+1]}, Q_{2,[s+1]}\rangle \\ &\quad = \langle \mathcal{A}(Q_{2,[s+1]}, Q_{2,[s+1]})Q_{1,[s+1]}, Q_{1,[s+1]}\rangle \\ &\quad \le \langle \mathcal{A}(Q_{2,[s+1]}, Q_{2,[s+1]})Q_{1,[s+1]}, Q_{1,[s+2]}\rangle.\end{aligned}$$

The third inequality is just a shift of index from the first inequality.

It only remains to prove the claims (28) and(29). We shall prove only (28), as the argument for (29) is similar. Write

$$\Delta Q_{1,[s]} := Q_{1,[s+1]} - Q_{1,[s]}.$$

Observe that

$$\begin{aligned}&\langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}, Q_{1,[s+1]}\rangle \\ &\quad = \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}, Q_{1,[s]} + \Delta Q_{1,[s]}\rangle \\ &\quad = \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}, Q_{1,[s]}\rangle + \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}, \Delta Q_{1,[s]}\rangle.\end{aligned}$$

The second term is nonnegative because $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})$ is symmetric and positive semi-definite. By using (26), we now write

$$\begin{aligned}&Q_{1,[s]}^\top \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]} \\ &\quad = (Q_{1,[s]}^\top \mathcal{A}^2(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]})^{\frac{1}{2}} - Q_{1,[s]}^\top \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]}.\end{aligned}$$

We claim that

$$(Q_{1,[s]}^\top \mathcal{A}^2(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]})^{\frac{1}{2}} \succeq Q_{1,[s]}^\top \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]} \tag{30}$$

in the sense of the Loewner partial order and thus the matrix

$$Q_{1,[s]}^\top \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}$$

is positive semi-definite. To prove (30), observe that

$$Q_{1,[s]}^\top \mathcal{A}^2(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]} - (Q_{1,[s]}^\top \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]})^2$$

$$= Q_{1,[s]}^{\top}\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})(I - Q_{1,[s]}Q_{1,[s]}^{\top})\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]}.$$

Since $Q_{1,[s]} \in \mathscr{S}(p_1, q_1)$, the matrix $I - Q_{1,[s]}Q_{1,[s]}^{\top}$ is itself symmetric and positive semi-definite. Recall that the square roots of partially ordered symmetric and semi-definite matrices preserve the ordering. We have thus proved all claims. □

Because $\{Q_{1,[s]}\} \subset \mathscr{S}(p_1, q_1)$ and $\{Q_{2,[s]}\} \subset \mathscr{S}(p_2, q_2)$, we see that both sequences $\{\text{trace}(P_{1,[s]})\}$ and $\{\text{trace}(P_{2,[s]})\}$ are bounded. By the monotonicity proved in (27), the sequences $\{\text{trace}(P_{1,[s]})\}$ and $\{\text{trace}(P_{2,[s]})\}$ must converge.

3.2 Diminishing increments

We first prove the diminishing increment between successive iterates. Define

$$h(Q_1, \widetilde{Q}_1; Q_2, \widetilde{Q}_2) := \langle \mathcal{B}(Q_1, \widetilde{Q}_1)Q_2, \widetilde{Q}_2\rangle = \langle \mathcal{A}(Q_2, \widetilde{Q}_2)Q_1, \widetilde{Q}_1\rangle. \tag{31}$$

It might be informative to summarize all the inequalities involved in the above proof as one sequence of telescoping relationships

$$\begin{array}{ll} h(Q_{1,[s]}, Q_{1,[s]}; Q_{2,[s]}, Q_{2,[s]}) & \le h(Q_{1,[s]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s]}) \\ \le h(Q_{1,[s+1]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s]}) & \le h(Q_{1,[s+1]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s+1]}) \\ \le h(Q_{1,[s+1]}, Q_{1,[s+1]}; Q_{2,[s+1]}, Q_{2,[s+1]}) & \le h(Q_{1,[s+1]}, Q_{1,[s+2]}; Q_{2,[s+1]}, Q_{2,[s+1]}). \end{array} \tag{32}$$

Observe that in (32) each inequality corresponds to alternating one variable at a time. In this way, the value of h is being pushed higher per change of the variables. This chain clearly indicates that it forms one complete cycle, ensuring that the telescoping behavior repeats for all s and, hence, the convergence of every h value in the chain.

