SIAM J. MATRIX ANAL. APPL.
Vol. 41, No. 3, pp. 957–983

AN EIGENVALUE-BASED METHOD FOR THE UNBALANCED PROCRUSTES PROBLEM*

LEI-HONG ZHANG†, WEI HONG YANG‡, CHUNGEN SHEN§, AND JIAQI YING¶

Abstract. In this paper, we propose a novel eigenvalue-based approach to solving the unbalanced orthogonal Procrustes problem. By making effective use of the necessary condition for the global minimizer and the orthogonal constraint, we shall first show that the unbalanced Procrustes problem can be equivalently transformed into an eigenvalue minimization whose solution can be computed by solving a related eigenvector-dependent nonlinear eigenvalue problem. Through the exploitation of certain techniques in the nonlinear eigenvalue computations, we adapt the standard self-consistent field (SCF) iteration to solve the resulting nonlinear eigenvalue problem. Theoretical convergence analysis of this customized SCF iteration is performed, and practical strategies for a more efficient numerical implementation are discussed. Our numerical experience on preliminary tests indicates that the proposed eigenvalue-based SCF iteration is a promising method for the unbalanced orthogonal Procrustes problem.

Key words. Procrustes problem, nonlinear eigenvalue problem, self-consistent field iterations

AMS subject classifications. 65F15, 65H17, 90C30

DOI. 10.1137/19M1270872

1. Introduction.

Given matrices $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{m\times \ell}$ with $n \geq \ell$, the Procrustes problem minimizes the square of the Frobenius norm $\|CX - D\|_F$ over the Stiefel manifold $\mathbb{O}^{n\times\ell} := \{X \in \mathbb{R}^{n\times\ell} | X^{\mathrm{T}}X = I_\ell\}$, i.e.,

$$\min_{X\in\mathbb{O}^{n\times\ell}} \|CX - D\|_F^2. \tag{1.1}$$

Equivalently, by expanding $\|CX - D\|_F^2$ and ignoring the constant $\|D\|_F^2$, (1.1) can be written as a problem to minimize a quadratic function over $\mathbb{O}^{n\times\ell}$, i.e.,

$$\min_{X\in\mathbb{O}^{n\times\ell}} \{f(X) := \operatorname{trace}(X^{\mathrm{T}}AX) + 2\operatorname{trace}(X^{\mathrm{T}}B)\}, \tag{1.2}$$

where $A = C^{\mathrm{T}}C$ and $B = -C^{\mathrm{T}}D$.

Originally, the terminology of the Procrustes problem (with $n = \ell$) is due to Hurley and Cattell [19], who tried to find an estimated transformation matrix X to relate a factor structure C from factor analysis to a specific hypothetical matrix D (see also [17]). The special case with $n = \ell$ is commonly called the balanced Procrustes

*Received by the editors June 27, 2019; accepted for publication (in revised form) by B. Vandereycken March 27, 2020; published electronically July 1, 2020.
https://doi.org/10.1137/19M1270872
Funding: The work of the first author was partially supported by the National Natural Science Foundation of China grant 11671246, and the National Key R & D Program of China (2018YFB0204404). The work of the second author was supported by the National Natural Science Foundation of China grant 11971118.

†School of Mathematical Sciences and Institute of Computational Science, Soochow University, Suzhou 215006, Jiangsu, China, and School of Mathematics, Shanghai University of Finance and Economics, Shanghai, 200433, China (longzlh@suda.edu.cn).
‡School of Mathematical Sciences, Fudan University, Shanghai, 200433, People's Republic of China (whyang@fudan.edu.cn).
§College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China (shenchungen@gmail.com).
¶School of Mathematics, Shanghai University of Finance and Economics, Shanghai 200433, China (candy201314l@qq.com).

problem [14, 16, 30, 38, 49], and until now, there are several variants of (1.1) arising from statistics, psychometrika, biometrics, data mining, and other fields. For example, instead of the Frobenius norm, [41] considers the solutions of the Procrustes problem in a family of orthogonally invariant norms; [18, 3] discuss the solutions X subject to the set of symmetric matrices and certain closed cones (the cone of symmetric positive semidefinite and the cone of (symmetric) elementwise nonnegative matrices), respectively.

