



Trace ratio optimization with an application to multi-view learning

Li Wang¹ · Lei-Hong Zhang² · Ren-Cang Li^{3,4} 

Received: 11 January 2021 / Accepted: 19 September 2022 / Published online: 7 October 2022
© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2022

Abstract

A trace ratio optimization problem over the Stiefel manifold is investigated from the perspectives of both theory and numerical computations. Necessary conditions in the form of nonlinear eigenvalue problem with eigenvector dependency (NEPv) are established and a numerical method based on the self-consistent field (SCF) iteration with a postprocessing step is designed to solve the NEPv and the method is proved to be always convergent. As an application to multi-view subspace learning, a new framework and its instantiated concrete models are proposed and demonstrated on real world data sets. Numerical results show that the efficiency of the proposed numerical methods and effectiveness of the new orthogonal multi-view subspace learning models.

Keywords Trace ratio · Stiefel manifold · Nonlinear eigenvalue problem with eigenvector dependency · NEPv · SCF · Multi-view subspace learning

Mathematics Subject Classification 58C40 · 65F30 · 65H17 · 65K05 · 90C26 · 90C32

Wang is supported in part by NSF DMS-2009689. Zhang is supported in part by the National Natural Science Foundation of China NSFC-12071332. Li is supported in part by NSF DMS-1719620 and DMS-2009689.

✉ Ren-Cang Li
rcli@uta.edu

Li Wang
li.wang@uta.edu

Lei-Hong Zhang
longzlh@suda.edu.cn

¹ Department of Mathematics and Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019-0408, USA

² School of Mathematical Sciences, Soochow University, Suzhou 215006, Jiangsu, China

³ Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019-0408, USA

⁴ Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong

1 Introduction

We are concerned with the following trace ratio maximization problem

$$\max_{X^T X = I_k} f_\theta(X), \quad (1.1a)$$

where $1 \leq k < n$, I_k is the $k \times k$ identity matrix, and

$$f_\theta(X) = \frac{\text{trace}(X^T A X + X^T D)}{[\text{trace}(X^T B X)]^\theta}, \quad (1.1b)$$

$A, B \in \mathbb{R}^{n \times n}$ are symmetric and B is positive semi-definite with $\text{rank}(B) > n - k$, $D \in \mathbb{R}^{n \times k}$, matrix variable $X \in \mathbb{R}^{n \times k}$, and parameter $0 \leq \theta \leq 1$. The condition that $\text{rank}(B) > n - k$ ensures the denominator of $f_\theta(X)$ is always positive for any X such that $X^T X = I_k$.

Problem (1.1) is a maximization problem on the Stiefel manifold [1]:

$$\mathbb{O}^{n \times k} = \{X \in \mathbb{R}^{n \times k} : X^T X = I_k\}.$$

Previously studied special cases include 1) $D = 0$ and $\theta = 1$ from Fisher's linear discriminant analysis (LDA) [31, 46, 47] in the setting of supervised machine learning; 2) $A = 0$ and $\theta = 1/2$ from orthogonal canonical correlation analysis (OCCA) [48]; 3) $B = I_n$ or $\theta = 0$ for which (1.1) is a fundamental problem in numerical linear algebra, optimization, and applied statistics [5, 8, 12–14, 16, 18, 28, 32, 49–51]. For our purpose in Sect. 5, problem (1.1) will appear as a subproblem that has to be solved repeatedly for a novel orthogonal multi-view subspace learning framework.

Our goal is to investigate problem (1.1) as a maximization problem on the Stiefel manifold $\mathbb{O}^{n \times k}$ in both theory and numerical computation. Our major contributions are as follows: (1) We transform the KKT condition of (1.1) with respect to the Stiefel manifold equivalently into a nonlinear eigenvalue problem with eigenvector dependency (NEPv), a term that was coined in [6]; (2) We establish crucial necessary conditions, beyond the KKT condition, of local and global maximizers in terms of the extreme eigenvalues of the NEPv; (3) We characterize the role of D in how precisely it pins maximizers down, which is important because when $D = 0$, any maximizer represents a class of many associated with an element of the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$, the set of all k dimensional subspaces of \mathbb{R}^n ; (4) A numerical method based on the self-consistent field (SCF) iteration for the NEPv with post-processing is proposed to efficiently solve (1.1) as a consequence of our theoretical results, and the method is always convergent; (5) As an application, we establish a new orthogonal multi-view subspace learning framework and solve it alternately with our method for (1.1) serving as the computational workhorse.

The rest of this paper is organized as follows. In Sect. 2, we derive the KKT condition, its associated NEPv, and important theoretical issues to lay the foundation for the rest of the paper. In Sect. 3, we investigate the role of D in pinning down the maximizers. In Sect. 4, we propose our SCF method for problem (1.1) and conduct

a detailed convergence analysis. An application to multi-view subspace learning is carried out in Sect. 5. Results of numerical experiments are reported in Sect. 6. Finally, we draw our conclusions in Sect. 7.

Notation. $\mathbb{R}^{m \times n}$ is the set of $m \times n$ real matrices and $\mathbb{R}^n = \mathbb{R}^{n \times 1}$. $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all ones. $\|\mathbf{x}\|_2$ is the 2-norm of vector $\mathbf{x} \in \mathbb{R}^n$. For $B \in \mathbb{R}^{m \times n}$, $\mathcal{R}(B)$ is the column subspace and its singular values are denoted by $\sigma_i(B)$ for $i = 1, \dots, \min\{m, n\}$ arranged in the nonincreasing order, and

$$\|B\|_2 = \sigma_1(B), \quad \|B\|_F = \sqrt{\sum_{i=1}^{\text{rank}(B)} [\sigma_i(B)]^2}, \quad \|B\|_{\text{trace}} = \sum_{i=1}^{\text{rank}(B)} \sigma_i(B)$$

are the spectral norm, the Frobenius norm, and the trace norm (also known as the nuclear norm) of B , respectively. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, $\text{eig}(A) = \{\lambda_i(A)\}_{i=1}^n$ denotes the set of its eigenvalues (counted by multiplicities) arranged in the nonincreasing order; $A \succ 0$ ($\succeq 0$) means that A is positive definite (semi-definite). MATLAB-like notation is used to access the entries of a matrix: $X_{(i:j,k:\ell)}$ to denote the submatrix of a matrix X , consisting of the intersections of rows i to j and columns k to ℓ , and when $i : j$ is replaced by $:$, it means all rows, similarly for columns.

2 KKT condition and associated NEPv

We start by finding out the first order optimality condition, also known as the KKT condition, for problem (1.1). To that end, we will need to find the gradient of f_θ on Stiefel manifold $\mathbb{O}^{n \times k}$. It is known that the gradient of f_θ on the manifold at X is given by [1, (3.35)], [8, Corollary 1]

$$\text{grad } f_{\theta|\mathbb{O}^{n \times k}}(X) = \Pi_X \left(\frac{\partial f_\theta(X)}{\partial X} \right) = \frac{\partial f_\theta(X)}{\partial X} - X \text{sym} \left(X^T \frac{\partial f_\theta(X)}{\partial X} \right), \quad (2.1)$$

where $\Pi_X(Z) := Z - X \text{sym}(X^T Z)$ and $\text{sym}(X^T Z) = (X^T Z + Z^T X)/2$. With (2.1), computing the gradient is just a matter of computing the partial derivative $\partial f_\theta(X)/\partial X$ for which all entries of X are treated as independent variables. We have

$$\frac{\partial f_\theta(X)}{\partial X} = \frac{2}{[\text{trace}(X^T B X)]^\theta} \left[A X + \frac{D}{2} - \theta f_1(X) B X \right],$$

where $f_1(X)$ is simply $f_\theta(X)$ in (1.1b) with $\theta = 1$. Finally, we obtain the KKT condition $\text{grad } f_{\theta|\mathbb{O}^{n \times k}}(X) = 0$, or equivalently,

$$\frac{2}{[\text{trace}(X^T B X)]^\theta} \left[A X + \frac{D}{2} - \theta f_1(X) B X \right] = X \widehat{\Lambda}, \quad (2.2a)$$

$$X \in \mathbb{O}^{n \times k}, \quad \widehat{\Lambda}^T = \widehat{\Lambda} \in \mathbb{R}^{k \times k}. \quad (2.2b)$$

An explicit expression for $\widehat{\Lambda}$ can be obtained by pre-multiplying Eq. (2.2a) by X^T . Equation (2.2a) does not appear in the form of an NEPv because of the isolated term D . Next, we introduce

$$E(X) = \frac{2}{[\text{trace}(X^T B X)]^\theta} \left[A + \frac{DX^T + XD^T}{2} - \theta f_1(X)B \right] \tag{2.3}$$

and consider the following NEPv

$$E(X)X = X\Lambda, \quad X \in \mathbb{O}^{n \times k}. \tag{2.4}$$

Pre-multiply (2.4) by X^T to get $\Lambda = X^T E(X)X$, which is always symmetric.

Remark 2.1 A KKT condition equivalent to (2.2) can also be obtained by working with $\ln f_\theta(X) = \ln(\text{trace}(X^T A X + X^T D)) - \theta \ln(\text{trace}(X^T B X))$.

Our first theorem establishes an equivalency relation between the KKT condition (2.2) and NEPv (2.4).

Theorem 2.1 $X \in \mathbb{O}^{n \times k}$ is a KKT point of (1.1), i.e., it satisfies (2.2), if and only if it is an orthonormal basis matrix of a k -dimensional invariant subspace of $E(X)$ and $X^T D$ is symmetric.

Proof Suppose that X satisfies (2.2). Pre-multiply (2.2a) by X^T and then solve for $X^T D$ to conclude that $X^T D$ is symmetric. Next, upon using $X^T X = I_k$, we have

$$\begin{aligned} E(X)X &= X\widehat{\Lambda} + \frac{1}{[\text{trace}(X^T B X)]^\theta} X D^T X = X \left(\widehat{\Lambda} + \frac{D^T X}{[\text{trace}(X^T B X)]^\theta} \right) \\ &=: X\Lambda, \end{aligned}$$

which gives (2.4). On the other hand, suppose that (2.4) holds and $X^T D$ is symmetric. We expand (2.4) and rearrange the terms to get

$$\begin{aligned} \text{LHS of (2.2a)} &= -\frac{1}{[\text{trace}(X^T B X)]^\theta} X D^T X + X\Lambda \\ &= X \left(\Lambda - \frac{D^T X}{[\text{trace}(X^T B X)]^\theta} \right) =: X\widehat{\Lambda}, \end{aligned}$$

which gives (2.2a) and also $\widehat{\Lambda}$ is symmetric because both Λ and $D^T X$ are symmetric. □

The following lemma plays a key role in our analysis later in this paper, where and henceforth

$$g_\theta(X) = \frac{\text{trace}(X^T A X)}{[\text{trace}(X^T B X)]^\theta}. \tag{2.5}$$

Lemma 2.1 For $X, \widehat{X} \in \mathbb{O}^{n \times k}$, if

$$\text{trace}(\widehat{X}^T E(X) \widehat{X}) \geq \text{trace}(X^T E(X) X), \quad (2.6)$$

then

$$f_\theta(X) + \gamma \leq g_\theta(\widehat{X}) + \frac{\text{trace}(\widehat{X}^T D X^T \widehat{X})}{[\text{trace}(\widehat{X}^T B \widehat{X})]^\theta}, \quad (2.7)$$

where $\alpha = \text{trace}(X^T A X)$, $\delta = \text{trace}(X^T D)$, $\beta = \text{trace}(X^T B X)$, $\widehat{\beta} = \text{trace}(\widehat{X}^T B \widehat{X})$, and

$$\gamma = \frac{\alpha + \delta}{\widehat{\beta}^\theta \beta} \left[(1 - \theta)\beta + \theta\widehat{\beta} - \beta^{1-\theta}\widehat{\beta}^\theta \right]. \quad (2.8)$$

Furthermore, if inequality (2.6) is strict, then so is inequality (2.7).

Proof It can be verified that

$$\text{trace}(X^T E(X) X) = 2(1 - \theta) f_\theta(X).$$

Let $\widehat{\alpha} = \text{trace}(\widehat{X}^T A \widehat{X})$. By assumption (2.6), we have

$$\begin{aligned} 2(1 - \theta) f_\theta(X) &\leq \text{trace}(\widehat{X}^T E(X) \widehat{X}) \\ &\leq \frac{2}{\beta^\theta} \left[\widehat{\alpha} + \text{trace}(\widehat{X}^T D X^T \widehat{X}) - \theta f_1(X) \widehat{\beta} \right], \\ (1 - \theta) f_\theta(X) \beta^\theta &\leq \widehat{\alpha} + \text{trace}(\widehat{X}^T D X^T \widehat{X}) - \theta f_1(X) \widehat{\beta}, \\ (1 - \theta) f_\theta(X) \frac{\beta^\theta}{\widehat{\beta}^\theta} &\leq \frac{\widehat{\alpha}}{\widehat{\beta}^\theta} + \frac{\text{trace}(\widehat{X}^T D X^T \widehat{X})}{\widehat{\beta}^\theta} - \theta f_\theta(X) \frac{\widehat{\beta}^{1-\theta}}{\beta^{1-\theta}}, \end{aligned}$$

implying

$$g_\theta(\widehat{X}) + \frac{\text{trace}(\widehat{X}^T D X^T \widehat{X})}{\widehat{\beta}^\theta} \geq f_\theta(X) + \gamma,$$

where

$$\begin{aligned} \gamma &= (1 - \theta) f_\theta(X) \frac{\beta^\theta}{\widehat{\beta}^\theta} + \theta f_\theta(X) \frac{\widehat{\beta}^{1-\theta}}{\beta^{1-\theta}} - f_\theta(X) \\ &= (1 - \theta) \frac{\alpha + \delta}{\widehat{\beta}^\theta} + \theta \frac{\alpha + \delta}{\beta} \widehat{\beta}^{1-\theta} - \frac{\alpha + \delta}{\beta^\theta} \\ &= \frac{\alpha + \delta}{\widehat{\beta}^\theta \beta} \left[(1 - \theta)\beta + \theta\widehat{\beta} - \beta^{1-\theta}\widehat{\beta}^\theta \right]. \end{aligned}$$

This proves inequality (2.7), and it is strict if inequality (2.6) is strict. \square

Lemma 2.1 is rather general and valid for all $\theta \in \mathbb{R}$ actually. As the first consequence of Lemma 2.1, we have the next theorem, where $0 \leq \theta \leq 1$ is imposed to ensure γ of (2.8) is nonnegative. It lays the foundation of our SCF iteration for NEPv (2.4) in Sect. 4, which iterates from the current approximation X to the next one \tilde{X} , while the objective value is increased.

Theorem 2.2 *Given $X \in \mathbb{O}^{n \times k}$, suppose either $\theta \in \{0, 1\}$, or $\text{trace}(X^T A X + X^T D) \geq 0$ when $0 < \theta < 1$. If (2.6) holds for $\hat{X} \in \mathbb{O}^{n \times k}$, then*

$$f_\theta(X) \leq g_\theta(\hat{X}) + \frac{\text{trace}(\hat{X}^T D X^T \hat{X})}{[\text{trace}(\hat{X}^T B \hat{X})]^\theta} \tag{2.9}$$

$$\leq g_\theta(\hat{X}) + \frac{\|\hat{X}^T D\|_{\text{trace}}}{[\text{trace}(\hat{X}^T B \hat{X})]^\theta} = f_\theta(\tilde{X}), \tag{2.10}$$

where $\tilde{X} = \hat{X}(UV^T)$ defined in terms of SVD $\hat{X}^T D = U \Sigma V^T$ [15]. Furthermore, if inequality (2.6) is strict, then so is the first inequality in (2.9).