Theorem 1 *Assume that $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})$ and $\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})$ are positive definite for all but finitely many s. Then $\Delta Q_{1,[s]}$ and $\Delta Q_{2,[s]}$ converge to zero.*

Proof The chain (32) allows us establish the differences

$$\begin{aligned} &h(Q_{1,[s]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s]}) - h(Q_{1,[s]}, Q_{1,[s]}; Q_{2,[s]}, Q_{2,[s]}) \\ &\quad = \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s]}, \Delta Q_{1,[s]}\rangle, \\ &h(Q_{1,[s+1]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s]}) - h(Q_{1,[s]}, Q_{1,[s+1]}; Q_{2,[s]}, Q_{2,[s]}) \\ &\quad = \langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})Q_{1,[s+1]}, \Delta Q_{1,[s]}\rangle. \end{aligned}$$

The interlacing of the chain, together with the convergence of traces, ascertains that these differences converge to zero simultaneously. Taking the difference of the above two equations again, we see that

$$\langle \mathcal{A}(Q_{2,[s]}, Q_{2,[s]})\Delta Q_{1,[s]}, \Delta Q_{1,[s]}\rangle \to 0.$$

A similar argument ensures that

$$\langle \mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})\Delta Q_{2,[s]}, \Delta Q_{2,[s]}\rangle \to 0.$$

Under the assumption that $\mathcal{A}(Q_{2,[s]}, Q_{2,[s]})$ and $\mathcal{B}(Q_{1,[s+1]}, Q_{1,[s+1]})$ are positive definite, the assertion is proved. □

3.3 Convergence of orbits

The Stiefel manifold $\mathscr{S}(p, q)$ is a Riemannian manifold. The group action by $\mathscr{O}(q)$ on $\mathscr{S}(p, q)$ is proper and free. Thus the set $\mathscr{S}(p, q)/\mathscr{O}(q)$ is a differentiable manifold. The quotient map

$$\pi : \mathscr{S}(p, q) \to \mathscr{S}(p, q)/\mathscr{O}(q)$$

is a submersion, that is, its differential

$$d\pi_Q : \mathfrak{T}_Q\mathscr{S}(p, q) \to \mathfrak{T}_{[Q]}(\mathscr{S}(p, q)/\mathscr{O}(q))$$

is surjective for every $Q \in \mathscr{S}(p, q)$. On the other hand, the orbit $[Q]$ is itself a submanifold in $\mathscr{S}(p, q)$ with

$$\mathfrak{T}_Q[Q] := \{QK \in \mathbb{R}^{p\times q} | K \in \mathbb{R}^{q\times q} \text{ is skew-symmetric}\} \subset \mathfrak{T}_Q\mathscr{S}(p, q).$$

The null space of $d\pi_Q$ is precisely $\mathfrak{T}_Q[Q]$. It follows that $d\pi_Q$ induces an isomorphism, denoted by $d\pi_Q^*$,

$$d\pi_Q^* : \mathfrak{T}_Q\mathscr{S}(p, q)/\mathfrak{T}_Q[Q] \to \mathfrak{T}_{[Q]}(\mathscr{S}(p, q)/\mathscr{O}(q)). \tag{33}$$

We thus can identify each element of $\mathfrak{T}_{[Q]}(\mathscr{S}(p, q)/\mathscr{O}(q))$ with an element of $\mathfrak{T}_Q\mathscr{S}(p, q)/\mathfrak{T}_Q[Q]$. Furthermore, because the group action by $\mathscr{S}(q)$ is isometric, we can impose an inner product on $\mathfrak{T}_{[Q]}(\mathscr{S}(p, q)/\mathscr{O}(q))$ in the same way as an inner product on $\mathfrak{T}_Q\mathscr{S}(p, q)/\mathfrak{T}_Q[Q]$ via the inverse map ${d\pi_Q^*}^{-1}$.