The minimization (1.1) with $n > \ell$ is referred to as the unbalanced Procrustes problem [8, 12, 14, 19, 17, 49]. Applications of the unbalanced Procrustes problem can arise, for example, in the orthogonal least squares regression (OLSR) for feature extraction [51, 30], the multidimensional similarity structure analysis (SSA) [6, Chapter 19], and the Maxbet problem [10, 26] in the canonical correlation analysis. The model OLSR is a supervised learning method in data mining, which aims at finding an orthogonal transformation matrix $X \in \mathbb{O}^{n\times\ell}$ to project high dimensional data (with dimension n) into a lower dimensional space (with dimension $\ell \ll n$). The mathematical formulation of OLSR is exactly an unbalanced Procrustes problem (1.1), and in section 6.2, we will give a detailed description of OLSR and apply our proposed method to solve it. In SSA, two configurations represented by matrices C and D are *similar* if they can be brought to a complete match by rigid motions and dilations [6, Chapter 19], and the Procrustes problem (1.1) is used to fit one configuration C to the other D. As pointed out in [6, section 19.8] two configurations are generally of the same dimensionality (i.e., $n = \ell$), but sometimes it is useful for them to have different dimensionality. A possible treatment of the latter leads to the unbalanced Procrustes problem, and in section 6.3, we will consider such an example. Another application is from solving a subproblem [26, section 4] of Maxbet in the form of (1.2), in which ℓ is the reduced dimension of a high dimensional random variable, and usually $\ell \ll n$.

For numerical methods, it is known that the balanced Procrustes problem admits a closed form solution which can be computed by the polar decomposition (e.g., [16]). Unfortunately, for the unbalanced case, there is no direct way to compute the solution X. In fact, the special case with $\ell = 1$ is widely known as the trust-region subproblem of the trust-region method in optimization (see, e.g., [31, 29]); even for this special case, it may admit a local but nonglobal minimizer [28], and computing the global solution is generally not an easy task. Due to the lack of closed form solution, necessary or sufficient conditions for the local and/or global minimizer of the unbalanced Procrustes problem (or the minimization (1.2)) have been developed; particularly, [8, 14, 50] establish a couple of useful and in-depth conditions that can be used to check if a computed one can be a local or a global solution.

For the general case $n > \ell > 1$, certain numerical approaches for computing an approximation to (1.2) or (1.1) are available. Besides resorting to generic optimization methods built on the Stiefel manifold (e.g., [1, 2, 12, 42]), several structure-exploiting algorithms have been proposed. For example, [5] presents an iterative method based on relaxation which produces a sequence that gradually reduces the objective function; in each iteration, the method finds a best plane rotation in a particularly chosen plane, and the resulting subproblem of computing the best plane rotation is called a planner Procrustes problem (i.e., $m = n = 2$ and $\ell = 1$ in (1.1)). In [49], a successive projection minimization approach is introduced in which, by alternatively fixing all but one column of X, each iteration solves a trust-region subproblem (i.e., (1.2) with $\ell = 1$) to update the approximation. Other numerical algorithms include the generalized power iteration in [30] and Newton-type iterations in [38]. However, none of them is an eigenvalue-based method so that the modern power of the state-

of-the-art eigensolvers can be exploited.