Proof In Lemma 2.1, we note $\gamma \equiv 0$ in the case $\theta \in \{0, 1\}$, and in the case $0 < \theta < 1$ we will have $\gamma \geq 0$ because $\alpha + \delta = \text{trace}(X^T A X + X^T D) \geq 0$ by assumption and¹ $(1 - \theta)\beta + \theta\hat{\beta} - \beta^{1-\theta}\hat{\beta}^\theta \geq 0$. Hence inequality (2.9) holds. To prove the inequality in (2.10), we note, by von Neumann’s trace inequality [42] (see also [17, p. 182], [36, 6.81]), that

$$\text{trace}(\hat{X}^T D X^T \hat{X}) \leq \sum_{i=1}^k \sigma_i(\hat{X}^T D) \sigma_i(X^T \hat{X}) \leq \sum_{i=1}^k \sigma_i(\hat{X}^T D) = \|\hat{X}^T D\|_{\text{trace}},$$

yielding the inequality in (2.10). To see the equality in (2.10), we notice that $\tilde{X}^T D = V U^T \hat{X}^T D = V \Sigma V^T$ yielding $\text{trace}(\tilde{X}^T D) = \|\hat{X}^T D\|_{\text{trace}}$, and that the trace is invariant with respect to similarity transformations. \square

Remark 2.2 In Theorem 2.2, if also $\hat{X}^T D \geq 0$, then $U = V$ and $\tilde{X} = \hat{X}$.

The proof of the inequality in (2.10) above can be adapted to yield the next lemma.

Lemma 2.2 *Given $X \in \mathbb{O}^{n \times k}$, we have*

$$\max_{Q \in \mathbb{O}^{k \times k}} f_\theta(XQ) = g_\theta(X) + \frac{\|X^T D\|_{\text{trace}}}{[\text{trace}(X^T B X)]^\theta},$$

and $Q_{\text{opt}} = UV^T$ is a global maximizer, where $U, V \in \mathbb{O}^{k \times k}$ are from the SVD $X^T D = U \Sigma V$.

¹ This is a classical inequality. A quick proof goes as follows. Suppose $\beta > 0$ (otherwise the inequality clearly holds). Let $x = \hat{\beta}/\beta$. It suffices to show $(1 - \theta) + \theta x \geq x^\theta$ for all $x \geq 0$. Since x^θ is concave for $0 < \theta < 1$, the curve of x^θ as a function of x is at or below its tangent line at $x = 1$ and hence $x^\theta \leq 1 + \theta(x - 1)$, as was to be shown.

Lemma 2.3 ([48, Lemma 3]) *For any $H \in \mathbb{R}^{k \times k}$, we have $|\text{trace}(H)| \leq \sum_{i=1}^k \sigma_i(H)$. If $|\text{trace}(H)| = \sum_{i=1}^k \sigma_i(H)$, then either $H \succeq 0$ when $\text{trace}(H) \geq 0$, or $H \preceq 0$ when $\text{trace}(H) \leq 0$.*

Our next theorem presents necessary conditions for local or global maximizers of (1.1).

Theorem 2.3 *Let $X_{\text{opt}} \in \mathbb{O}^{n \times k}$ be a local or global maximizer of (1.1).*

- (a) *If X_{opt} is a global maximizer, then $X_{\text{opt}}^T D \geq 0$;*
 (b) *If $X_{\text{opt}}^T D \geq 0$, and if also $\text{trace}(X_{\text{opt}}^T A X_{\text{opt}} + X_{\text{opt}}^T D) \geq 0$ in the case when $0 < \theta < 1$, then X_{opt} is an orthonormal basis matrix of the invariant subspace associated with the k largest eigenvalues of $E(X_{\text{opt}})$.*

Proof If $X_{\text{opt}} \in \mathbb{O}^{n \times k}$ is a global maximizer, then by Lemma 2.2

$$f_{\theta}(X_{\text{opt}}) = \max_{Q \in \mathbb{O}^{k \times k}} f_{\theta}(X_{\text{opt}} Q) = g_{\theta}(X_{\text{opt}}) + \frac{\|X_{\text{opt}}^T D\|_{\text{trace}}}{[\text{trace}(X_{\text{opt}}^T B X_{\text{opt}})]^{\theta}},$$

implying $\text{trace}(X_{\text{opt}}^T D) = \|X_{\text{opt}}^T D\|_{\text{trace}}$, which in turn implies $X_{\text{opt}}^T D \geq 0$ by Lemma 2.3. This proves item (a).

Next we prove item (b). Since X_{opt} is a KKT point of problem (1.1), $\mathcal{R}(X_{\text{opt}})$ is an invariant subspace of $E(X_{\text{opt}})$ by Theorem 2.1. Therefore there is an orthogonal matrix Q such that the columns of $X_{\text{opt}} Q \equiv [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k]$ are eigenvectors of $E(X_{\text{opt}})$ associated with its eigenvalues $\mu_1 \geq \mu_2 \geq \dots \geq \mu_k$.

Let $E(X_{\text{opt}}) = U \Lambda U^T$ be the eigen-decomposition of $E(X_{\text{opt}})$, where $U = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$, $U^T U = I_n$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Further, we can choose this eigen-decomposition such that $\mathbf{v}_j = \mathbf{u}_{i_j}$ and $\mu_j = \lambda_{i_j}$ for $1 \leq j \leq k$. It goes as follows: $i_1 = \min\{i : \lambda_i = \mu_1\}$ and recursively, $i_j = \min\{i : \lambda_i = \mu_j, i > i_{j-1}\}$ for $j = 2, \dots, k$. Thus,

$$X_{\text{opt}} = [\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_{k-1}}, \mathbf{u}_{i_k}] Q^T. \quad (2.11)$$

Assume, to the contrary, that $\mathcal{R}(X_{\text{opt}})$ is not an eigenspace associated with the k largest eigenvalues of $E(X_{\text{opt}})$. Then $\mu_k = \lambda_{i_k} < \lambda_k$. Necessarily $i_k > k$. At least one of \mathbf{u}_j for $1 \leq j \leq k$ does not appear among $\mathbf{v}_j = \mathbf{u}_{i_j}$ for $1 \leq j \leq k$ and let \mathbf{u}_{ℓ} be such one. Consider for $0 < \varepsilon < 1$

$$X_{\varepsilon} = [\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_{k-1}}, \sqrt{1 - \varepsilon^2} \mathbf{u}_{i_k} + \varepsilon s \mathbf{u}_{\ell}] Q^T \quad (2.12)$$

which goes to X_{opt} as ε goes to 0, where $s = \pm 1$ such that $s \mathbf{u}_{\ell}^T D \mathbf{q}_k \geq 0$ and \mathbf{q}_k is the last column of Q . It can be verified that $X_{\varepsilon}^T X_{\varepsilon} = I_k$ and

$$\text{trace}(X_{\varepsilon}^T E(X_{\text{opt}}) X_{\varepsilon}) = \sum_{j=1}^k \lambda_{i_j} + \varepsilon^2 (\lambda_{\ell} - \lambda_{i_k})$$

$$\begin{aligned} &\geq \text{trace}(X_{\text{opt}}^T E(X_{\text{opt}}) X_{\text{opt}}) + \varepsilon^2(\lambda_k - \lambda_{i_k}) \\ &> \text{trace}(X_{\text{opt}}^T E(X_{\text{opt}}) X_{\text{opt}}). \end{aligned} \tag{2.13}$$

By Lemma 2.1 and noticing that $\gamma = 0$ for $\theta \in \{0, 1\}$ and $\gamma \geq 0$ for $0 < \theta < 1$ because $\text{trace}(X_{\text{opt}}^T A X_{\text{opt}} + X_{\text{opt}}^T D) \geq 0$ is assumed for the case, we have

$$f_\theta(X_{\text{opt}}) < g_\theta(X_\varepsilon) + \frac{\text{trace}(X_\varepsilon^T D X_{\text{opt}}^T X_\varepsilon)}{[\text{trace}(X_\varepsilon^T B X_\varepsilon)]^\theta}. \tag{2.14}$$

We get from (2.11) and (2.12) that

$$\begin{aligned} X_{\text{opt}}^T X_\varepsilon &= Q[\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_{k-1}}, \mathbf{u}_{i_k}]^T \left[\mathbf{u}_{i_1}, \dots, \mathbf{u}_{i_{k-1}}, \sqrt{1 - \varepsilon^2} \mathbf{u}_{i_k} + \varepsilon s \mathbf{u}_\ell \right] Q^T \\ &= Q \underbrace{\text{diag}(1, 1, \dots, 1, \sqrt{1 - \varepsilon^2})}_{=:\Omega} Q^T, \end{aligned}$$

and

$$\begin{aligned} \text{trace}(X_\varepsilon^T D X_{\text{opt}}^T X_\varepsilon) &= \text{trace}(X_\varepsilon^T D Q \Omega Q^T) = \text{trace}(Q^T X_\varepsilon^T D Q \Omega) \\ &= \sum_{j=1}^{k-1} (Q^T X_\varepsilon^T D Q)_{(j,j)} + \sqrt{1 - \varepsilon^2} (Q^T X_\varepsilon^T D Q)_{(k,k)}, \end{aligned} \tag{2.15}$$

where $(Q^T X_\varepsilon^T D Q)_{(j,j)}$ denotes the (j, j) th entry of $Q^T X_\varepsilon^T D Q$. Next, we note

$$\begin{aligned} X_\varepsilon^T D &= X_{\text{opt}}^T D + Q \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\frac{\varepsilon^2}{1+\sqrt{1-\varepsilon^2}} \mathbf{u}_{i_k}^T D + \varepsilon s \mathbf{u}_\ell^T D \end{bmatrix}, \\ Q^T X_\varepsilon^T D Q &= Q^T X_{\text{opt}}^T D Q + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\frac{\varepsilon^2}{1+\sqrt{1-\varepsilon^2}} \mathbf{u}_{i_k}^T D + \varepsilon s \mathbf{u}_\ell^T D \end{bmatrix} Q \\ &= \begin{bmatrix} \mathbf{u}_{i_1}^T D \\ \vdots \\ \mathbf{u}_{i_{k-1}}^T D \\ \mathbf{u}_{i_k}^T D \end{bmatrix} Q + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\frac{\varepsilon^2}{1+\sqrt{1-\varepsilon^2}} \mathbf{u}_{i_k}^T D + \varepsilon s \mathbf{u}_\ell^T D \end{bmatrix} Q. \end{aligned} \tag{2.16}$$

Recall that $X_{\text{opt}}^T D \geq 0$, and thus $Q^T X_{\text{opt}}^T D Q \geq 0$ and, as a result, its (k, k) th entry $(Q^T X_{\text{opt}}^T D Q)_{(k,k)} = \mathbf{u}_{i_k}^T D \mathbf{q}_k \geq 0$. There are two cases to consider

1. **Case** $(Q^T X_{\text{opt}}^T D Q)_{(k,k)} = \mathbf{u}_{i_k}^T D \mathbf{q}_k > 0$. Then we have

$$\lim_{\varepsilon \rightarrow 0^+} (Q^T X_\varepsilon^T D Q)_{(k,k)} = (Q^T X_{\text{opt}}^T D Q)_{(k,k)} > 0,$$

implying $(Q^T X_\varepsilon^T D Q)_{(k,k)} > 0$ for sufficiently tiny ε ;

2. **Case** $(Q^T X_{\text{opt}}^T D Q)_{(k,k)} = \mathbf{u}_{i_k}^T D \mathbf{q}_k = 0$. It follows from (2.16) that

$$(Q^T X_\varepsilon^T D Q)_{(k,k)} = \varepsilon s \mathbf{u}_\ell^T D \mathbf{q}_k \geq 0$$

by the choice of s we made earlier.

In summary, we always have $(Q^T X_\varepsilon^T D Q)_{(k,k)} \geq 0$ for sufficiently tiny ε . Therefore, for sufficiently tiny $\varepsilon > 0$, we have by (2.15)

$$\begin{aligned} \text{trace}(X_\varepsilon^T D X_{\text{opt}} X_\varepsilon) &\leq \sum_{j=1}^k (Q^T X_\varepsilon^T D Q)_{(j,j)} \\ &= \text{trace}(Q^T X_\varepsilon^T D Q) = \text{trace}(X_\varepsilon^T D). \end{aligned} \quad (2.17)$$

Combine (2.14) and (2.17) to get $f_\theta(X_{\text{opt}}) < f_\theta(X_\varepsilon)$ for sufficiently tiny $\varepsilon > 0$, contradicting that X_{opt} is a local maximizer. \square

Remark 2.3 Our proof for Theorem 2.3(b) is quite laborious, chiefly because we take care of both local and global maximizers with the same argument. Just for the case of a global maximizer alone, it can be significantly simplified. In fact, by Theorem 2.2 and Lemma 2.2, we have from (2.13)

$$f_\theta(X_{\text{opt}}) < g_\theta(X_\varepsilon) + \frac{\|X_\varepsilon^T D\|_{\text{trace}}}{[\text{trace}(X_\varepsilon^T B X_\varepsilon)]^\theta} = \max_{Q \in \mathbb{O}^{k \times k}} f_\theta(X_\varepsilon Q),$$

contradicting that $X_{\text{opt}} \in \mathbb{O}^{n \times k}$ is a global maximizer.

3 The role of D

When $D = 0$, $f_\theta(X Q) \equiv f_\theta(X)$ for any $X \in \mathbb{O}^{n \times k}$ and $Q \in \mathbb{O}^{k \times k}$, as in the LDA case for which $\theta = 1$ as well. In such a case, f_θ is actually a function on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$, the collection of all k -dimensional subspaces in \mathbb{R}^n . Any global maximizer X_{opt} is a representative of a class

$$\mathbb{X}_{\text{opt}} := \{X_{\text{opt}} Q : Q \in \mathbb{O}^{k \times k}\} \quad (3.1)$$

of maximizers. As a result, maximizers are not unique. Fortunately, often any maximizer is just as good as another in applications such as LDA.

In general if $D \neq 0$, then $f_\theta(XQ) \neq f_\theta(X)$. The global maximizers of (1.1) cannot be characterized as simple as we just did for the case $D = 0$. Our goal in this section is to characterize the maximizers of (1.1) for a general D . In particular, our main result imply that if X_{opt} is a global maximizer and if $\text{rank}(X_{\text{opt}}^T D) = k$, then X_{opt} is the unique maximizer within \mathbb{X}_{opt} in (3.1) in the sense that

$$f_\theta(X) < f_\theta(X_{\text{opt}}) \text{ for any } X \in \mathbb{X}_{\text{opt}} \text{ but } X \neq X_{\text{opt}}.$$

To achieve our goal, we will investigate, for a given $\mathcal{X}_* \in \mathcal{G}_k(\mathbb{R}^n)$,

$$\max_{X \in \mathbb{O}^{n \times k}, \mathcal{R}(X) = \mathcal{X}_*} f_\theta(X). \tag{3.2}$$

Lemma 3.1 *Given $\mathcal{X} \in \mathcal{G}_k(\mathbb{R}^n)$, the singular values of $X^T D$ are independent of the choice of $X \in \mathbb{O}^{n \times k}$ subject to $\mathcal{R}(X) = \mathcal{X}$, and as a result, $\text{rank}(X^T D)$ is a constant for any $X \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(X) = \mathcal{X}$.*

Proof Pick a particular $X_0 \in \mathbb{O}^{n \times k}$ such that $\mathcal{R}(X_0) = \mathcal{X}$. Any $X \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(X) = \mathcal{X}$ takes the form $X_0 Q$ for some $Q \in \mathbb{O}^{k \times k}$. The conclusion is a simple consequence of $[(X_0 Q)^T D]^T [(X_0 Q)^T D] = [X_0^T D]^T [X_0^T D]$, which has nothing to do with Q . □

Owing to this lemma, we define the D -rank of $\mathcal{X} \in \mathcal{G}_k(\mathbb{R}^n)$ with respect to $D \in \mathbb{R}^{n \times k}$ by

$$\text{rank}_D(\mathcal{X}) = \text{rank}(X^T D),$$

for any $X \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(X) = \mathcal{X}$. Our main result in this section is Theorem 3.1 below whose proof is deferred to the end of this section after we develop a concrete version of it in Theorem 3.2.