We now specify the inner product over the quotient space $\mathfrak{T}_Q\mathscr{S}(p, q)/\mathfrak{T}_Q[Q]$. The notation

$$[\![X]\!] := \{X + V | V \in \mathfrak{T}_Q[Q]\}$$

of an orbit in the quotient space $\mathfrak{T}_Q\mathscr{S}(p, q)/\mathfrak{T}_Q[Q]$ is not to be confused with the orbit $[Q]$ in the quotient manifold $\mathscr{S}(p, q)/\mathscr{O}(q)$. By the decomposition

$$\mathfrak{T}_Q\mathscr{S}(p, q) = \mathfrak{T}_Q[Q] \oplus (\mathfrak{T}_Q[Q])^\perp,$$

we quickly see from (10) that if $Q_\perp$ denotes the matrix in $\mathscr{S}(p, p - q)$ such that the augmented matrix $[Q, Q_\perp]$ is in $\mathscr{O}(p)$, then

$$(\mathfrak{T}_Q[Q])^\perp = \{Q_\perp\Psi | \Psi \in \mathbb{R}^{(p-q)\times q} \text{ is arbitrary}\}.$$

Any element $X \in \mathfrak{T}_Q\mathscr{S}(p, q)$ has a unique decomposition

$$X = QK_X + Q_\perp\Psi_X,$$

where $QK_X \in \mathfrak{T}_Q[Q]$ and $Q_\perp \Psi_X \in (\mathfrak{T}_Q[Q])^\perp$. For any $[\![X]\!], [\![Y]\!] \in \mathfrak{T}_Q\mathscr{S}(p,q)/\mathfrak{T}_Q[Q]$, define

$$\langle [\![X]\!], [\![Y]\!] \rangle := \langle \Psi_X, \Psi_Y \rangle, \tag{34}$$

where, without causing ambiguity, we use the same inner product notation $\langle \cdot, \cdot \rangle$ on both sides.

Lemma 8 *The map $\langle [\![X]\!], [\![Y]\!] \rangle$ defined in (34) is independent of its representatives and is an inner product over the quotient space $\mathfrak{T}_Q\mathscr{S}(p,q)/\mathfrak{T}_Q[Q]$.*

Proof Suppose that $X_1, X_2 \in [\![X]\!]$ and $Y_1, Y_2 \in [\![Y]\!]$. Then

$$\langle [\![X_1]\!], [\![Y_1]\!] \rangle - \langle [\![X_2]\!], [\![Y_2]\!] \rangle = \langle [\![X_1]\!] - [\![X_2]\!], [\![Y_1]\!] \rangle + \langle [\![X_2]\!], [\![Y_1]\!] - [\![Y_2]\!] \rangle.$$

Since $X_1 - X_2 \in \mathfrak{T}_Q[Q]$, it must be that $\Psi_{X_1} = \Psi_{X_2}$. It follows that

$$\langle [\![X_1]\!] - [\![X_2]\!], [\![Y_1]\!] \rangle = \langle \Psi_{X_1} - \Psi_{X_2}, \Psi_{Y_1} \rangle = 0.$$

Likewise, since $\Psi_{Y_1} = \Psi_{Y_2}$, the second term is also zero.

The conjugate symmetry and linearity of(34) are obvious. We also check that if $\langle [\![X]\!], [\![X]\!] \rangle = 0$, then $\Psi_X = 0$, implying that $X \in \mathfrak{T}_Q[Q]$. □

Corollary 2 *The quotient manifold $\mathscr{S}(p,q)/\mathscr{O}(q)$ is a Riemannian manifold equipped with an inner product that is isometric to the inner product in $\mathfrak{T}_Q\mathscr{S}(p,q)/\mathfrak{T}_Q[Q]$ defined by (34). Indeed, its tangent space $\mathfrak{T}_{[Q]}(\mathscr{S}(p,q)/\mathscr{O}(q))$ can be identified with the subspace $(\mathfrak{T}_Q[Q])^\perp$.*

It thus follows that $\mathscr{S}(p,q)/\mathscr{O}(q)$ can be endowed with a metric, known as the geodesics, for measuring the distance between orbits. That is,

$$d([Q],[\tilde{Q}]) = \inf_{[\gamma]} \left\{ \int_0^1 \|[\gamma(t)]'\| dt \right\}, \tag{35}$$

where $[\gamma(t)]$ is a continuously differentiable curve on the manifold $\mathscr{S}(p,q)/\mathscr{O}(q)$ joining $[Q]$ and $[\tilde{Q}]$, $[\gamma(t)]'$ is the corresponding tangent vector in $\mathfrak{T}_{[\gamma(t)]}(\mathscr{S}(p,q)/\mathscr{O}(q))$, and $\|\cdot\|$ is the norm induced by the underlying inner product.