In this paper, we propose a self-consistent-field (SCF) iteration for the unbalanced Procrustes problem (1.1) or its equivalent minimization (1.2). The SCF iteration is an efficient method for solving a class of eigenvector-dependent nonlinear eigenvalue problems (NEPvs) [7], and can exploit the presently used relatively mature techniques in eigenvalue computations. It was originally used in electronic structure calculations (e.g., [25, 27, 33, 43]), and appears very useful in many applications arising recently from data science (e.g., [4, 24, 40, 44, 45, 48, 46, 47]). A general discussion on SCF together with some convergence results can be found in, e.g., [7]. For the unbalanced Procrustes problem (1.1) and/or (1.2), our contributions in this paper include the following three aspects: (i) by making use of the necessary conditions for the global minimizer and the orthogonal constraint $X \in \mathbb{O}^{n\times\ell}$, we construct an equivalent eigenvalue minimization of (1.2) for which a related NEPv can be established, (ii) by the general SCF iteration scheme and the special structure of (1.2), we propose a customized SCF iteration and establish the convergence, and (iii) we introduce a subspace acceleration procedure to improve the performance of the basic SCF. Numerical simulations are carried out and demonstrate that our eigenvalue-based SCF iteration is an efficient method for the unbalanced orthogonal Procrustes problem, particularly for the case $\ell \ll n$.

We organize the remainder of this paper as follows: In section 2, we first summarize some basic optimality conditions for the local and/or global minimizer of (1.2). Relying on the necessary conditions, in section 3, we then transform the problem (1.2) into an eigenvalue minimization, and also propose the eigenvector-dependent NEVp. The basic form of our SCF iteration is introduced in this section. In section 4, we shall present the convergence analysis of the SCF iteration. To make the basic SCF iteration more efficient, in section 5, we then discuss two practical treatments: the inexact SCF iteration and the subspace acceleration. Numerical results of our practical SCF iteration are reported in section 6, and final remarks are drawn in section 7.

Notation. Throughout this paper, all vectors are column vectors and are typeset in bold lowercase letters. For a matrix $C \in \mathbb{R}^{n\times m}$, C^{T} and $\mathcal{R}(C)$ denote its transpose and range space, respectively. The $n \times n$ identity matrix is I_n, and $\boldsymbol{e}_j$ is the jth column of an identity matrix whose size is determined by the context. If C is square, we denote the set of its eigenvalues by $\mathrm{eig}(C)$. When all of the eigenvalues of C are real, we order them as

$$\lambda_1(C) \le \lambda_2(C) \le \cdots \le \lambda_n(C).$$

The eigendecomposition of A is given by $A = U\Theta U^{\mathrm{T}}$ and $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ with the arranged eigenvalues

$$\theta_1 \le \theta_2 \le \cdots \le \theta_n,$$

where $\theta_i = \lambda_i(A)$ is reserved in this paper. When A is positive semidefinite (positive definite), we denote it by $A \succcurlyeq 0$ ($\succ 0$). The notation $\|\cdot\|_{\mathrm{ui}}$ stands for any unitary invariant norm, and it holds that

$$\|XYZ\|_{\mathrm{ui}} \le \|X\|_2 \, \|Y\|_{\mathrm{ui}} \, \|Z\|_2.$$

To simplify our presentation, we shall also adopt MATLAB-like convention to access the (i, j)th entry of A as $A_{(i,j)}$.

2. Optimality conditions. For the trust-region subproblem (i.e., $\ell = 1$), the optimality conditions due to Gay [15] and Moré and Sorensen [29] (see also [35] and [31, Theorem 4.1]) are well known. The generalization for (1.2) or, equivalently, (1.1) has been discussed in [8, 14], and we summarize some basic facts below.

2.1. Local optimality conditions.

LEMMA 2.1 (first-order optimality [8, Theorem 3.8] and [14, Theorem 3.1]). *If $X \in \mathbb{O}^{n\times\ell}$ is a local minimizer, then there is a symmetric matrix $\Lambda \in \mathbb{R}^{\ell\times\ell}$ such that*

$$AX + X\Lambda = -B. \tag{2.1}$$

Note from (2.1) and the constraint $X^{\mathrm{T}}X = I_\ell$ that $\Lambda = -X^{\mathrm{T}}B - X^{\mathrm{T}}AX$ is indeed determined by the local solution X, and also the matrix $X^{\mathrm{T}}B = B^{\mathrm{T}}X$ is symmetric. Moreover, we will further see in Theorem 2.4 that if X is a global minimizer, the matrix $X^{\mathrm{T}}AX + \Lambda = -X^{\mathrm{T}}B$ is positive semidefinite.