Theorem 3.1 *Given $\mathcal{X}_* \in \mathcal{G}_k(\mathbb{R}^n)$, let $r = \text{rank}_D(\mathcal{X}_*)$. The maximizer X_{opt} of (3.2) admits the decomposition*

$$X_{\text{opt}} = X_{\mathcal{X}_*} + Y_{\mathcal{X}_*}, \tag{3.3}$$

where $X_{\mathcal{X}_*}$ having $\text{rank}(X_{\mathcal{X}_*}) = r$ is unique while $Y_{\mathcal{X}_*}$ with $\text{rank}(Y_{\mathcal{X}_*}) = k - r$ has a freedom of $\mathbb{O}^{(k-r) \times (k-r)}$.

To make Theorem 3.1 concrete, we will explicitly construct $X_{\mathcal{X}_*}$ and $Y_{\mathcal{X}_*}$ in (3.3). To this end, we pick a particular $X_* \in \mathbb{O}^{n \times k}$ such that $\mathcal{R}(X_*) = \mathcal{X}_*$ and keep it fixed. Then any $X \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(X) = \mathcal{X}_*$ takes the form $X = X_* Q$ for some $Q \in \mathbb{O}^{k \times k}$ and vice versa. With this X_* , (3.2) can be equivalently reformulated as

$$\max_{Q \in \mathbb{O}^{k \times k}} f_\theta(X_* Q). \tag{3.4}$$

Lemma 3.2 Let $S \in \mathbb{R}^{k \times k}$ with SVD $S = U \Sigma V^T$, where $U, V \in \mathbb{O}^{k \times k}$ and

$$\Sigma = \text{diag}(\mu_1 I_{k_1}, \dots, \mu_{t-1} I_{k_{t-1}}, \mu_t I_{k_t}) \quad \text{with} \quad (3.5a)$$

$$\mu_1 > \dots > \mu_{t-1} > \mu_t \geq 0, \quad \sum_{i=1}^t k_i = k. \quad (3.5b)$$

Let $r = \text{rank}(S)$, which is k if $\mu_t > 0$ or $k - k_t$ if $\mu_t = 0$. The maximizers of

$$\max_{Q \in \mathbb{O}^{k \times k}} \text{trace}(Q^T S) \quad (3.6)$$

are given by

$$Q_{\text{opt}} = U_{(:,1:r)} V_{(:,1:r)}^T + U_{(:,r+1:k)} W V_{(:,r+1:k)}^T, \quad (3.7)$$

where² $W \in \mathbb{O}^{(k-r) \times (k-r)}$ is arbitrary.

Proof By von Neumann's trace inequality [42],

$$\text{trace}(Q^T S) \leq \sum_{j=1}^k \sigma_j(Q^T) \sigma_j(S) = \sum_{j=1}^k \sigma_j(S) = \|S\|_{\text{trace}} = \|Q^T S\|_{\text{trace}},$$

where for the equality to hold, by Lemma 2.3, we need $Q^T S \geq 0$ and vice versa. Any such Q which we will characterize in a moment is a maximizer of (3.6). Now for any $Q \in \mathbb{O}^{k \times k}$ such that $Q^T S \geq 0$, we have

$$V^T(Q^T S)V = \underbrace{V^T Q^T U}_{=: Z} \Sigma =: Z \Sigma \geq 0.$$

In particular, $Z \Sigma$ is symmetric, i.e., $Z \Sigma = (Z \Sigma)^T = \Sigma Z^T$, from which we get, upon using $Z \in \mathbb{O}^{k \times k}$,

$$\begin{aligned} Z \Sigma Z &= (Z \Sigma) Z = (\Sigma Z^T) Z = \Sigma, \\ Z \Sigma^2 &= (Z \Sigma) \Sigma = (\Sigma Z^T) \Sigma = (\Sigma Z^T)(Z \Sigma Z) = \Sigma^2 Z, \end{aligned}$$

i.e., Σ^2 and Z commute, which implies $Z = \text{diag}(Z_1, \dots, Z_t)$, where $Z_i \in \mathbb{O}^{k_i \times k_i}$. Again use $Z \Sigma = \Sigma Z^T$ to conclude $Z_i^T = Z_i$ for $1 \leq i \leq t-1$ and $Z_t^T = Z_t$ too if $\mu_t > 0$. Furthermore

$$0 \leq Z \Sigma = \text{diag}(\mu_1 Z_1, \dots, \mu_{t-1} Z_{t-1}, \mu_t Z_t)$$

² By convention, when $r = k$, W is a null matrix and the term $U_{(:,r+1:k)} W V_{(:,r+1:k)}^T$ disappears from (3.7) altogether.

yields that $Z_i > 0$ for $1 \leq i \leq t - 1$ and $Z_t > 0$ too if $\mu_t > 0$. Hence $Z_i = I_{k_i}$ for $1 \leq i \leq t - 1$, and $Z_t = I_{k_t}$ too if $\mu_t > 0$ but otherwise $Z_t \in \mathbb{O}^{k_t \times k_t}$ arbitrary. Specifically,

$$Z = \begin{cases} I_k, & \text{if } r = k, \\ \text{diag}(I_r, Z_t), & \text{if } r < k, \end{cases} \tag{3.8}$$

where $Z_t \in \mathbb{O}^{(k-r) \times (k-r)}$ in the case $r < k$ is arbitrary. Finally any maximizer of (3.6) is given by $Q = UZ^T V^T$ with Z as characterized in (3.8). \square

Remark 3.1 The decomposition of Q_{opt} as the sum of two terms in (3.7) is constructed in terms of the SVD of S as specified in the lemma. As they appear, both terms are SVD-dependent! However, they are not. In fact, the first term $U_{(:,1:r)} V_{(:,1:r)}^T$ is the subunitary factor of the polar decomposition of S and the factor is unique [23, 25, 27], independent of any variation in SVD $S = U \Sigma V^T$ so long as $\Sigma_{(1:r,1:r)} > 0$. The second term represents a set of matrices of the form $U_{\perp} W V_{\perp}^T$, where $W \in \mathbb{O}^{(k-r) \times (k-r)}$ is arbitrary, and $U_{\perp}, V_{\perp} \in \mathbb{O}^{k \times (k-r)}$ are any orthonormal basis matrices of the subspaces $\mathcal{R}(S)^{\perp}, \mathcal{R}(S^T)^{\perp}$, respectively.

With the help of Lemma 3.2, we present a concrete version of Theorem 3.1 in Theorem 3.2 below.

Theorem 3.2 Given $X_* \in \mathbb{O}^{n \times k}$, let $X_*^T D = U \Sigma V^T$ be the SVD of $X_*^T D$, where $U, V \in \mathbb{O}^{k \times k}$ and $\Sigma_{(1:r,1:r)} > 0$, where $r = \text{rank}(X_*^T D)$. For any maximizer Q_{opt} of (3.4),

$$X_* Q_{\text{opt}} = X_* U_{(:,1:r)} V_{(:,1:r)}^T + X_* U_{(:,r+1:k)} W V_{(:,r+1:k)}^T, \tag{3.9}$$

for which the first term $X_* U_{(:,1:r)} V_{(:,1:r)}^T$ has rank r and the second term has rank $k - r$ and a freedom in $W \in \mathbb{O}^{(k-r) \times (k-r)}$. Moreover,

$$[X_* Q_{\text{opt}}]^T D = V \Sigma V^T \geq 0, \quad \text{trace}([X_* Q_{\text{opt}}]^T D) = \|X_*^T D\|_{\text{trace}}.$$

Proof By Lemma 2.2, we have

$$\max_{Q \in \mathbb{O}^{k \times k}} f_{\theta}(X_* Q) = g_{\theta}(X_*) + \frac{\text{trace}([X_* Q]^T D)}{[\text{trace}(X_*^T B X_*)]^{\theta}}.$$

Hence the optimizers of (3.4) are the same as those of

$$\max_{Q \in \mathbb{O}^{k \times k}} \text{trace}([X_* Q]^T D) = \max_{Q \in \mathbb{O}^{k \times k}} \text{trace}(Q^T X_*^T D).$$

By Lemma 3.2 with $S = X_*^T D$, Q_{opt} takes the form of (3.7), yielding (3.9). The rest of claims of the theorem are simple consequences. \square

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1 For any particularly chosen $X_* \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(X_*) = \mathcal{X}_*$, adopting the notation of Theorem 3.2, we find $X_{\text{opt}} = X_* Q_{\text{opt}}$ as given by (3.9). We claim that the first term $X_* U_{(:,1:r)} V_{(:,1:r)}^T$ is independent of choices of X_* , although it is constructed by a particularly chosen X_* . First we note $r = \text{rank}(X_*^T D)$ depends on \mathcal{X}_* only by Lemma 3.1. Second, the product $U_{(:,1:r)} V_{(:,1:r)}^T$ does not change with the inherent variations in SVD, as we argued in Remark 3.1. Third, suppose a different $\tilde{X}_* \in \mathbb{O}^{n \times k}$ satisfying $\mathcal{R}(\tilde{X}_*) = \mathcal{X}_*$ is chosen. Then $\tilde{X}_* = X_* \tilde{Q}$ for some $\tilde{Q} \in \mathbb{O}^{k \times k}$. We have

$$\tilde{X}_*^T D = (X_* \tilde{Q})^T D = \tilde{Q}^T X_*^T D = (\tilde{Q}^T U) \Sigma V^T.$$

As we just argued that the product “ $U_{(:,1:r)} V_{(:,1:r)}^T$ ” does not change with the inherent variations in SVD”, we conclude that the first term in (3.9) corresponding to \tilde{X}_* is given by

$$\tilde{X}_* (\tilde{Q}^T U)_{(:,1:r)} V_{(:,1:r)}^T = (X_* \tilde{Q}) (\tilde{Q}^T U_{(:,1:r)}) V_{(:,1:r)}^T = X_* U_{(:,1:r)} V_{(:,1:r)}^T,$$

having nothing to do with \tilde{Q} , as expected. □

The last terms in (3.3) and its concrete version in (3.9) disappear altogether if $r := \text{rank}_D(\mathcal{X}_*) = k$. Hence we have the following corollary.

Corollary 3.1 *Problem (3.2) has a unique maximizer if $\text{rank}_D(\mathcal{X}_*) = k$.*

4 Self-consistent field iteration

In what follows we will limit problem (1.1) to the case:

$$\text{there exists } X \in \mathbb{O}^{n \times k} \text{ such that } \text{trace}(X^T A X + X^T D) \geq 0. \tag{4.1}$$

This assumption ensures that, at optimality, the objective value is nonnegative when $0 < \theta < 1$ but not really needed for our results to hold when $\theta \in \{0, 1\}$, however. It evidently holds if $A \geq 0$, because $\text{trace}(X^T A X + X^T D) \geq 0$ for $X = UV^T$ where U, V are from the SVD of $D = U \Sigma V^T$.

We argue that having (4.1) doesn't lose much generality even for $\theta \in \{0, 1\}$. In fact, it can be verified that for $X \in \mathbb{O}^{n \times k}$

$$f_0(X) = \text{trace}(X^T [A + \alpha I_n] X + X^T D) - k\alpha,$$

$$f_1(X) = \frac{\text{trace}(X^T [A + \alpha B] X + X^T D)}{\text{trace}(X^T B X)} - \alpha.$$

By choosing a sufficiently large $\alpha > 0$, we can make $A + \alpha I_n \geq 0$ and $A + \alpha B \geq 0$ (assuming $B \succ 0$ otherwise it may be possible that $A + \alpha B \not\geq 0$ for any $\alpha > 0$) and hence transform problem (1.1) for $\theta \in \{0, 1\}$ to an equivalent one that satisfies assumption (4.1).

4.1 SCF

Based on the KKT condition in Theorem 2.1, the monotonicity claims in Theorem 2.2 and Lemma 2.2, and the necessary conditions in Theorem 2.3 for a local/global maximizer, an SCF iteration as outlined in Algorithm 4.1 is rather natural.

Algorithm 4.1 SCF iteration for problem (1.1) satisfying (4.1)

Input: $X_0 \in \mathbb{O}^{n \times k}$, such that $\text{trace}(X_0^T A X_0 + X_0^T D) \geq 0$ if $0 < \theta < 1$ but otherwise not required;

Output: a maximizer of (1.1).

- 1: **for** $i = 1, 2, \dots$ until convergence **do**
 - 2: construct $E_i = E(X_{i-1})$ as in (2.3);
 - 3: compute the partial eigen-decomposition $E_i \widehat{X}_i = \widehat{X}_i \Lambda_{i-1}$ for the k largest eigenvalues of E_i and their associated eigenvectors, or simply some $\widehat{X}_i \in \mathbb{O}^{n \times k}$ such that $\text{trace}(\widehat{X}_i^T E_i \widehat{X}_i) > \text{trace}(X_{i-1}^T E_i X_{i-1})$;
 - 4: compute SVD: $\widehat{X}_i^T D = U_i \Sigma_i V_i^T$;
 - 5: $X_i = \widehat{X}_i U_i V_i^T$;
 - 6: **end for**
 - 7: **return** the last X_i as a maximizer of (1.1).
-

A few comments are in order for Algorithm 4.1.

- (1) It is required initially $\text{trace}(X_0^T A X_0 + X_0^T D) \geq 0$ if $0 < \theta < 1$ but not necessary for $\theta \in \{0, 1\}$, in order to ensure that $\{f_\theta(X_i)\}_{i=0}^\infty$ is monotonically increasing (see Theorem 4.2 in the next subsection).

The case when $A \geq 0$ can be easily dealt with as follows: 1) compute SVD $X_0^T D = U \Sigma V^T$, and update X_0 to $X_0(UV^T)$. With the updated X_0 , we have $\text{trace}(X_0^T A X_0 + X_0^T D) = \text{trace}(X_0^T A X_0) + \text{trace}(\Sigma) \geq 0$. This case is ubiquitous in data science applications such as multi-view learning in Sect. 5 that motivates our study in the first place, where A is often some type of variances and hence positive semidefinite. In general, when A is just symmetric and possibly indefinite, what can we do if we don't have such an initial X_0 but (4.1) is known to hold in the case $0 < \theta < 1$? One remedy is to set $\theta = 0$ (or 1) and iterate until some X_i with $\text{trace}(X_i^T A X_i + X_i^T D) \geq 0$ and then switch back to the original θ . But we caution that this remedy could still fail because even with $\theta = 0$ (or 1), the algorithm does not guarantee to find a global maximizer, i.e., there is no guarantee to have some X_i with $\text{trace}(X_i^T A X_i + X_i^T D) \geq 0$. When that happens, we may have to try with one or more random X_0 as the last resort to increase chance of success.

- (2) At line 3, we offer two options to obtain \widehat{X}_i . Evidently, \widehat{X}_i associated with the k largest eigenvalues of E_i maximizes $\text{trace}(\widehat{X}_i^T E_i \widehat{X}_i)$. But as we will show later in Theorem 4.2 that the objective value will still increase as long as $\text{trace}(\widehat{X}_i^T E_i \widehat{X}_i) > \text{trace}(X_{i-1}^T E_i X_{i-1})$. This is an important observation, especially for a large scale problem where an iterative method has to be used to compute the partial eigen-decomposition of E_i and the convergence to a very accurate partial eigen-decomposition may not be cost effective. When that is the case, we can afford to compute partial eigen-decompositions with gradually increased accuracy as the for-loop progresses, namely, use less accurate partial eigen-decompositions

at the beginning many for-loops to save work and more and more accurate partial eigen-decompositions as X_i comes closer and closer to the target. Such an adaptive strategy is a delicate issue and often the best strategy is problem-dependent. Further study on this is out of the scope of this paper and should be pursued elsewhere.