Finally, we are ready to apply the above concept to our algorithm (24).

Theorem 2 *Let $\{(Q_{1,[s]}, Q_{2,[s]})\}$ be the sequence generated by the scheme (24). Let d_i, $i = 1, 2$, be the metrics defined by (35) on $\mathscr{S}(p_i,q_i)/\mathscr{O}(q_i)$, respectively. Then $d_i([Q_{i,[s+1]}], [Q_{i,[s]}])$ converges to zero.*

Proof To measure the distance between $[Q_{i,[s+1]}]$ and $[Q_{i,[s]}]$, it suffices to consider by the isometric isomorphism described above the pullback curve $\gamma_i(t)$ on $\mathscr{S}(p_i,q_i)$ connecting $Q_{i,[s+1]}$ and $Q_{i,[s]}$ with the identities

$$\|[\gamma_i(t)]'\| = \|d\pi^{-1}_{\gamma_i(t)}[\gamma_i(t)]'\| = \|[\![\gamma_i'(t)]\!]\| = \|\gamma_i(t)_\perp \Psi_{\gamma_i'(t)}\| = \|\Psi_{\gamma_i'(t)}\|.$$

On one hand, note that by Lemma 8 the metrics d_i are independent of the representatives. On the other hand, note that by Theorem 1 we have already shown that in one "cross-section", namely, the sequence $\{(Q_{1,[s]}, Q_{2,[s]})\}$ generated by (24), the increments $\Delta Q_{1,[s]}$ and $\Delta Q_{2,[s]}$ converge to zero. It follows by continuity that $d_i([Q_{i,[s+1]}], [Q_{i,[s]}])$ converges to zero for $i = 1, 2$. □

Recall the following lemma from the theory of parameter continuation [9, Theorem 7.1.1] which concerns geometrically isolated solutions to a generic polynomial system.

Lemma 9 *Let $P(\boldsymbol{z}; \boldsymbol{q})$ be a system of n polynomials in variables $\boldsymbol{z} \in \mathbb{C}^n$ and parameters $\boldsymbol{q} \in \mathbb{C}^m$. Let $\mathcal{N}(\boldsymbol{q})$ denote the number of geometrically isolated solutions to $P(\boldsymbol{z}; \boldsymbol{q}) = 0$ over the algebraically closed complex space. Then,*

1. *$\mathcal{N}(\boldsymbol{q})$ is finite, and it is the same, say $\mathcal{N}$, for almost all $\boldsymbol{q} \in \mathbb{C}^m$;*
2. *For all $\boldsymbol{q} \in \mathbb{C}^m$, $\mathcal{N}(\boldsymbol{q}) \leq \mathcal{N}$;*
3. *The subset of $\mathbb{C}^m$ where $\mathcal{N}(\boldsymbol{q}) = \mathcal{N}$ is a Zariski open set. That is, the exceptional subset of $\boldsymbol{q} \in \mathbb{C}^m$, where $\mathcal{N}(\boldsymbol{q}) < \mathcal{N}$, is an affine algebraic set contained within an algebraic set of codimension one.*

Since $\mathbb{R}^n$ (indeed, the closure of any infinite subset) is Zariski dense in $\mathbb{C}^n$, the above statements hold for almost all parameters $\mathbf{q} \in \mathbb{R}^m$, except that the number of real-valued isolated solutions varies as a function of $\mathbf{q}$ and is no longer a constant. For our applications, we only need the fact that the number of real roots of a polynomial system is finite and that they are geometrically isolated for generic $\mathbf{q}$.