Note that the tangent space $\mathcal{T}_X\mathbb{O}^{n\times\ell}$ at any $X \in \mathbb{O}^{n\times\ell}$ can be expressed as, e.g., [2, p. 42]

$$\mathcal{T}_X\mathbb{O}^{n\times\ell} := \{H \in \mathbb{R}^{n\times\ell} \,:\, H^{\mathrm{T}}X + X^{\mathrm{T}}H = 0\} \tag{2.2a}$$

$$= \{H = XK + (I_n - XX^{\mathrm{T}})J \,:\, K = -K^{\mathrm{T}} \in \mathbb{R}^{\ell\times\ell}, J \in \mathbb{R}^{n\times\ell}\}. \tag{2.2b}$$

With the standard inner product (or the Frobenius inner product) induced on $\mathcal{T}_X\mathbb{O}^{n\times\ell}$

$$\langle Z, Y\rangle = \mathrm{trace}(Z^{\mathrm{T}}Y) \quad \forall\, Z, Y \in \mathcal{T}_X\mathbb{O}^{n\times\ell},$$

it is known that the orthogonal projection of $Z \in \mathbb{R}^{n\times\ell}$ onto the tangent space $\mathcal{T}_X\mathbb{O}^{n\times\ell}$ at X is given by [2, eq. (3.35)]

$$\Pi_X(Z) := Z - X\frac{X^{\mathrm{T}}Z + Z^{\mathrm{T}}X}{2}. \tag{2.3}$$

LEMMA 2.2 (second-order optimality [8, Theorem 3.2] and [14, p. 609]). *If X is a local minimizer of* (1.2)*, then*

$$\mathrm{trace}(H^{\mathrm{T}}AH) + \mathrm{trace}(H\Lambda H^{\mathrm{T}}) \geq 0 \quad \forall H \in \mathcal{T}_X\mathbb{O}^{n\times\ell};$$

moreover, if the strict inequality holds for any nonzero tangent vector $H \in \mathcal{T}_X\mathbb{O}^{n\times\ell}$ then X is a strict local minimizer.

It is worth mentioning that Lemma 2.2 is just the generalization of the second-order optimality condition [28, Lemma 2.1] for the trust-region subproblem (see also [35, 29] and [31, Theorem 4.1]), i.e., $\ell = 1$. In the general case $\ell > 1$, we can see that the elements of Λ are just the Lagrangian multipliers, which are uniquely given by X.

With the representation of the tangent vector given in (2.2b), we can further refine, as in [8] for the Procrustes problem (1.1), the second-order optimality condition of Lemma 2.2 as follows.

COROLLARY 2.3. *If X is a local minimizer of* (1.2)*, then*

(i) *for any skew-symmetric $K \in \mathbb{R}^{\ell\times\ell}$ and any $J \in \mathbb{R}^{n\times\ell}$, it holds that*

$$\begin{aligned}&\mathrm{trace}\left(J^{\mathrm{T}}\left[A + X\Lambda X^{\mathrm{T}} + (2I_n - XX^{\mathrm{T}})BX^{\mathrm{T}}\right]J\right) - \mathrm{trace}(X^{\mathrm{T}}J\Lambda J^{\mathrm{T}}X)\\ &+ \mathrm{trace}(J\Lambda J^{\mathrm{T}}) + 2\,\mathrm{trace}\left(J^{\mathrm{T}}(XB^{\mathrm{T}}X + B)K\right) - \mathrm{trace}(K^{\mathrm{T}}X^{\mathrm{T}}BK) \geq 0,\end{aligned} \tag{2.4}$$

and moreover, if the strict inequality holds for any skew-symmetric $K \in \mathbb{R}^{\ell\times\ell}$ and any $J \in \mathbb{R}^{n\times\ell}$ with $XK + (I_n - XX^{\mathrm{T}})J \neq 0$, then X is a strict local minimizer;

(ii) *if* $\ell \geq 2$, *then* $\operatorname{trace}(X^{\mathrm{T}}B) \leq 0$.