- (3) Lines 4 and 5 execute $\max_Q f(\widehat{X}_i Q)$ yielding X_i according to Theorem 3.2 (with $W = I$ in (3.7) always). X_i is not uniquely defined if $\text{rank}(\widehat{X}_i^T D) < k$. But that non-uniqueness doesn't affect the corresponding objective value.
- (4) A natural stopping criterion to end the for-loop is to use the normalized residual of NEPv (2.4):

$$\frac{[\text{trace}(X_i^T B X_i)]^\theta}{2\sqrt{k}} \cdot \frac{\|E(X_i)X_i - X_i(X_i^T E(X_i)X_i)\|_F}{\|A\|_2 + \theta \|f_1(X_i)\| \|B\|_2 + \|D\|_2} \leq \tau_{01}, \quad (4.2)$$

where τ_{01} is a preset tolerance. For computational convenience, it will be just fine to replace the spectral norms $\|A\|_2$, $\|B\|_2$, and $\|D\|_2$ in (4.2) by their corresponding 1-norm.

Theorem 4.1 below presents two properties about approximation X_i .

Theorem 4.1 *Let $\{X_i\}_{i=0}^\infty$ be generated by Algorithm 4.1.*

- (a) $X_i^T D \succeq 0$ and $\text{trace}(X_i^T D) = \|X_i^T D\|_{\text{trace}}$ for $i \geq 0$.
- (b) *If the eigenvalue gap*

$$\lambda_k(E(X_{i-1})) - \lambda_{k+1}(E(X_{i-1})) > 0,$$

then any two orthonormal eigenbasis matrices \widehat{X}_i and \widehat{Y}_i associated with k largest eigenvalues of $E(X_{i-1})$ satisfy $\widehat{Y}_i = \widehat{X}_i Q$ for some $Q \in \mathbb{O}^{k \times k}$. Furthermore, if, additionally, $\text{rank}(D^T \widehat{X}_i) = k$ (which is independent of Q), then the next approximation X_i from line 5 of Algorithm 4.1 is uniquely determined regardless of any inherent freedom in SVD.

Proof The conclusions in item (a) follows from $X_i^T D = V_i \Sigma_i V_i^T$.

Consider item (b). Since the eigenvalue gap is positive, the eigenspace associated with the k largest eigenvalues of $E(X_{i-1})$ is unique [38, p. 244], and thus the first claim $\widehat{Y}_i = \widehat{X}_i Q$ follows. The second claim is a consequence of Theorem 3.2. \square

4.2 Convergence analysis

Much of our analysis is similar to the one for the OCCA case [48]: $A = 0$ and $\theta = 1/2$. But the complete characterization on what the limits of X_i may look like in the rank-deficient situation, i.e., $\text{rank}(X_*^T D) < k$, in Theorem 4.3 is entirely new even for the OCCA case.

Theorem 4.2 *Let the sequence $\{X_i\}_{i=0}^\infty$ be generated by Algorithm 4.1. The following statements hold.*

- (a) *The sequence $\{f_\theta(X_i)\}_{i=0}^\infty$ is monotonically increasing and convergent;*

(b) If

$$\text{trace}(\widehat{X}_i^T E(X_{i-1})\widehat{X}_i) > \text{trace}(X_{i-1}^T E(X_{i-1})X_{i-1}), \tag{4.3}$$

then $f_\theta(X_{i-1}) < f_\theta(X_i)$;

(c) Let $\{X_i\}_{i \in \mathbb{I}}$ be any convergent subsequence of $\{X_i\}_{i=0}^\infty$, converging to X_* . Then X_* satisfies the first order optimality condition in (2.2) and the necessary condition in Theorem 2.3 for a global maximizer: $X_*^T D \geq 0$ and X_* is an orthonormal basis matrix of the invariant subspace associated with the k largest eigenvalues of $E(X_*)$.

Proof In Algorithm 4.1, we require initially $\text{trace}(X_0^T A X_0 + X_0^T D) \geq 0$ for the case $0 < \theta < 1$ so that all subsequent $\text{trace}(X_i^T A X_i + X_i^T D) \geq 0$ for the case $0 < \theta < 1$. As a result, $\{f_\theta(X_i)\}_{i=0}^\infty$ is monotonically increasing for all $0 \leq \theta \leq 1$. In fact, for $i = 1$, by Theorem 2.2 we conclude that

$$f_\theta(X_0) \leq g_\theta(\widehat{X}_1) + \frac{\|\widehat{X}_1^T D\|_{\text{trace}}}{[\text{trace}(\widehat{X}_1^T B \widehat{X}_1)]^\theta} = f_\theta(X_1).$$

As a consequence, it guarantees $\text{trace}(X_1^T A X_1 + X_1^T D) \geq 0$ when $0 < \theta < 1$. In particular,

$$\text{trace}(\widehat{X}_1^T E(X_0)\widehat{X}_1) > \text{trace}(X_0^T E(X_0)X_0) \implies f_\theta(X_0) < f_\theta(X_1).$$

Inductively, we conclude that $\text{trace}(X_i^T A X_i + X_i^T D) \geq 0$ for all i when $0 < \theta < 1$ and that $\{f_\theta(X_i)\}_{i=0}^\infty$ is monotonically increasing for all $0 \leq \theta \leq 1$, and, furthermore, that if (4.3) holds then $f_\theta(X_{i-1}) < f_\theta(X_i)$. This proves items (a) and (b).

To prove item (c), we consider the subsequence $\{X_{i+1}\}_{i \in \mathbb{I}}$, which, as a bounded sequence in $\mathbb{R}^{n \times k}$, has a convergent subsequence $\{X_{i+1}\}_{i \in \widehat{\mathbb{I}}}$, where $\widehat{\mathbb{I}} \subseteq \mathbb{I}$. Let

$$Z = \lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} X_{i+1} \in \mathbb{O}^{n \times k}.$$

As a result, by Theorem 4.1(a) $Z^T D \geq 0$ which we will need in a moment. According to $E(X_i)X_{i+1} = X_{i+1}(Q_{i+1}^T A_i Q_{i+1})$ for $i \in \widehat{\mathbb{I}}$ where $Q_{i+1} = U_{i+1}V_{i+1}^T$, it holds that

$$E(X_*)Z = ZM, \quad M = Z^T E(X_*)Z. \tag{4.4}$$

Also, from Theorem 4.1(a), we know $Y^T D = D^T Y \geq 0$ for $Y \in \{X_*, Z\}$, and

$$f_\theta(X_*) = \lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} f_\theta(X_i) = \lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} f_\theta(X_{i+1}) = f_\theta(Z). \tag{4.5}$$

Since $E(X_i)X_{i+1} = X_{i+1}(Q_{i+1}^T A_i Q_{i+1})$ and X_{i+1} associates with the k largest eigenvalues of $E(X_i)$, we conclude that Z is an orthonormal eigenbasis matrix of

$E(X_*)$ associated with its k largest eigenvalues. We claim that X_* is also one, too, because, otherwise, we would have

$$\text{trace}(Z^T E(X_*)Z) > \text{trace}(X_*^T E(X_*)X_*).$$

which, by Theorem 2.2 and Remark 2.2 (recall $Z^T D \geq 0$), yields $f_\theta(Z) > f_\theta(X_*)$, contradicting (4.5). Hence X_* is indeed an orthonormal eigenbasis matrix of $E(X_*)$ associated with its k largest eigenvalues, implying

$$E(X_*)X_* = X_*\Lambda_*$$

for some $k \times k$ symmetric Λ_* whose eigenvalues consists of the k largest eigenvalues of $E(X_*)$. Consequently, by Theorem 2.1, X_* satisfies the first order optimality in (2.2). Regarding the necessary conditions in Theorem 2.3 for a global maximizer, we notice that $X_*^T D \geq 0$ is a result of Theorem 4.1(a) and we have already shown that X_* is an orthonormal eigenbasis matrix of $E(X_*)$ associated with its k largest eigenvalues. □

To further analyze the convergence of the sequence $\{X_i\}_{i=0}^\infty$, we now introduce the distance metric on Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$. Let $\mathcal{X} = \mathcal{R}(X)$ and $\mathcal{Y} = \mathcal{R}(Y)$, where $X, Y \in \mathbb{R}^{n \times k}$ with $X^T X = Y^T Y = I_k$. The canonical angles $\theta_1(\mathcal{X}, \mathcal{Y}) \geq \dots \geq \theta_k(\mathcal{X}, \mathcal{Y})$ between \mathcal{X} and \mathcal{Y} are defined by

$$0 \leq \theta_i(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_i(X^T Y) \leq \frac{\pi}{2} \quad \text{for } 1 \leq i \leq k,$$

and accordingly, $\Theta(\mathcal{X}, \mathcal{Y}) = \text{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \dots, \theta_k(\mathcal{X}, \mathcal{Y}))$. It is known that

$$\text{dist}_2(\mathcal{X}, \mathcal{Y}) := \|\sin \Theta(\mathcal{X}, \mathcal{Y})\|_2 \quad (4.6)$$

is a unitarily invariant metric on $\mathcal{G}_k(\mathbb{R}^n)$ [40, p. 95].

The following lemma is an equivalent restatement of [30, Lemma 4.10] (see also [19, Proposition 7]) in the context of Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$.

Lemma 4.1 ([30, Lemma 4.10]) *Let $\mathcal{X}_* \in \mathcal{G}_k(\mathbb{R}^n)$ be an isolated accumulation point of the sequence $\{\mathcal{X}_i \in \mathcal{G}_k(\mathbb{R}^n)\}_{i=0}^\infty$, in the metric (4.6), such that, for every subsequence $\{\mathcal{X}_i\}_{i \in \mathbb{I}}$ converging to \mathcal{X}_* , there is an infinite subset $\widehat{\mathbb{I}} \subseteq \mathbb{I}$ satisfying $\text{dist}_2(\mathcal{X}_i, \mathcal{X}_{i+1}) \rightarrow 0$ as $\mathbb{I} \ni i \rightarrow \infty$. Then the entire sequence $\{\mathcal{X}_i\}_{i=0}^\infty$ converges to \mathcal{X}_* .*

Theorem 4.3 *Let the sequence $\{X_i\}_{i=0}^\infty$ be generated by Algorithm 4.1, and let X_* be an accumulation point of $\{X_i\}_{i=0}^\infty$. Let $0 \leq X_*^T D = V \Sigma V^T$ be the SVD of $X_*^T D$ such that $\Sigma_{(1:r, 1:r)} > 0$, where $r := \text{rank}(X_*^T D)$. Suppose that $\mathcal{R}(X_*)$ is an isolated accumulation point of $\{\mathcal{R}(X_i)\}_{i=0}^\infty$ in the metric (4.6), and that the eigenvalue gap assumption,*

$$\lambda_k(E(X_*)) - \lambda_{k+1}(E(X_*)) > 0, \quad (4.7)$$

holds.

- (a) The entire sequence $\{\mathcal{R}(X_i)\}_{i=0}^\infty$ converges to $\mathcal{R}(X_*)$.
- (b) If $r = k$, then $\{X_i\}_{i=0}^\infty$ converges to X_* (in the standard Euclidean metric).
- (c) In general for $r < k$, $\{X_i\}_{i=0}^\infty$ converges to the set

$$\mathbb{X}_* = \left\{ X_* V_{(:,1:r)} V_{(:,1:r)}^T + X_* V_{(:,r+1:k)} W V_{(:,r+1:k)}^T : W \in \mathbb{O}^{(k-r) \times (k-r)} \right\} \tag{4.8}$$

in the sense that

$$\min_{X \in \mathbb{X}_*} \|X_i - X\|_2 \rightarrow 0 \text{ as } i \rightarrow \infty. \tag{4.9}$$

Proof Suppose that $\{X_i\}_{i \in \mathbb{I}}$ is a subsequence converging to X_* . $\{X_{i+1}\}_{i \in \widehat{\mathbb{I}}}$, as a bounded sequence in $\mathbb{R}^{n \times k}$, has a convergent subsequence $\{X_{i+1}\}_{i \in \widehat{\mathbb{I}}}$, where $\widehat{\mathbb{I}} \subset \mathbb{I}$. Let

$$Z = \widehat{\lim}_{i \rightarrow \infty} X_{i+1} \in \mathbb{O}^{n \times k}.$$

It can be seen that $\{\mathcal{R}(X_i)\}_{i \in \mathbb{I}}$ converges to $\mathcal{R}(X_*)$ and $\{\mathcal{R}(X_{i+1})\}_{i \in \widehat{\mathbb{I}}}$ converges to $\mathcal{R}(Z)$ in the metric (4.6). As in the proof of Theorem 4.2, we will have (4.4) and conclude that both X_* and Z are orthonormal eigenbasis matrices associated with the k largest eigenvalues of $E(X_*)$, and $\mathcal{R}(Z) = \mathcal{R}(X_*)$ by the eigenvalue gap assumption (4.7). Hence

$$\lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} \text{dist}_2(\mathcal{R}(X_i), \mathcal{R}(X_{i+1})) = \text{dist}_2(\mathcal{R}(X_*), \mathcal{R}(Z)) = 0.$$

By Lemma 4.1, $\{\mathcal{R}(X_i)\}_{i=0}^\infty$ converges to $\mathcal{R}(X_*)$. This proves item (a).

With additionally $\text{rank}(X_*^T D) = k$ and the conclusion we just proved, we know that the limit of any convergent subsequence of $\{X_i\}_{i=0}^\infty$ takes the form of $X_* Q$ for some $Q \in \mathbb{O}^{k \times k}$ because all limits share the same column space $\mathcal{R}(X_*)$. Moreover, Theorem 4.2(a) implies that $f_\theta(X_*) = f_\theta(X_* Q)$. Noticing that

$$f_\theta(X_*) = g_\theta(X_*) + \frac{\text{trace}(X_*^T D)}{[\text{trace}(X_*^T B X_*)]^\theta}, \quad f_\theta(X_* Q) = g_\theta(X_*) + \frac{\text{trace}(Q^T X_*^T D)}{[\text{trace}(X_*^T B X_*)]^\theta},$$

we find $\text{trace}(Q^T X_*^T D) = \text{trace}(X_*^T D) = \|X_*^T D\|_{\text{trace}}$ since $X_*^T D > 0$, i.e., this $Q \in \mathbb{O}^{k \times k}$ maximizes $\text{trace}(G^T (X_*^T D))$ over $G \in \mathbb{O}^{k \times k}$, and thus, by Lemma 3.2, Q is the unitary polar factor of $X_*^T D$, yielding $Q = I_k$. This completes the proof of item (b).