Our polynomial system (3) is parameterized by W which, in turns, is related to the given order-k tensor T, but is under-determined. In no way can the solutions be isolated. However, the solutions are structured into orbital varieties. By regarding (3) as a system over the quotient manifold $\mathscr{S}(p_1, q_1)/\mathscr{O}(q_1) \times \mathscr{S}(p_2, q_2)/\mathscr{O}(q_2)$, we have the same number of equations and unknowns (orbits). Lemma 9 can be applied, that is, there is a generic behavior of the solution set, interpreted as orbits, to (3) with respect to the parameter W.

Corollary 3 *For generic $W \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$, there are finitely many and geometrically isolated orbital varieties in the form (22) for the polynomial system (3).*

To complete our proof of convergence of the orbit $\{([Q_{1,[s]}], [Q_{2,[s]}])\}$, we recall the following result first proved in [3, Lemma 4.3].

Lemma 10 *Let $\{a_s\}$ be a bounded sequence of real numbers with the property $|a_{s+1} - a_s| \to 0$ as $s \to \infty$. If the accumulation points for the sequence are isolated, then $\{a_s\}$ converges to a unique limit point.*

Theorem 3 *For generic $W \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$, the sequence of orbits $\{([Q_{1,[s]}], [Q_{2,[s]}])\}$ whose representatives $\{(Q_{1,[s]}, Q_{2,[s]})\}$ are generated by (24) converges to a single limit point.*

Proof The sequence $\{([Q_{1,[s]}], [Q_{2,[s]}])\}$ is clearly bounded, so it must have accumulation points. The accumulation points must satisfy the polynomial system (3). By Corollary 3, they are isolated for generic W. Replacing the role of $\{a_s\}$ in Lemma 10 by

$\{[Q_{i,[s]}]\}$, $i = 1, 2$, together with the fact that the metric $d_i([Q_{i,[s+1]}], [Q_{i,[s]}]) \to 0$ as $s \to \infty$, the sequence $\{[Q_{i,[s]}]\}$ converges to a unique limit point. □

The convergence of the sequence of orbits does not necessarily imply that the sequence of any representatives from the orbits converges. Our sequence $\{Q_{i,[s]}\}$, however, is a cross-section with the identity 1_{q_i} fixed across all orbits.

Corollary 4 *The sequence of iterates* $\{(Q_{1,[s]}, Q_{2,[s]})\}$ *generated by the scheme (24) converges.*

4 Numerical examples

Thus far, the discussion has been on the theoretical side, especially since we have to resort to Riemannian geometry and quotient manifolds to address the convergence of the orbital variety. Now that we have completed the theory for the general case, it might be illuminating to consider a few examples to further demonstrate the intriguing dynamics involved in the general theory discussed above. We present our examples from three perspectives and point out a few open questions along the way.

First, we relate our general scheme to the conventional notion of power iteration and subspace iteration.

Example 1 The iterative scheme corresponding to the case $q_1 = q_2 = 1$ looks very much like the conventional power method

$$\begin{cases} \mathcal{A}(\mathbf{q}_{2,[s]}, \mathbf{q}_{2,[s]})\mathbf{q}_{1,[s]} = \mathbf{q}_{1,[s+1]}\lambda_{1,[s+1]}, \\ \mathcal{B}(\mathbf{q}_{1,[s+1]}, \mathbf{q}_{1,[s+1]})\mathbf{q}_{2,[s]} = \mathbf{q}_{2,[s+1]}\lambda_{2,[s+1]}, \end{cases}$$

where $\lambda_{1,[s]}, \lambda_{2,[s]}$ are simply the norms of the vectors generated by the product on the left-hand side. The main difference is that the matrices $\mathcal{A}(\mathbf{q}_{2,[s]}, \mathbf{q}_{2,[s]})$ and $\mathcal{B}(\mathbf{q}_{1,[s+1]}, \mathbf{q}_{1,[s+1]})$ used to generate the iterates are not stationary. This is a special type nonlinear eigenvalue problem. The group action by $\mathcal{O}(1) = \{1, -1\}$ is simply a change of sign.

Example 2 Suppose that the operators $\mathcal{A} \in \mathbb{R}^{p_1 \times p_1}$ and $\mathcal{B} \in \mathbb{R}^{p_2 \times p_2}$ are invariant in s. The iterative scheme

$$\begin{cases} \mathcal{A}Q_{1,[s]} = Q_{1,[s+1]}P_{1,[s+1]}, \\ \mathcal{B}Q_{2,[s]} = Q_{2,[s+1]}P_{2,[s+1]} \end{cases}$$

is equivalent to the simultaneous subspace iteration. Specifically, after $\mathcal{A}$ and $\mathcal{B}$ are applied to the subspaces spanned by the columns of $Q_{1,[s]}$ and $Q_{2,[s]}$, respectively, the bases of the new subspaces are represented by the orthonormal vectors of $Q_{1,[s+1]}$ and $Q_{2,[s+1]}$. Normally, the new bases would be obtained from the QR decomposition, but our scheme uses the bases from the polar decomposition. It is well known in the literature that the simultaneous q-dimensional subspace iteration will converge to the subspace spanned by the eigenvectors corresponding to the first q dominant eigenvalues. In our case, even though the polar decomposition is used in place of

the QR decomposition, the same convergence result prevails. The group action by orthogonal matrices simply changes one orthonormal basis to another one, but does not change the corresponding subspace.

Second, in Lemma 6, we have defined the orbital variety as a solution to the system (2). In Sect. 2.3, we have cautioned that, although each orbit $\mathscr{X}$ is isomorphic to $\mathscr{O}(q_1) \times \mathscr{O}(q_2)$, there might be multiple disjoint orbits. We now construct an example to demonstrate this point.

Example 3 Consider the case $(Q_1, Q_2) \in \mathscr{S}(3, 2) \times \mathscr{S}(2, 1)$. Let $Z \in \mathbb{R}^{3\times 3}$ be an arbitrary symmetric and positive definite matrix such that its eigenvalues are distinct and that its smallest eigenvalue is sufficiently far away from zero. Define

$$W = \begin{bmatrix} Z & 0_3 \\ 0_3 & \mathrm{I}_3 \end{bmatrix},$$

which is considered as an order-4 tensor in $\mathbb{R}^{3\times 3\times 2\times 2}$. Take $Q_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Then

$$\mathcal{A}(Q_2, Q_2) = Z.$$

Take Q_1 to be any two of the three orthonormal eigenvectors of Z. Then

$$\mathscr{B}(Q_1, Q_1) = \begin{bmatrix} \mathrm{trace}(Q_1^\top Z Q_1) & 0 \\ 0 & 2 \end{bmatrix},$$

which has Q_2 as its eigenvector. It is readily observable that (3) is satisfied with this pair (Q_1, Q_2). There are three ways to choose Q_1. These orbits are disjoint because the eigenvectors of Z are linearly independent.

Third, we experiment with some non-trivial data and demonstrate numerically the limiting behavior that we have described in the theory.

Example 4 Consider the scenario that $k = 5$ and $\prod_{\ell=1}^{3} r_{\beta_\ell} = 70$, whereas we search for $(Q_1, Q_2) \in \mathscr{S}(5, 3) \times \mathscr{S}(8, 4)$ to maximize the functional $g(Q_1, Q_2)$ in (6). We randomly generate a test data $\Omega \in \mathbb{R}^{70\times 40}$ and carry out the iteration in the scheme (24) until the difference between two consecutive iterates is less than 10^{-10}. With this fixed Ω, we repeat this experiment 100 times with randomly generated starting values.

Echoing what we have described in Example 3 where there are multiple orbits, we find in one of the test data that the iterates produce two optimal values, each of which can be reached with significant probability, as is evinced in Fig. 1. Such a phenomenon should not be a surprise because we are dealing with a non-convex optimization problem. What is not clear is how the number of optimal values, which is two in this case but may vary in other cases, depends on the problem data Ω.