We now prove item (c). Let $X_{*\perp} \in \mathbb{O}^{n \times (n-k)}$ such that $[X_*, X_{*\perp}]$ is orthogonal. We expand X_i as

$$X_i = X_* (X_*^T X_i) + X_{*\perp} (X_{*\perp}^T X_i) =: X_* C_i + X_{*\perp} S_i. \tag{4.10}$$

It can be seen that $\|S_i\|_2 = \|\sin(\mathcal{R}(X_i), \mathcal{R}(X_*))\|_2 \rightarrow 0$ as $i \rightarrow \infty$ by item (a). The singular values of C_i are the cosines of the canonical angles between $\mathcal{R}(X_i)$ and

$\mathcal{R}(X_*)$, which all go to 1 as $i \rightarrow \infty$ by item (a). In other words, $C_i C_i^T \rightarrow I_k$ as $i \rightarrow \infty$. So we can write

$$C_i C_i^T = I_k + F_i^{(1)}, \quad \lim_{i \rightarrow \infty} F_i^{(1)} = 0. \quad (4.11)$$

By how X_i are defined in the algorithm and by (4.10), we have

$$0 \leq X_i^T D = C_i^T X_*^T D + S_i^T X_{*\perp}^T D = C_i^T V_{(:,1:r)} \Sigma_{(1:r,1:r)} V_{(:,1:r)}^T + S_i^T X_{*\perp}^T D.$$

Since $S_i^T X_{*\perp}^T D \rightarrow 0$ as $i \rightarrow \infty$, we conclude that

$$\lim_{i \rightarrow \infty} C_i^T V_{(:,1:r)} \Sigma_{(1:r,1:r)} V_{(:,1:r)}^T = \lim_{i \rightarrow \infty} X_i^T D = X_*^T D = V_{(:,1:r)} \Sigma_{(1:r,1:r)} V_{(:,1:r)}^T$$

which implies $\lim_{i \rightarrow \infty} C_i^T V_{(:,1:r)} = V_{(:,1:r)}$. So we can write

$$C_i^T V_{(:,1:r)} = V_{(:,1:r)} + F_i^{(2)}, \quad \lim_{i \rightarrow \infty} F_i^{(2)} = 0. \quad (4.12)$$

As a result, we get $V_{(:,1:r)} = C_i^T V_{(:,1:r)} - F_i^{(2)}$ and

$$\begin{aligned} C_i V_{(:,1:r)} &= C_i C_i^T V_{(:,1:r)} - C_i F_i^{(2)} \\ &= V_{(:,1:r)} + F_i^{(1)} V_{(:,1:r)} - C_i F_i^{(2)} =: V_{(:,1:r)} + F_i^{(3)}, \end{aligned} \quad (4.13)$$

and $\lim_{i \rightarrow \infty} F_i^{(3)} = 0$. Using (4.11) and (4.13), we get

$$[C_i^T V_{(:,r+1:k)}]^T V_{(:,1:r)} = V_{(:,r+1:k)}^T C_i V_{(:,1:r)} = V_{(:,r+1:k)}^T F_i^{(3)}, \quad (4.14)$$

$$\begin{aligned} [C_i^T V_{(:,r+1:k)}]^T [C_i^T V_{(:,r+1:k)}] &= V_{(:,r+1:k)} C_i C_i^T V_{(:,r+1:k)} \\ &= I_k + V_{(:,r+1:k)} F_i^{(1)} V_{(:,r+1:k)}. \end{aligned} \quad (4.15)$$

The distance between two subspaces $\mathcal{R}(C_i^T V_{(:,r+1:k)})$ and $\mathcal{R}(V_{(:,r+1:k)})$ can be given by [24, Lemma 2.1], [39, Chapter 1]

$$\left\| \left\{ [C_i^T V_{(:,r+1:k)}]^T [C_i^T V_{(:,r+1:k)}] \right\}^{-1/2} [C_i^T V_{(:,r+1:k)}]^T V_{(:,1:r)} \right\|_2$$

which goes to 0 as $i \rightarrow \infty$ by (4.14) and (4.15), i.e.,

$$\lim_{i \rightarrow \infty} \mathcal{R}(C_i^T V_{(:,r+1:k)}) = \mathcal{R}(V_{(:,r+1:k)}) \quad (4.16)$$

in the metric (4.6). Equation (4.16) together with (4.15) imply

$$C_i^T V_{(:,r+1:k)} = V_{(:,r+1:k)} W_i^T + F_i^{(4)} \text{ for some } W_i \in \mathbb{O}^{(k-r) \times (k-r)}, \quad (4.17)$$

and $\lim_{i \rightarrow \infty} F_i^{(4)} = 0$. As a result, we get $V_{(:,r+1:k)} = C_i^T V_{(:,r+1:k)} W_i - F_i^{(4)} W_i$ and

$$\begin{aligned} C_i V_{(:,r+1:k)} &= C_i C_i^T V_{(:,r+1:k)} W_i - C_i F_i^{(4)} W_i \\ &= V_{(:,r+1:k)} W_i + F_i^{(1)} V_{(:,r+1:k)} W_i - C_i F_i^{(4)} W_i \\ &=: V_{(:,r+1:k)} W_i + F_i^{(5)}, \end{aligned} \tag{4.18}$$

and $\lim_{i \rightarrow \infty} F_i^{(5)} = 0$. With (4.11)–(4.18) in mind, we have from (4.10) and $V_{(:,1:r)} V_{(:,1:r)}^T + V_{(:,r+1:k)} V_{(:,r+1:k)}^T = I_k$ that

$$\begin{aligned} X_i &= X_* C_i V_{(:,1:r)} V_{(:,1:r)}^T + X_* C_i V_{(:,r+1:k)} V_{(:,r+1:k)}^T + X_{*\perp} S_i \\ &= X_* V_{(:,1:r)} V_{(:,1:r)}^T + X_* V_{(:,r+1:k)} W_i V_{(:,r+1:k)}^T \end{aligned} \tag{4.19a}$$

$$+ X_* F_i^{(3)} V_{(:,1:r)}^T + X_* F_i^{(5)} V_{(:,r+1:k)}^T + X_{*\perp} S_i. \tag{4.19b}$$

The sum of the two terms in (4.19a) belongs to \mathbb{X}_* of (4.8) and each of the three terms in (4.19b) goes to 0 and hence the limiting equation (4.9) holds. \square

What is remarkable about Theorem 4.3 is that we start with an accumulation point X_* which always exists because $\mathbb{O}^{n \times k}$ is a bounded set in $\mathbb{R}^{n \times k}$ and thus is compact, and end up with the conclusions that $\{\mathcal{R}(X_i)\}_{i=0}^\infty$ converges to $\mathcal{R}(X_*)$ and that $\{X_i\}_{i=0}^\infty$ converges to X_* under the conditions that $\text{rank}(X_*^T D) = k$ and $\lambda_k(E(X_*)) > \lambda_{k+1}(E(X_*))$. In general, X_i arbitrarily approaches \mathbb{X}_* of (4.8) as $i \rightarrow \infty$. The set \mathbb{X}_* is uniquely determined by $\mathcal{R}(X_*)$, independent of a particular accumulation point X_* .

A quantitative convergence estimate like [48, ineq. (24)] can be derived, in a similar way there, to obtain

$$\text{dist}_2(\mathcal{R}(X_{i+1}), \mathcal{R}(X_*)) \leq c_0 \|X_i - X_*\|_{\text{trace}},$$

where c_0 is a constant dependent on A, B , and D , and it will be inevitably overestimated to be too big to be of much use, as for [48, ineq. (24)]. For that reason, we will simply skip it. Recently, the authors of [3] proposed an approach to estimate the true SCF convergence rate for an NEPv satisfying the invariance property $E(XQ) \equiv E(X)$ for any $X \in \mathbb{O}^{n \times k}$ and $Q \in \mathbb{O}^{k \times k}$. The approach is not straightforwardly applicable to our NEPv (2.4) because $E(\cdot)$ in (2.3) does not have this invariance property. In what follows, we explain an idea from the forthcoming paper [29] for the case when $X_*^T D > 0$. It is much more complicated to deal with the general case when $X_*^T D \geq 0$ (see [29] for detail). If $X_*^T D > 0$, then $X^T D$ is nonsingular for any $X \in \mathbb{O}^{n \times k}$ such that $\|\sin \Theta(\mathcal{R}(X), \mathcal{R}(X_*))\|_2$ is sufficiently tiny, which is a reasonable assumption for studying local convergence rate. Suppose, in the rest of this paragraph, $\|\sin \Theta(\mathcal{R}(X), \mathcal{R}(X_*))\|_2$ is sufficiently tiny such that $X^T D$ is nonsingular. Then $X^T D$ has a unique polar decomposition [23, 25, 27], and hence

$$\tilde{E}(X) := E(X \Gamma(X^T D)),$$

is well-defined, where $\Gamma(X^T D)$ is the orthogonal polar factor of $X^T D$. It can be seen that the usual SCF, $\tilde{E}(X_{i-1})X_i = X_i \Lambda_i$ on $\tilde{E}(X)$, is same as the SCF in Algorithm 4.1 if starting with the same initial X_0 . It can be proved nontrivially [29] that $\tilde{E}(XQ) \equiv \tilde{E}(X)$ for any $Q \in \mathbb{O}^{k \times k}$ and $X \in \mathbb{O}^{n \times k}$ such that $X^T D$ is nonsingular. Therefore the results in [3] can be applied.

5 Application to multi-view learning

Different from classical machine learning, multi-view learning aims to learn from multiple views of the same object in order to leverage their complementary and redundant information to boost learning performances [4]. For example, in the classification of Internet advertisement on Internet pages [20], the geometry of the image (if available) as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text are considered as different views of a page. Due to the heterogeneity of multiple views, learning from multi-view data is challenging, even though they conceal more information. Multi-view subspace learning is the most popularly studied methodology designed to narrow the heterogeneity gap [35] by learning proper representations of the multiple views in a common latent subspace. In what follows, we will first briefly introduce the problem formulation of multi-view subspace learning and related works, and then propose our new learning model and an efficient alternating iterative method based on our earlier SCF iteration in Algorithm 4.1 to numerically solve the model.

5.1 Problem formulation and related work

Multi-view subspace learning seeks a common latent space via some unknown transformation on each view so that certain learning criteria over the given multi-view data set are optimized with respect to these transformations. Let $\{(z_i^{(1)}, \dots, z_i^{(v)}, y_i)\}_{i=1}^m$ be a multi-view data set of v views and m instances, where the i th data points $z_i^{(s)} \in \mathbb{R}^{n_s}$ of all views ($1 \leq s \leq v$) share the same class label $y_i \in \{0, 1\}^c$ of c classes, whose r th entry $(y_i)_{(r)} = 1$ if the i th data points belong to class r and otherwise $(y_i)_{(r)} = 0$. Linear transformations are often used to perform feature extraction. Specifically, we look for projection matrix $P_s \in \mathbb{R}^{n_s \times k}$ for view s to transform $z_i^{(s)}$ from \mathbb{R}^{n_s} to its latent representation $u_i^{(s)} = P_s^T z_i^{(s)}$ in the common space \mathbb{R}^k .

Represent the m data points of view s by $Z_s = [z_1^{(s)}, \dots, z_m^{(s)}] \in \mathbb{R}^{n_s \times m}$, its latent representation by $U_s = [u_1^{(s)}, \dots, u_m^{(s)}] = P_s^T Z_s \in \mathbb{R}^{k \times m}$, and pack the projection matrices P_s to get

$$P = [P_1^T, P_2^T, \dots, P_v^T]^T \text{ with } P_s \in \mathbb{R}^{n_s \times k} \text{ for } 1 \leq s \leq v. \quad (5.1)$$

We will require $1 \leq k \leq \min_s n_s$ due to the orthogonality constraints to be imposed on each P_s later.

Several existing methods [7, 37, 41] in the literature explored both the inter-view correlations and the intra-view class separability from the labeled multi-view data. Some important statistical quantities are summarized as follows:

1. the sample cross-covariance matrices $C_{s,t} = \frac{1}{m} Z_s H_m Z_t^T$, and, in particular, the covariance matrices $C_{s,s} = \frac{1}{m} Z_s H_m Z_s^T$,
2. the between-class scatter matrices $S_b^{(s)} = Z_s (Y^T \Gamma^{-1} Y - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T) Z_s^T$,
3. the within-class scatter matrices $S_w^{(s)} = Z_s (I_m - Y^T \Gamma^{-1} Y) Z_s^T$,
4. the class centers scatter matrices across views $M_{s,t} = Z_s Y^T \Gamma^{-1} H_c \Gamma^{-1} Y Z_t^T$,

where centering matrix $H_m = I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T$, label matrix $Y = [\mathbf{y}_1, \dots, \mathbf{y}_m]$, and $\Gamma = Y Y^T$.

Most existing methods are often formulated as a trace maximization problem in the form

$$\max_{P^T \Psi P = I_k} \text{trace}(P^T \Phi P) \tag{5.2}$$

which is equivalent to a generalized eigenvalue problem (GEP) for matrix pencil $\Phi - \lambda \Psi$ [10, 15], where $\Phi = \Phi^T$ and $\Psi = \Psi^T > 0$ have the following block structures

$$\Phi = \begin{matrix} & \begin{matrix} n_1 & n_2 & \dots & n_v \end{matrix} \\ \begin{matrix} n_1 \\ n_2 \\ \vdots \\ n_v \end{matrix} & \begin{bmatrix} \Phi_{1,1} & \Phi_{1,2} & \dots & \Phi_{1,v} \\ \Phi_{2,1} & \Phi_{2,2} & \dots & \Phi_{2,v} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{v,1} & \Phi_{v,2} & \dots & \Phi_{v,v} \end{bmatrix} \end{matrix}, \quad \Psi = \begin{matrix} & \begin{matrix} n_1 & n_2 & \dots & n_v \end{matrix} \\ \begin{matrix} n_1 \\ n_2 \\ \vdots \\ n_v \end{matrix} & \begin{bmatrix} \Psi_1 & & & \\ & \Psi_2 & & \\ & & \ddots & \\ & & & \Psi_v \end{bmatrix} \end{matrix}, \tag{5.3}$$

with $\Phi_{s,t}$ and Ψ_s taken to be $C_{s,t}$, $M_{s,t}$, $S_b^{(s)}$, and $S_w^{(s)}$, depending on different learning objectives:

- multiset canonical correlation analysis (MCCA) [43] is (5.2) with $\Phi_{s,t} = C_{s,t}$ and $\Psi_s = C_{s,s}$, $\forall s, t$,
- generalized multi-view analysis (GMA) [37] is (5.2) with $\Phi_{s,t} = \alpha C_{s,t}$, $\forall s \neq t$, $\Phi_{s,s} = S_b^{(s)}$, and $\Psi_s = S_w^{(s)}$, $\forall s$,
- multi-view linear discriminant analysis (MLDA) [41] is (5.2) with $\Phi_{s,t} = \alpha C_{s,t}$, $\forall s \neq t$, $\Phi_{s,s} = S_b^{(s)}$, and $\Psi_s = C_{s,s}$, $\forall s$, and
- multi-view modular discriminant analysis (MvMDA) [7] is (5.2) with $\Phi_{s,t} = M_{s,t}$ and $\Psi_s = S_w^{(s)}$, $\forall s, t$,

where $\alpha \geq 0$ is a pre-defined parameter to weigh the importance of class separability. In these methods, only $\Psi \geq 0$ is guaranteed, and there is a possibility that Ψ may be singular. When that happens, often Ψ is regularized by adding γI with a tiny γ to it. In MCCA, $\Phi \geq 0$ always holds, but it may be indefinite for the other three. In all of these methods, the diagonal blocks of Φ are always positive semi-definite, i.e., all $\Phi_{s,s} \geq 0$, which makes the first point in our previous comments for Algorithm 4.1 a non-issue.

5.2 Proposed model and alternating iteration

We propose a new formulation for supervised multi-view subspace learning as

$$\max_{\{P_s \in \mathbb{O}^{n_s \times k}\}_{s=1}^v} \phi_\theta(P) \quad \text{with} \quad \phi_\theta(P) := \frac{\text{trace}(P^T \Phi P)}{[\text{trace}(P^T \Psi P)]^\theta}, \quad (5.4)$$

an *orthogonal multi-view subspace learning model* (OMvSL), where Φ and Ψ generally will have the block structures as in (5.3) with each block taken to be $C_{s,t}$, $M_{s,t}$, $S_b^{(s)}$, or $S_w^{(s)}$, depending on learning scenarios as the above existing methods, and $0 \leq \theta \leq 1$ is an adjustable parameter that can be fine tuned to yield the best contrastive effect between $\text{trace}(P^T \Phi P)$ and $\text{trace}(P^T \Psi P)$. It is noted that $\theta = 0$ means no contrastive comparison at all, $\theta = 1$ means the other extreme as in LDA, while $0 < \theta < 1$ means comparing $\text{trace}(P^T \Phi P)$ against $\text{trace}(P^T \Psi P)$ fractionally. Function $\phi_\theta(P)$ is well-defined if at least one of the inequalities $\text{rank}(\Psi_s) > n_s - k$ for $1 \leq s \leq v$ is valid.

Comparing with (5.2), the new model (5.4) possesses two unique properties:

1. Linear projection matrices P_s are orthonormal for $s = 1, \dots, v$. This is a preferred property for metric preservation and data visualization, and has been explored for unsupervised learning in, e.g., [9, 48]. However, orthogonality constraints are incompatible with constraint in (5.2) if both are imposed. A workaround in the past is to compute a solution P to (5.2) and orthogonalize each corresponding portion of P to generate a projection matrices for the view, but it may produce suboptimal performance [9].
2. Trace ratio formulation (5.4) is a more essential formulation for general feature extraction problem than ratio trace formulation (5.2) [44] since it naturally solves the above-mentioned incompatibility issue. The introduced θ , as a super-parameter, can adjust the relative importance of $\text{trace}(P^T \Phi P)$ against that of $\text{trace}(P^T \Psi P)$. In our later numerical experiments, we will investigate the impact of θ in terms of classification accuracy.

Model (5.4) is a maximization problem over the Cartesian product of v Stiefel manifolds $\mathbb{O}^{n_s \times k}$. The KKT conditions can be derived straightforwardly by examining the partial gradients with respect to each P_s on $\mathbb{O}^{n_s \times k}$ along the line of derivations in Sect. 2. In fact, for any fixed s and fixed $P_{s'}$ for $s' \neq s$, the objective of (5.4) becomes a function of P_s alone:

$$\chi_{s;\theta}(P_s) := \frac{\text{trace}(P_s^T A_s P_s) + \text{trace}(P_s^T D_s)}{[\text{trace}(P_s^T B_s P_s)]^\theta}, \quad (5.5a)$$

where, with $\alpha_s = \text{trace}(P_{[s]}^T \Phi_{[s]} P_{[s]})$ and $\beta_s = \sum_{s' \neq s} \text{trace}(P_{s'}^T \Psi_{s'} P_{s'})$,

$$A_s = \Phi_{ss} + (\alpha_s/k)I_{n_s}, \quad B_s = \Psi_s + (\beta_s/k)I_{n_s}, \quad D_s = 2 \sum_{s' \neq s} \Phi_{s,s'} P_{s'}, \quad (5.5b)$$

$\Phi_{[s]}$ is Φ after crossing out its s th block-row and s th block-column, and $P_{[s]}$ is P of (5.1) after crossing out its s th block. The dependency of A_s , B_s , and D_s on $P_{s'}$ for $s' \neq s$ is suppressed for clarity. The KKT conditions of (5.4) can be made to consist of v coupled NEPv:

$$E_s(P_s) P_s = P_s A_s, \quad P_s \in \mathbb{O}^{n_s \times k} \quad \text{for } 1 \leq s \leq v, \tag{5.6a}$$

where

$$E_s(P_s) = \frac{2}{[\text{trace}(P_s^T B_s P_s)]^\theta} \left[A_s + \frac{D_s P_s^T + P_s D_s^T}{2} - \theta \chi_{s;1}(P_s) B_s \right]. \tag{5.6b}$$

They are coupled because of the dependency of A_s , B_s , and D_s on $P_{s'}$ for $s' \neq s$. Individually, (5.6) is the KKT condition for

$$\max_{P_s \in \mathbb{O}^{n_s \times k}} \chi_{s;\theta}(P_s), \quad \text{given } P_{s'} \text{ for } s' \neq s. \tag{5.7}$$

Along the line of reasoning in Sect. 2, we can get the next theorem, as an extension of Theorem 2.3.

Theorem 5.1 *Let $\{P_s^{\text{opt}} \in \mathbb{O}^{n_s \times k}\}_{s=1}^v$ be a local or global maximizer of (5.4) and let A_s , B_s , and D_s in (5.6) be evaluated at $\{P_s^{\text{opt}}\}_{s=1}^v$.*

- (a) *If $\{P_s^{\text{opt}}\}_{s=1}^v$ is a global maximizer, then $(P_s^{\text{opt}})^T D_s \geq 0$ for $1 \leq s \leq v$;*
- (b) *Suppose $\phi_\theta(\{P_s^{\text{opt}}\}_{s=1}^v) \geq 0$ when $0 < \theta < 1$ but otherwise not required for $\theta \in \{0, 1\}$. If $(P_s^{\text{opt}})^T D_s \geq 0$, then P_s^{opt} is an orthonormal basis matrix of the invariant subspace associated with the k largest eigenvalues of $E_s(P_s^{\text{opt}})$.*

Proof If $\{P_s^{\text{opt}}\}_{s=1}^v$ is a global maximizer, then P_s^{opt} for a fixed s is a global maximizer of (5.7). Item (a) is a consequence of Theorem 2.3(a). Notice that $\phi_\theta(\{P_s^{\text{opt}}\}_{s=1}^v) \geq 0$ implies $\text{trace}([P_s^{\text{opt}}]^T A_s P_s^{\text{opt}}) + \text{trace}([P_s^{\text{opt}}]^T D_s) \geq 0$. Apply Theorem 2.3(b) to (5.7) to conclude the proof of item (b). □

5.3 Alternating iteration

Similar to Sect. 4, in what follows we will limit problem (5.4) to the case:

$$\text{there exists } \{P_s \in \mathbb{O}^{n_s \times k}\}_{s=1}^v \text{ such that } \text{trace}(P^T \Phi P) \geq 0. \tag{5.8}$$

This assumption is automatically satisfied if all $\Phi_{s,s} \geq 0$, as in all existing multi-view learning methods reviewed in Sect. 5.1, because

$$\text{trace}(P^T \Phi P) = \sum_{s=1}^v \text{trace}(P_s^T \Phi_{s,s} P_s) \geq 0.$$

Algorithm 5.1 OMvSL θ : Orthogonal Multi-view Subspace Learning via θ -Trace Ratio**Input:** Φ and Ψ as in (5.3), $1 \leq k \leq \min_s n_s$, and tolerance ϵ ;**Output:** $\{P_s \in \mathbb{O}^{n_s \times k}\}_{s=1}^v$ that approximately solves (5.4).

- 1: pick $\{P_s^{(0)} \in \mathbb{O}^{n_s \times k}\}_{s=1}^v$ satisfying $\text{trace}([P^{(0)}]^T \Phi P^{(0)}) \geq 0$ if $0 < \theta < 1$ but otherwise not required for $\theta \in \{0, 1\}$, where $P^{(0)} = [(P_1^{(0)})^T, \dots, (P_v^{(0)})^T]^T$;
- 2: $i = 0$, and evaluate the objective of (5.4) at $\{P_s^{(0)}\}_{s=1}^v$ to ϕ ;
- 3: **repeat**
- 4: **for** $s = 1$ to v **do**
- 5: form (5.7) with either $P_{s'} = P_{s'}^{(i)}$, $\forall s' \neq s$ for the Jacobi-style updating, or $P_{s'} = P_{s'}^{(i+1)}$, $1 \leq s' < s$ and $P_{s'} = P_{s'}^{(i)}$, $s < s' \leq v$ for the Gauss-Seidel-style updating;
- 6: solve (5.7) by Algorithm 4.1 (with $P_s^{(i)}$ as an initial guess) for its maximizer $P_s^{(i+1)}$;
- 7: **end for**
- 8: $\phi_0 = \phi$, and evaluate the objective of (5.4) at $\{P_s^{(i+1)}\}_{s=1}^v$ to ϕ ;
- 9: $i = i + 1$;
- 10: **until** $|\phi - \phi_0| \leq \epsilon \phi$;
- 11: **return** the last $\{P_s^{(i)} \in \mathbb{O}^{n_s \times k}\}_{s=1}^v$.

Again assumption (5.8) is not really needed for our results to hold in the case when $\theta \in \{0, 1\}$. Generic optimization methods for optimizing a smooth function over the Cartesian product of the Stiefel manifolds $\mathbb{O}^{n_s \times k}$ are available and can be applied. For example, classical optimization algorithms such as the steepest ascent or trust-region methods over the Euclidean space have been extended to the general Riemannian manifolds in, e.g., [1]. But these methods do not make use of the special trace-fractional structure. In what follows, we propose to solve (5.4) by maximizing its objective alternatingly over $\{P_s \in \mathbb{O}^{n_s \times k}\}_{s=1}^v$ in either the Jacobi-style or Gauss–Seidel-style updating as outlined in Algorithm 5.1, where the SCF iteration in Algorithm 4.1 serves as the computational engine to solve each subproblem (5.7) over just one P_s at line 6.

Algorithm 5.1 requires initially $\text{trace}([P^{(0)}]^T A P^{(0)}) \geq 0$ if $0 < \theta < 1$ but otherwise not required for $\theta \in \{0, 1\}$, similarly to what we previously remarked for Algorithm 4.1. Note that $\text{trace}([P^{(0)}]^T A P^{(0)}) \geq 0$ is guaranteed to hold if all $\Phi_{s,s} \geq 0$. The condition guarantees that the objective (5.4) is monotonically increasing for the Gauss–Seidel-style updating. In cases when we don't have an initial guess $P^{(0)}$ satisfying $\text{trace}([P^{(0)}]^T \Phi P^{(0)}) \geq 0$ for the case $0 < \theta < 1$, we suggest to set $\theta = 0$ (or 1) and iterate until some $P^{(i)}$ such that $\text{trace}([P^{(i)}]^T \Phi P^{(i)}) \geq 0$ and then switch back to the original θ . Unfortunately, it is not clear if the monotonicity property in the objective holds for the Jacobi-style updating even with $\text{trace}([P^{(0)}]^T \Phi P^{(0)}) \geq 0$. In all of our numerical experiments in Sect. 6.2, we simply take $P_s^{(0)}$ to be the first k columns of I_{n_s} for reproducibility and didn't encounter any convergence issue nonetheless for both the Jacobi-style and Gauss-Seidel-style updating. In practice, we may simply take random $\{P_s^{(0)}\}_{s=1}^v$ if no better one is known.

Remark 5.1 Later in Sect. 6, Algorithm 5.1 will be applied with the blocks of Φ and Ψ realized as in the multi-view learning methods GMA [37], MLDA [41], and MvMDA [7]. The resulting (5.4) will be referred to as OGMA, OMLDA, and OMvMDA, respectively.

Next we will discuss the convergence of Algorithm 5.1. With the Jacobi-style updating, Algorithm 5.1 generates a sequence $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ and the same can be said for with the Gauss–Seidel-style updating. But for the convenience of convergence analysis, we shall expand the sequence by inserting $v - 1$ additional intermediate approximations

$$\{P_1^{(i+1)}, \dots, P_s^{(i+1)}, P_{s+1}^{(i)}, \dots, P_v^{(i)}\}, s = 1, 2, \dots, v - 1$$

into between $\{P_s^{(i)}\}_{s=1}^v$ and $\{P_s^{(i+1)}\}_{s=1}^v$ in the case of the Gauss–Seidel-style updating. We then re-index the expanded sequence and still denote it by $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$.

Theorem 5.2 *Let the sequence $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ be generated by Algorithm 5.1, and let $\{P_s^{(*)}\}_{s=1}^v$ be an accumulation point of $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$. Evaluate $A_s, B_s,$ and D_s in (5.6) at $\{P_{s'}^{(*)}, s' \neq s\}$ for each s to $A_s^{(*)}, B_s^{(*)},$ and $D_s^{(*)}$, respectively.*

- (a) $(P_s^{(*)})^T D_s^{(*)} \geq 0$ for $1 \leq s \leq v$.
- (b) $\{\phi_\theta(\{P_s^{(i)}\}_{s=1}^v)\}_{i=0}^\infty$ is monotonically increasing in the case of the Gauss-Seidel-style updating and thus convergent.

Proof Item (a) holds because $(P_s^{(i)})^T D_s \geq 0$ is designed to hold in Algorithm 4.1. Because of our expansion in the sequence of approximations in notation for the Gauss–Seidel-style updating, $\{P_s^{(i)}\}_{s=1}^v$ differs from $\{P_s^{(i+1)}\}_{s=1}^v$ in just one of the $P_s^{(i)}$, and that particular $P_s^{(i)}$ is updated by Algorithm 4.1 so that the objective value is increased. Hence item (b) holds. □

We caution that the monotonicity in objective value is only proved for the case of the Gauss–Seidel-style updating but not the Jacobi-style updating. As a tradeoff in the for-loop of Algorithm 5.1, the v subproblems (5.7) for the Jacobi-style updating are completely independent and can be solved in parallel, while those for the Gauss–Seidel-style updating have to be solved sequentially.

We introduce a metric for the Cartesian product of v Grassmann manifolds:

$$\text{dist}_2(\{P_s\}_{s=1}^v, \{Q_s\}_{s=1}^v) = \sum_{s=1}^v \|\sin \Theta(P_s, Q_s)\|_2 \tag{5.9}$$

for $(P_1, \dots, P_v), (Q_1, \dots, Q_v) \in \mathcal{G}_k(\mathbb{R}^{n_1}) \times \dots \times \mathcal{G}_k(\mathbb{R}^{n_v})$. Again the following lemma, similar to Lemma 4.1, is an equivalent restatement of [30, Lemma 4.10] (see also [19, Proposition 7]) in the context of the Cartesian product of Grassmann manifolds $\mathcal{G}_k(\mathbb{R}^{n_s})$.

Lemma 5.1 ([30, Lemma 4.10]) *Let $\{P_s^{(*)}\} \in \mathcal{G}_k(\mathbb{R}^{n_s})\}_{s=1}^v$ be an isolated accumulation point of the sequence $\{\{P_s^{(i)} \in \mathcal{G}_k(\mathbb{R}^{n_s})\}_{s=1}^v\}_{i=0}^\infty$, in the metric (5.9), such that, for every subsequence $\{\{P_s^{(i)}\}_{s=1}^v\}_{i \in \mathbb{I}}$ converging to $\{P_s^{(*)}\}_{s=1}^v$, there is an infinite subset $\widehat{\mathbb{I}} \subseteq \mathbb{I}$ satisfying $\text{dist}_2(\{P_s^{(i)}\}_{s=1}^v, \{P_s^{(i+1)}\}_{s=1}^v) \rightarrow 0$ as $\widehat{\mathbb{I}} \ni i \rightarrow \infty$. Then the entire sequence $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ converges to $\{P_s^{(*)}\}_{s=1}^v$.*

Theorem 5.3 *To the conditions of Theorem 5.2 add these: $\{\mathcal{R}(P_s^{(*)})\}_{s=1}^v$ is an isolated accumulation point of $\{\{\mathcal{R}(P_s^{(i)})\}_{s=1}^v\}_{i=0}^\infty$ in the metric (5.9) and the eigenvalue gaps*

$$\lambda_k(E_s^{(*)}(P_s^{(*)})) - \lambda_{k+1}(E_s^{(*)}(P_s^{(*)})) > 0 \text{ for } 1 \leq s \leq v,$$

where each $E_s^{(*)}(P_s^{(*)})$ is defined as in (5.6b) with $A_s^{(*)}$, $B_s^{(*)}$, and $D_s^{(*)}$, and also³ $\{\phi_\theta(\{P_s^{(i)}\}_{s=1}^v)\}_{i=0}^\infty$ is convergent for the Jacobi-style updating. Let $r_s := \text{rank}((P_s^{(*)})^T D_s^{(*)})$ for $1 \leq s \leq v$.