In Fig. 2, we record the evolution of the h values defined in (31). The result is from only one run of the iteration, but is typical in all other runs. Our point is that the interlacing property (32), which is essential to our proof of convergence, is manifested

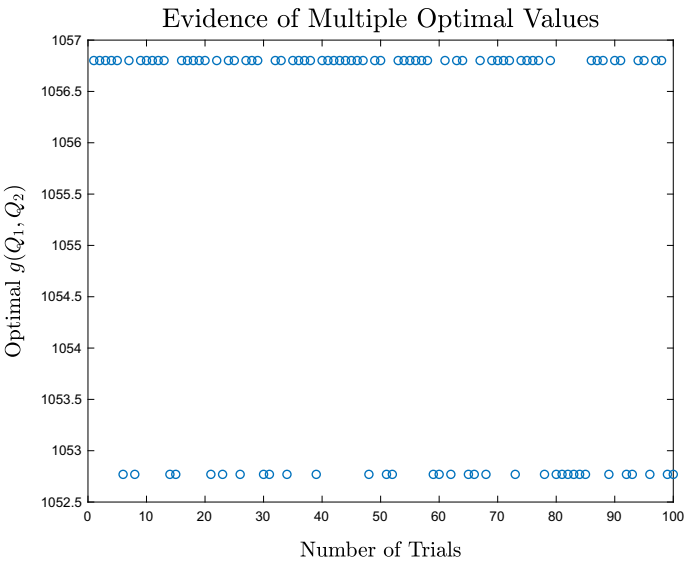

Fig. 1 Multiple optimal values obtained via repeated trials using random starting points

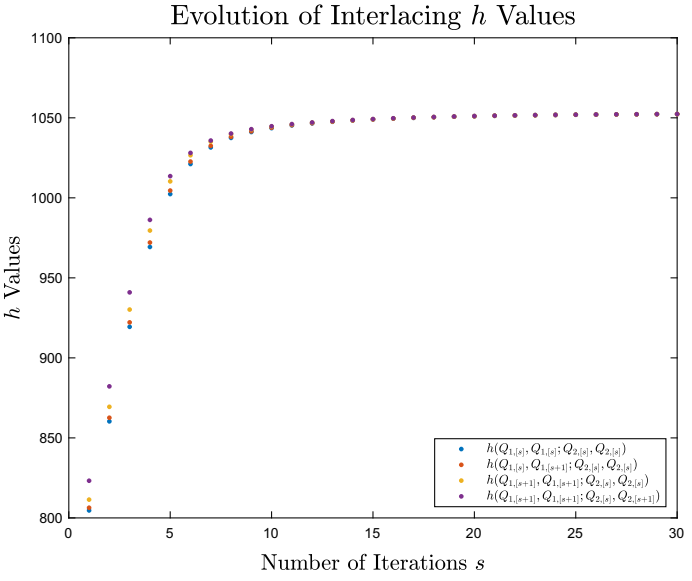

Fig. 2 A typical run showing monotone increasing and interlacing behavior of h values

in the drawing. For clarity, we display only the first 30 iterates. Stacked vertically on top of each other for each s and gradually increased to a common limit point are the first four h values in (32), where the variables are changed one at a time.

Finally, we have repeatedly mentioned that, apart from the nonlinearity induced by the tensor product and the normalization accomplished by the polar decomposition,