- (a) *The entire sequence $\{\{\mathcal{R}(P_s^{(i)})\}_{s=1}^v\}_{i=0}^\infty$ converges to $\{\mathcal{R}(P_s^{(*)})\}_{s=1}^v$.*
 (b) *If $r_s = k$ for all $1 \leq s \leq k$, then $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ converges to $\{P_s^{(*)}\}_{s=1}^v$ (in the metric of the Cartesian product of the Euclidean spaces $\mathbb{R}^{n_s \times k}$).*
 (c) *In general, for each s , let $[P_s^{(*)}]^T D_s^{(*)} = V_s \Sigma_s V_s^T$ be the singular value decomposition such that $(\Sigma_s)_{(1:r_s, 1:r_s)} > 0$, and define*

$$\mathbb{P}_s^{(*)} = \left\{ P_s^{(*)} (V_s)_{(:, 1:r_s)} (V_s)_{(:, 1:r_s)}^T + P_s^{(*)} (V_s)_{(:, r_s+1:k)} W_s (V_s)_{(:, r_s+1:k)}^T : W_s \in \mathbb{O}^{(k-r_s) \times (k-r_s)} \right\}.$$

Then $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ converges to the product $\mathbb{P}_1^{(*)} \times \cdots \times \mathbb{P}_v^{(*)}$ of sets, in the sense that

$$\min_{P_s \in \mathbb{P}_s^{(*)} \forall s} \sum_s \|P_s^{(i)} - P_s\|_2 \rightarrow 0 \text{ as } i \rightarrow \infty.$$

Proof Suppose that $\{\{P_s^{(i)}\}_{s=1}^v\}_{i \in \mathbb{I}}$ is a subsequence converging to $\{P_s^{(*)}\}_{s=1}^v$. Note that $\{\{P_s^{(i+1)}\}_{s=1}^v\}_{i \in \mathbb{I}}$, as a bounded sequence in $\mathbb{R}^{n_1 \times k} \times \cdots \times \mathbb{R}^{n_v \times k}$, has a convergent subsequence $\{\{P_s^{(i+1)}\}_{s=1}^v\}_{i \in \widehat{\mathbb{I}}}$, where $\widehat{\mathbb{I}} \subset \mathbb{I}$. Let

$$Z_s = \lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} P_s^{(i+1)} \in \mathbb{O}^{n_s \times k} \text{ for } 1 \leq s \leq v.$$

It can be seen that $\{\{\mathcal{R}(P_s^{(i)})\}_{s=1}^v\}_{i \in \mathbb{I}}$ converges to $\{\mathcal{R}(P_s^{(*)})\}_{s=1}^v$ and $\{\{\mathcal{R}(P_s^{(i+1)})\}_{s=1}^v\}_{i \in \widehat{\mathbb{I}}}$ converges to $\{\mathcal{R}(Z_s)\}_{s=1}^v$ in the metric (5.9).

For each s , we have $E_s(P_s^{(i)})P_s^{(i+1)} = P_s^{(i+1)}[(Q_s^{(i+1)})^T \Lambda_s^{(i)} Q_s^{(i+1)}]$ or possibly $P_s^{(i+1)} = P_s^{(i)}$ in the case of the Gauss–Seidel-style updating. Now letting $\widehat{\mathbb{I}} \ni i \rightarrow \infty$, we get $E_s(P_s^{(*)})Z_s = Z_s M_s$ or possibly $P_s^{(*)} = Z_s$. For the latter, we obviously have $\mathcal{R}(P_s^{(*)}) = \mathcal{R}(Z_s)$, and, for the former, as we argued in the proof of Theorem 4.2, we will also have $\mathcal{R}(P_s^{(*)}) = \mathcal{R}(Z_s)$. Hence for each s

$$\lim_{\widehat{\mathbb{I}} \ni i \rightarrow \infty} \text{dist}_2(\mathcal{R}(P_s^{(i)}), \mathcal{R}(P_s^{(i+1)})) = \text{dist}_2(\mathcal{R}(P_s^{(*)}), \mathcal{R}(Z_s)) = 0.$$

³ $\{\phi_\theta(\{P_s^{(i)}\}_{s=1}^v)\}_{i=0}^\infty$ is guaranteed convergent for the Gauss–Seidel-style updating by Theorem 5.2(b).

By Lemma 5.1, the entire sequence $\{\{\mathcal{R}(P_s^{(i)})\}_{s=1}^v\}_{i=0}^\infty$ converges to $\{\mathcal{R}(P_s^{(*)})\}_{s=1}^v$. This proves item (a).

With, additionally, $\text{rank}((P_s^{(*)})^T D_s^{(*)}) = k$ and the conclusion we just proved, we know that the limit of any convergent subsequence of $\{\{P_s^{(i)}\}_{s=1}^v\}_{i=0}^\infty$ takes the form of $\{P_s^{(*)} Q_s\}_{s=1}^v$ for some $Q_s \in \mathbb{O}^{k \times k}$ because all limits share the same column spaces, respectively. Moreover, since $\{\phi_\theta(\{P_s^{(i)}\}_{s=1}^v)\}_{i=0}^\infty$ is convergent (by assumption for the Jacobi-style updating or guaranteed for the Gauss–Seidel-style updating by Theorem 5.2(b)), we must have

$$\phi_\theta(\{P_s^{(*)}\}_{s=1}^v) = \phi_\theta(\{P_s^{(*)} Q_s\}_{s=1}^v).$$

It follows from (5.5) that

$$\text{trace}(Q_s^T [P_s^{(*)}]^T D_s^{(*)}) = \text{trace}([P_s^{(*)}]^T D_s^{(*)}) = \|[P_s^{(*)}]^T D_s^{(*)}\|_{\text{trace}}.$$

Hence $Q_s \in \mathbb{O}^{k \times k}$ maximizes $\text{trace}(G^T [P_s^{(*)}]^T D_s^{(*)})$ over $G \in \mathbb{O}^{k \times k}$ and therefore Q_s is the unitary polar factor of $[P_s^{(*)}]^T D_s^{(*)}$, yielding $Q_s = I_k$. This completes the proof of item (b). A proof of item (c) can be given in a similar way to that of Theorem 4.3(c). Detail is omitted. \square

6 Numerical experiments

In this section, we will perform two sets of numerical experiments. The first set demonstrates the basic behavior of the SCF iteration in Algorithm 4.1 for problem (1.1) on synthetic examples, and the second set demonstrates the effectiveness of our multi-view subspace learning model (5.4) solved by the alternating iteration in Algorithm 5.1 which uses Algorithm 4.1 as its computational workhorse. We compare ours against the state-of-the-art methods in machine learning for multi-view feature extraction on five real world data sets. All experiments were conducted in MATLAB 2018a on an Mac laptop using macOS Mojave with Intel Core i9 CPU (2.9 GHz) and 32 GB memory.

6.1 Experiments on synthetic problems

We first report numerical results on problem (1.1) solved by our proposed SCF iteration in Algorithm 4.1 on synthetic examples, where matrices A , B and D are randomly generated with varying $n \in [1000, 4000]$ and $k \in \{50, 100\}$. Specifically, for a given pair (n, k) , matrices A and D are synthesized in MATLAB as

```
X = randn(n, n); X = (X+X') ./ 2; [U, ~] = eig(X);
v = rand(n, 1) + 1e-6; A = U * diag(v) * U'; D = randn(n, k);
```

and B is generated similarly to A . With an increase of n by 1000 in the given interval, we generated 8 synthetic examples in total. Also varying $\theta \in [0, 1]$ with an increase

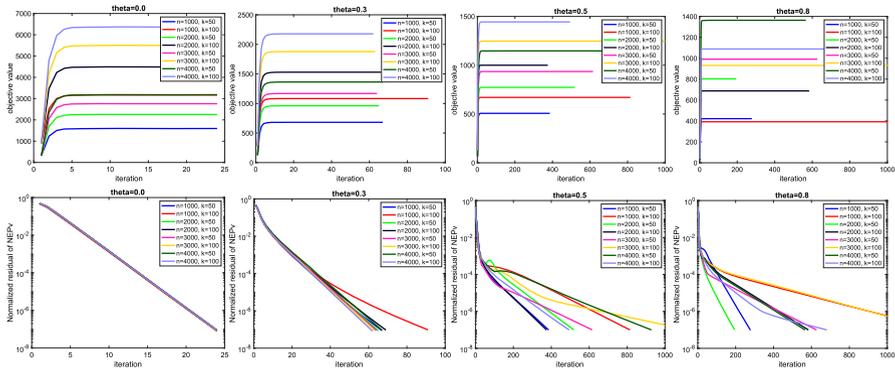


Fig. 1 Convergence curves of objective value and of normalized NEPV residual, defined as the left-hand side of (4.2), by Algorithm 4.1 on 8 synthetic examples for $\theta \in \{0, 0.3, 0.5, 0.8\}$

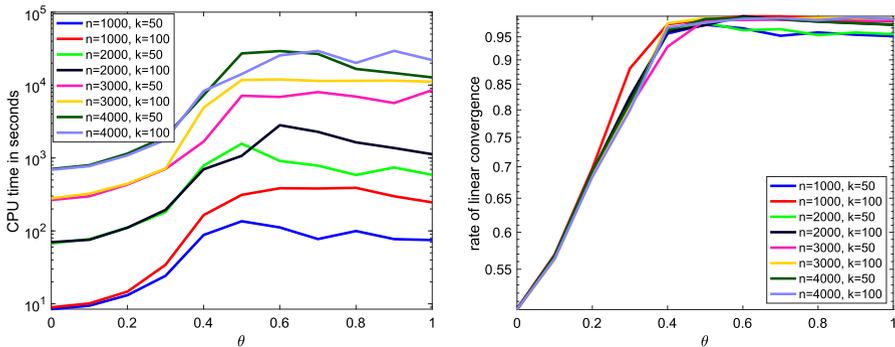


Fig. 2 CPU time and the estimated linear convergence rate by Algorithm 4.1 on 8 synthetic examples for $\theta \in [0, 1]$

by 0.1, we tested Algorithm 4.1 on a total of 88 problems (1.1). The stopping tolerance $\tau_{ol} = 10^{-7}$ in (4.2) and the maximum number of iterations is set to 10^3 .

Figure 1 displays the convergence curves of both objective function value and normalized NEPV residual, defined as the left-hand side of (4.2), by Algorithm 4.1 on 8 synthetic examples with selected $\theta \in \{0, 0.3, 0.5, 0.8\}$. As can be observed, most of the curves of normalized NEPV residual reach the preset tolerance $\tau_{ol} = 10^{-7}$ much earlier than the preset maximum number of iterations. For these synthetic examples, fewer numbers of iterations are required for smaller θ than larger ones. We point out that the tolerance 10^{-7} is often considered too tiny in machine learning applications. Also observe that all objective value curves are very much flat in fewer than 50 SCF iterations.

Figure 2 plots the CPU times by Algorithm 4.1 again on the 8 synthetic examples as θ varies. These times are well correlated with the size n . The larger n is, the more CPU time is consumed. For $\theta < 0.2$, the CPU times are comparable for all examples. As θ becomes large, more CPU times are consumed for the same (n, k) . This observation is consistent with our estimated rates of linear convergence, which are

always under 1 (demonstrating always convergence) but increase as θ does for those synthetic examples (demonstrating more iterations are needed for a larger θ than a smaller one). We caution the reader that in general, the rate of linear convergence by Algorithm 4.1 is unlikely an increasing function of θ for given A , B , and D .

6.2 Experiments on multi-view data for feature extraction

We will specialize the blocks of Φ and Ψ in (5.3) according to supervised multi-view subspace learning models GMA [37], MLDA [41], and MvMDA [7] as detailed in Sect. 5.1, yielding three different orthogonal MvSL (OMvSL) models in the form (5.4) that will be referred to, accordingly, as OGMA, OMLDA, and OMvMDA, where prefix “O” is for “Orthogonal” (as previously for OCCA [48]). Each of them can be solved by Algorithm 5.1 with either the Jacobi-style or Gauss–Seidel-style updating, leading to six OMvSL methods that will be distinguished further by a suffix “-J” or “-G”. For example, OGMA-J and OGMA-G are OMvSL (5.4) with Φ and Ψ specialised as in GMA and solved by Algorithm 5.1 with the Jacobi-style and Gauss–Seidel-style updating, respectively.

We evaluate the model (5.4) for multi-view feature extraction. Five data sets in Table 1 are used to evaluate the performance of the three proposed concrete models solved by the two updating schemes: OGMA-G, OGMA-J, OMLDA-G, OMLDA-J, OMvMDA-G, and OMvMDA-J, in terms of multi-view feature extraction by comparing them against their baseline counterparts: GMA, MLDA and MvMDA. We apply various feature descriptors to extract features of views, including CENTRIST [45], GIST [34], LBP [33], histogram of oriented gradient (HOG), color histogram (CH), and SIFT-SPM [21], from image data sets: Caltech101 [22] and Scene15 [21]. Multiple Features (mfeat) and Internet Advertisements (Ads) are publicly available from the UCI machine learning repository [11]. Dataset mfeat contains handwritten numeral data with six views including profile correlations (fac), Fourier coefficients of the character shapes (fou), Karhunen-Love coefficients (kar), morphological features (mor), pixel averages in 2×3 windows (pix), and Zernike moments (zer). Ads is used to predict whether or not a given hyperlink (associated with an image) is an advertisement and has three views: features based on the terms in the images URL, caption, and alt text (url+alt+caption), features based on the terms in the URL of the current site (origurl), and features based on the terms in the anchor URL (ancurl).

Except for MvMDA and its new variant: OMvMDA, all other models share the same trade-off parameter α to balance the pairwise correlations and supervised information. In our experiments, we tune $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ for proper balancing in supervised setting. To prevent possible singularity, we add a small value, e.g., 10^{-8} , to the diagonals of $\Psi_s \forall s$. For our proposed methods, an additional parameter θ is varied from 0 to 1 with an increase of 0.1. We also set the maximum number of iterations to 50 for both the SCF iteration of Algorithm 4.1 and the Jacobi-style or Gauss-Seidel-style updating of Algorithm 5.1. It is more of an empirical threshold observed as a good enough setting for multi-view feature extraction.

To evaluate the classification performance of compared methods, the 1-nearest neighbor classifier as the base classifier is employed. We run each method to

Table 1 Real world data sets, where the number of features for each view is shown inside the parentheses and ‘-’ for views not applicable

	mfeat	Caltech101-7	Caltech101-20	Scene15	Ads
Samples	2000	1474	2386	4310	3279
Classes	10	7	20	15	2
View 1	fac(216)	CENTRIST(254)	CENTRIST(254)	CENTRIST(254)	url+alt+caption(588)
View 2	fou(76)	GIST(512)	GIST(512)	GIST(512)	origurl(495)
View 3	kar(64)	LBP(1180)	LBP(1180)	LBP(531)	ancurl(472)
View 4	mor(6)	HOG(1008)	HOG(1008)	HOG(360)	–
View 5	pix(240)	CH(64)	CH(64)	SIFT-SPM(1000)	–
View 6	zer(47)	SIFT-SPM(1000)	SIFT-SPM(1000)	–	–

Table 2 Classification accuracy (\pm standard deviation) of multi-view feature extraction on the five real world data sets in Table 1 with 10% training and 90% testing over 10 random splits. The best θ is shown in the parentheses

Methods	mfeat	Caltech101-7	Caltech101-20	Scene15	Ads
GMA	93.99 \pm 0.87	93.25 \pm 1.04	81.16 \pm 0.94	61.41 \pm 1.30	92.59 \pm 1.76
OGMA-J	96.81 \pm 0.46 _(0.4)	95.14 \pm 0.59 _(0.4)	86.48 \pm 1.02 _(0.6)	79.90 \pm 0.80 _(1.0)	94.69 \pm 0.75 _(0.8)
OGMA-G	96.80 \pm 0.44 _(0.4)	95.07 \pm 0.56 _(0.5)	86.60 \pm 1.11 _(0.5)	79.90 \pm 1.02 _(1.0)	94.91 \pm 0.67 _(0.8)
MLDA	92.01 \pm 1.74	92.18 \pm 0.95	77.79 \pm 1.01	59.02 \pm 0.94	92.50 \pm 2.06
OMLDA-J	96.74 \pm 0.40 _(0.8)	94.68 \pm 0.48 _(0.8)	86.23 \pm 1.16 _(0.9)	81.42 \pm 1.07 _(1.0)	94.79 \pm 0.65 _(0.8)
OMLDA-G	96.82 \pm 0.38 _(0.8)	94.59 \pm 0.62 _(0.3)	86.09 \pm 1.22 _(0.9)	80.68 \pm 0.88 _(1.0)	94.76 \pm 0.74 _(0.8)
MvMDA	93.78 \pm 0.91	92.14 \pm 0.68	79.27 \pm 1.71	57.33 \pm 1.18	78.51 \pm 2.96
OMvMDA-J	96.62 \pm 0.31 _(0.5)	95.11 \pm 0.72 _(0.4)	85.69 \pm 0.87 _(0.4)	77.98 \pm 1.24 _(0.9)	94.02 \pm 1.54 _(0.6)
OMvMDA-G	96.63 \pm 0.37 _(0.4)	94.95 \pm 0.56 _(0.5)	85.76 \pm 1.00 _(0.4)	78.07 \pm 0.83 _(0.9)	93.52 \pm 0.59 _(0.0)

learn projection matrices with varying dimension of the common latent subspace $k \in \{2, 3, 5 : 5 : 30\}$ for all data sets except $k \in \{2, 3, 4, 5, 6\}$ for mfeat due to its smallest view mor having only 6 features. We split each data set into training and testing with ratio 10/90. The learned projection matrices are used to transform both training and testing data into the latent common space, and then the classifier is trained and tested in this space. Following [48], we employ the serial feature fusion strategy by concatenating projected features from all views. Classification accuracy is used to measure learning performance. Experimental results are reported in terms of the average and standard deviation over 10 randomly drawn splits.

Table 2 shows the classification accuracies and standard deviations of 9 multi-view feature extraction methods on five real world data sets over 10 random splits with 10% training and 90% testing. We have observed the following:

- (i) Our proposed methods consistently outperform their counterparts on all five data sets. The least improvement is about 2% on Ads, while the largest improvement about 20% occurs on Scene15.

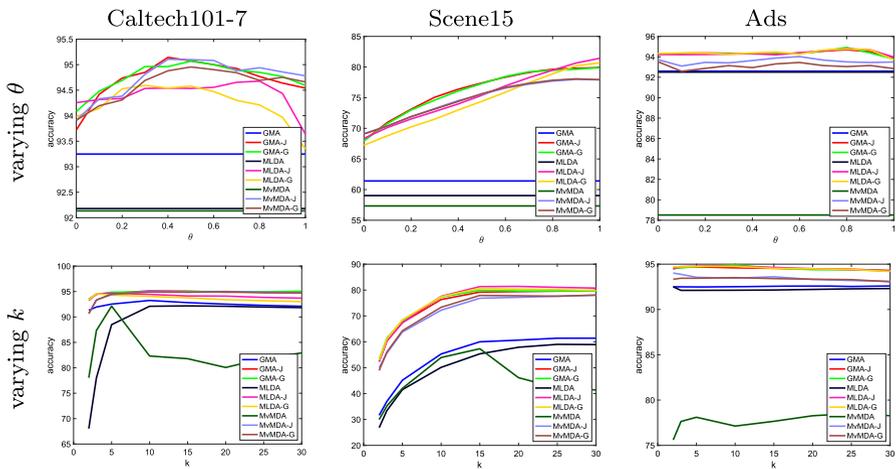


Fig. 3 Accuracies of compared methods on three data sets for $k \in \{2, 3, 4, 5, 6\}$ and $\theta \in [0, 1]$

(ii) The Jacobi-style and Gauss–Seidel-style updating on the same model achieve similar classification accuracies with differences within 0.8%.

This is great news for the Jacobi-style updating for which there is no guarantee that the objective value is monotonically increasing, unlike for the Gauss–Seidel-style updating.

(iii) MvMDA on Ads fails to produce a proper latent representation since its accuracy is 14% less than those of GMA and MLDA. However, both OMvMDA-J and OMvMDA-G perform very well on the same data set.

(iv) Accuracies by all three proposed models solved by two updating schemes are comparable and very good.

This is likely due to our trace ratio formulation with its orthogonality constraints in (5.4) that are known for their robustness [9, 48] as well as varying θ for weighting.

In Fig. 3 (the 1st row), we also report the classification accuracies of 9 methods with varying $\theta \in [0, 1]$. On Caltech101-7, Caltech101-20 and mfeat, the best results of our proposed methods are roughly around $\theta = 0.5$. However, different behaviors are found on Scene15 and Ads. θ does not show significantly impact on Ads, but we see better accuracy on Scene15 as θ increases. For almost all θ , our proposed methods consistently outperform baselines. This demonstrates that θ introduced in (5.4) can be useful to find better projection matrices for multi-view feature extraction. We further show the trend of classification accuracy by compared methods as the dimension k of common latent space increases in Fig. reffig:thetaspsk (the 2nd row). For any fixed k , our proposed methods outperform their counterparts. Importantly, all of our proposed methods nearly reach their best performances at fairly small k , while baseline methods have to use larger k to match that. This can be plausibly explained, namely, orthonormal bases retain less redundant information than non-orthonormal ones. We also observe that MvMDA behaves unstably for large k on Caltech101-7 and Scene15 since the accuracy suffers a significant drop, which does not happen to OMvMDA-G and OMvMDA-J. In summary, our proposed models not only demonstrate superior

performances to baseline methods but also are more robust to data noise and can achieve the same or better performance at smaller k . Small k implies fast computations if an iterative eigen-solver [2, 15, 26] is used in Algorithm 4.1 and that is extremely useful for large scale real world applications, such as cross-modal retrieval [7], for a fast response time due to less cost for computing pairwise distances in a lower dimensional latent space.

7 Conclusions and remarks

We have conducted an investigation, both in theory and numerical solutions, of the trace ratio maximization problem

$$\max_{X^T X = I_k} \frac{\text{trace}(X^T A X + X^T D)}{[\text{trace}(X^T B X)]^\theta}. \quad (7.1)$$

At least three special cases of it have been well studied in the past decades because of their immediate applications to data science. Our main results include an NEPv (nonlinear eigenvalue problem with eigenvector dependency, a term coined in [6]) formulation of its KKT condition, necessary conditions for its local and global maximizers, a complete picture of the role played by D on the maximizers, a guaranteed convergent self-consistent-field (SCF) iteration and its convergent analysis. As an application of these results, we propose a novel orthogonal multi-view subspace framework and experiment on its three instantiated models OGMA, OMLDA, and OMvMDA in either supervised or unsupervised setting. Numerical results demonstrate the new models outperform existing baselines.

Our convergence analysis results on the SCF-type iteration for the NEPv arising from the trace ratio maximization problem (7.1) are qualitative rather than quantitative, i.e., lacking a quantitative estimation on the actual rate of convergence, unfortunately. At the end of Sect. 4, we briefly commented on how to extend the approach in [3] which deals with NEPv with a unitarily invariance property to the NEPv here, that in general do not have the property, but only for the case when $X_*^T D$ is positive definite. In general when $X_*^T D$ is singular, it is much more involved. Looking ahead, in [29] we will develop a local convergence theory that covers the latter case and can yield quantitative estimations on the actual rate of convergence.

Although we have been limiting our discussion on real matrices, the developments in this paper can be straightforwardly extended to complex matrices with minor modifications, namely, replace all \mathbb{R} by \mathbb{C} (the set of complex numbers) and all transposes T by complex conjugate transposes H .

Acknowledgements The authors wish to thank the two anonymous referees for their constructive suggestions that greatly improved the presentation of this paper.

Author Contributions LW, L-HZ, R-CL All authors contribute equally.

Fundings Research was supported in part by United States National Science Foundation DMS-1719620 and DMS-2009689, and by the National Natural Science Foundation of China NSFC-12071332.

Data Availability Statement All used data are public available online.

Code Availability Available upon request.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: *Optimization Algorithms On Matrix Manifolds*. Princeton University Press, Princeton (2008)
2. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia (2000)
3. Bai, Z., Li, R.C., Lu, D.: Sharp estimation of convergence rate for self-consistent field iteration to solve eigenvector-dependent nonlinear eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **43**(1), 301–327 (2022)
4. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 423–443 (2018)
5. Borg, I., Lingoes, J.: *Multidimensional Similarity Structure Analysis*. Springer, New York (1987)
6. Cai, Y., Zhang, L.H., Bai, Z., Li, R.C.: On an eigenvector-dependent nonlinear eigenvalue problem. *SIAM J. Matrix Anal. Appl.* **39**(3), 1360–1382 (2018)
7. Cao, G., Iosifidis, A., Chen, K., Gabbouj, M.: Generalized multi-view embedding for visual recognition and cross-modal retrieval. *IEEE Trans. Cybern.* **48**(9), 2542–2555 (2018)
8. Chu, M.T., Trendafilov, N.T.: The orthogonally constrained regression revisited. *J. Comput. Graph. Stat.* **10**(4), 746–771 (2001)
9. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: survey, insights, and generalizations. *J. Mach. Learn. Res.* **16**, 2859–2900 (2015)
10. Demmel, J.: *Applied Numerical Linear Algebra*. SIAM, Philadelphia (1997)
11. Dua, D., Graff, C.: UCI machine learning repository (2017). <http://archive.ics.uci.edu/ml>
12. de Geer, J.P.V.: Linear relations among k sets of variables. *Psychometrika* **49**, 70–94 (1984)
13. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**(2), 303–353 (1999)
14. Eldén, L., Park, H.: A procrustes problem on the Stiefel manifold. *Numer. Math.* **82**, 599–619 (1999)
15. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 4th edn. Johns Hopkins University Press, Baltimore (2013)
16. Gower, J.C., Dijksterhuis, G.B.: *Procrustes Problems*. Oxford University Press, New York (2004)
17. Horn, R.A., Johnson, C.R.: *Topics in Matrix Analysis*. Cambridge University Press, Cambridge (1991)
18. Hurley, J.R., Cattell, R.B.: The Procrustes program: producing direct rotation to test a hypothesized factor structure. *Behav. Sci.* **7**, 258–262 (1962)
19. Kanzow, C., Qi, H.D.: A QP-free constrained Newton-type method for variational inequality problems. *Math. Program.* **85**, 81–106 (1999)
20. Kushmerick, N.: Learning to remove internet advertisements. In: *Proceedings of the Third Annual Conference on Autonomous Agents*, pp. 175–181 (1999)
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178. IEEE (2006)
22. Li, F.F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* **106**(1), 59–70 (2007)
23. Li, R.C.: A perturbation bound for the generalized polar decomposition. *BIT* **33**, 304–308 (1993)
24. Li, R.C.: On perturbations of matrix pencils with real spectra. *Math. Comput.* **62**, 231–265 (1994)
25. Li, R.C.: New perturbation bounds for the unitary polar factor. *SIAM J. Matrix Anal. Appl.* **16**, 327–332 (1995)
26. Li, R.C.: Rayleigh quotient based optimization methods for eigenvalue problems. In: Bai, Z., Gao, W., Su, Y. (eds.) *Matrix Functions and Matrix Equations*, Series in Contemporary Applied Mathematics.

- Lecture summary for 2013 Gene Golub SIAM Summer School vol. 19, pp. 76–108. World Scientific, Singapore (2015)
27. Li, W., Sun, W.: Perturbation bounds for unitary and subunitary polar factors. *SIAM J. Matrix Anal. Appl.* **23**, 1183–1193 (2002)
 28. Liu, X.G., Wang, X.F., Wang, W.G.: Maximization of matrix trace function of product Stiefel manifolds. *SIAM J. Matrix Anal. Appl.* **36**(4), 1489–1506 (2015)
 29. Lu, D., Li, R.C.: Convergence of SCF for NEPv without unitary invariance property (2022). Work-in-progress
 30. Moré, J.J., Sorensen, D.C.: Computing a trust region step. *SIAM J. Sci. Statist. Comput.* **4**(3), 553–572 (1983)
 31. Ngo, T., Bellalij, M., Saad, Y.: The trace ratio optimization problem for dimensionality reduction. *SIAM J. Matrix Anal. Appl.* **31**(5), 2950–2971 (2010)
 32. Nie, F., Zhang, R., Li, X.: A generalized power iteration method for solving quadratic problem on the Stiefel manifold. *Sci. China Info. Sci.* **60**, 112101:1–112101:10 (2017)
 33. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
 34. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**(3), 145–175 (2001)
 35. Peng, Y., Qi, J.: CM-GANs: cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **15**(1), 1–24 (2019)
 36. Seber, G.A.F.: *A Matrix Handbook for Statisticians*. Wiley, New York (2007)
 37. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2160–2167. IEEE (2012)
 38. Stewart, G.W.: *Matrix Algorithms, Vol. II: Eigensystems*. SIAM, Philadelphia (2001)
 39. Stewart, G.W., Sun, J.G.: *Matrix Perturbation Theory*. Academic Press, Boston (1990)
 40. Sun, J.G.: *Matrix Perturbation Analysis*. Academic Press, Beijing (1987). **(In Chinese)**
 41. Sun, S., Xie, X., Yang, M.: Multiview uncorrelated discriminant analysis. *IEEE Trans. Cybern.* **46**(12), 3272–3284 (2016)
 42. von Neumann, J.: Some matrix-inequalities and metrization of matrix-space. *Tomck. Univ. Rev.* **1**, 286–300 (1937)
 43. Vía, J., Santamaría, I., Pérez, J.: A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Netw.* **20**(1), 139–152 (2007)
 44. Wang, H., Yan, S., Xu, D., Tang, X., Huang, T.: Trace ratio vs. ratio trace for dimensionality reduction. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
 45. Wu, J., Rehg, J.M.: Where am i: Place instance and category recognition using spatial pact. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
 46. Zhang, L.H., Liao, L.Z., Ng, M.K.: Fast algorithms for the generalized Foley–Sammon discriminant analysis. *SIAM J. Matrix Anal. Appl.* **31**(4), 1584–1605 (2010)
 47. Zhang, L.H., Liao, L.Z., Ng, M.K.: Superlinear convergence of a general algorithm for the generalized Foley–Sammon discriminant analysis. *J. Optim. Theory Appl.* **157**(3), 853–865 (2013)
 48. Zhang, L.H., Wang, L., Bai, Z., Li, R.C.: A self-consistent-field iteration for orthogonal canonical correlation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2), 890–904 (2022). <https://doi.org/10.1109/TPAMI.2020.3012541>
 49. Zhang, L.H., Yang, W.H., Shen, C., Ying, J.: An eigenvalue-based method for the unbalanced Procrustes problem. *SIAM J. Matrix Anal. Appl.* **41**(3), 957–983 (2020)
 50. Zhang, Z., Du, K.: Successive projection method for solving the unbalanced procrustes problem. *Sci. China Math.* **49**(7), 971–986 (2006)
 51. Zhao, H., Wang, Z., Nie, F.: Orthogonal least squares regression for feature extraction. *Neurocomputing* **216**, 200–207 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.