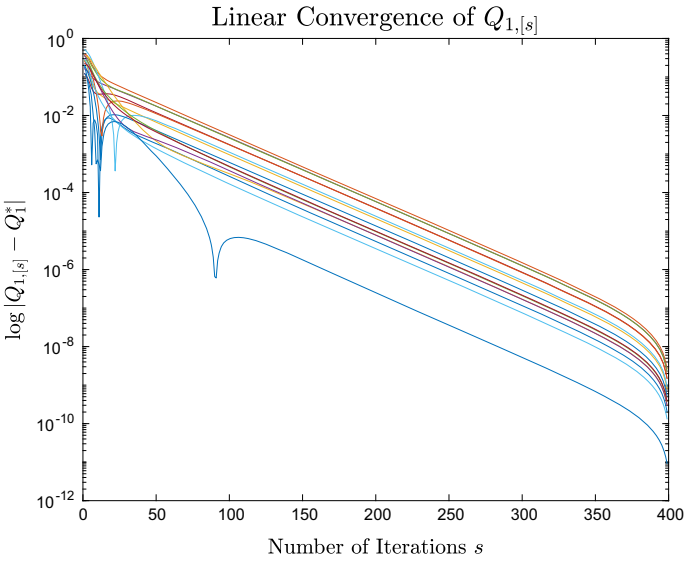

Fig. 3 A typical run showing linear convergence

the iterative scheme (24) resembles the conventional power method and the subspace iteration method. Thus the speed of convergence is expected to be linear at best. Taking empirical result from one typical run, we plot $\log |Q_{1,[s]} - Q_1^*|$ in Fig. 3, where the absolute value is taken entry by entry and Q_1^* denotes the limit point after convergence. The linearity of the logarithmic graph of all 15 entries is an indication that $Q_{1,[s]}$ converges linearly.

5 Conclusion

A basic mechanism needed for the simultaneous update of two factor matrices at a time in the low rank Tucker tensor approximation leads to a coupled nonlinear system of matrix equations. The solution set is invariant under the orthogonal group action and, hence, consists of orbital varieties. An iterative scheme analogous to the conventional power method for subspace iterations is proposed to track one cross-section of the orbits. The convergence analysis is much harder because the matrices used for iteration are themselves part of the dynamics. We resort to the notion of Riemannian manifolds and their quotients as a tool to address the challenges. In particular, we establish an isometric isomorphism (33) between the quotient space $\mathfrak{T}_Q\mathscr{S}(p,q)/\mathfrak{T}_Q[Q]$, which is the tangent space of the Stiefel manifold modulo the tangent space of one orbit, and the tangent space $\mathfrak{T}_{[Q]}(\mathscr{S}(p,q)/\mathscr{O}(q))$ of the Riemannian manifold $\mathscr{S}(p,q)/\mathscr{O}(q)$. Using the induced Riemannian metric, the convergence theory of the orbits as well as of the iterates is completed.

References

1. Bader, B.W., Kolda, T.G.: Efficient MATLAB computations with sparse and factored tensors. SIAM J. Sci. Comput. **30**(1), 205–231 (2008). https://doi.org/10.1137/060676489
2. Chu, M.T., Trendafilov, N.T.: The orthogonally constrained regression revisited. J. Comput. Graph. Stat. **10**(4), 746–771 (2001). https://doi.org/10.1198/106186001317243430
3. Chu, M.T., Trendafilov, J.L.: On a multivariate eigenvalue problem. I. algebraic theory and a power method. SIAM J. Sci. Comput. **14**(5), 1089–1106 (1993). https://doi.org/10.1137/0914066
4. Derksen, H., Kemper, G.: Computational Invariant Theory. Encyclopaedia of Mathematical Sciences Invariant Theory and Algebraic Transformation Groups VIII. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-48422-7
5. Hein, M., Bühler, T.: An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. CoRR arXiv:1012.0774 (2010)
6. Higham, N.J.: Computing the polar decomposition—with applications. SIAM J. Sci. Stat. Comput. **7**(4), 1160–1174 (1986). https://doi.org/10.1137/0907079
7. Langville, A.N., Stewart, W.J.: The Kronecker product and stochastic automata networks. J. Comput. Appl. Math. **167**(2), 429–447 (2004). https://doi.org/10.1016/j.cam.2003.10.010
8. Ruhe, A.: Algorithms for the nonlinear eigenvalue problem. SIAM J. Numer. Anal. **10**, 674–689 (1973). https://doi.org/10.1137/0710059
9. Sommese, A.J., Wampler, I.C.W.: The Numerical Solution of Systems of Polynomials Arising in Engineering and Science. World Scientific Publishing Co. Pte. Ltd., Hackensack (2005). https://doi.org/10.1142/9789812567727
10. Van Loan, C.F.: The ubiquitous Kronecker product. J. Comput. Appl. Math. **123**(1–2), 85–100 (2000). https://doi.org/10.1016/S0377-0427(00)00393-9
11. Van Loan, C.F.: Structured matrix problems from tensors. In: Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications, Lecture Notes in Math., vol. 2173, pp. 1–63. Springer, Cham (2016)