Proof. For item (i), it is a straightforward task to verify (2.4) by using the facts that (a) $X^{\mathrm{T}}X = I_\ell$, (b) Lemma 2.1, (c) Lemma 2.2, and (d) the representation of $\mathcal{T}_X\mathbb{O}^{n\times\ell}$ in (2.2b). The detailed manipulations are omitted.

For (ii), we particularly choose $J = 0$ in (2.4), or equivalently, $H = XK$ in Lemma 2.2 to have

$$\operatorname{trace}(K^{\mathrm{T}}X^{\mathrm{T}}BK) \leq 0 \quad \forall K = -K^{\mathrm{T}} \in \mathbb{R}^{\ell\times\ell}.$$

By the symmetry of $X^{\mathrm{T}}B$, let $X^{\mathrm{T}}B = W^{\mathrm{T}}\Sigma W$ be the spectral decomposition of $X^{\mathrm{T}}B$, and $\Sigma = \operatorname{diag}(\mu_1, \ldots, \mu_\ell)$. Choose $K = K_{12} := W \operatorname{diag}(\Xi, 0, \ldots, 0)W^{\mathrm{T}}$, where $\Xi = \left[\begin{smallmatrix} 0 & \zeta \\ -\zeta & 0 \end{smallmatrix}\right]$ for $\zeta > 0$; then it follows that

$$(\mu_1 + \mu_2)\zeta^2 \leq 0 \Longrightarrow \mu_1 + \mu_2 \leq 0.$$

Similarly, we can define K_{ij} for $1 \leq i < j \leq \ell$ and add all of the resulting inequalities to have

$$(\ell - 1)(\mu_1 + \cdots + \mu_\ell)\zeta^2 \leq 0 \Longrightarrow \mu_1 + \cdots + \mu_\ell = \operatorname{trace}(X^{\mathrm{T}}B) \leq 0.$$

This completes the proof. ∎

As the special case of $\ell = 1$, we remark that Corollary 2.3(ii) may not hold. As an example with $A = \operatorname{diag}(1, -1)$ and $B = [0, 1]^{\mathrm{T}}$, the minimization

$$\min_{x_1^2 + x_2^2 = 1} x_1^2 - x_2^2 + 2x_2$$

has a local minimizer $X = [0, 1]^{\mathrm{T}}$, but $B^{\mathrm{T}}X = 1 > 0$.

2.2. Global optimality conditions. Suppose X_{opt} is a global minimizer. Then for any orthogonal $Q \in \mathbb{R}^{\ell\times\ell}$, we have

$$\begin{aligned} f(X_{\mathrm{opt}}) \leq f(X_{\mathrm{opt}}Q) &= \operatorname{trace}(Q^{\mathrm{T}}X_{\mathrm{opt}}^{\mathrm{T}}AX_{\mathrm{opt}}Q) + 2\operatorname{trace}(Q^{\mathrm{T}}X_{\mathrm{opt}}^{\mathrm{T}}B) \\ &= \operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}}AX_{\mathrm{opt}}) + 2\operatorname{trace}(Q^{\mathrm{T}}X_{\mathrm{opt}}^{\mathrm{T}}B). \end{aligned}$$

This leads to $\operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}}B) \leq \operatorname{trace}(Q^{\mathrm{T}}X_{\mathrm{opt}}^{\mathrm{T}}B)$ for any orthogonal $Q \in \mathbb{R}^{\ell\times\ell}$. Now, let $X_{\mathrm{opt}}^{\mathrm{T}}B = U\Gamma U^{\mathrm{T}}$ be the spectral decomposition (note that $X_{\mathrm{opt}}^{\mathrm{T}}B$ is symmetric), and we can conclude $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell) \preccurlyeq 0$ because, otherwise, we can choose $Q = U \operatorname{diag}(-\operatorname{sign}(\gamma_1), \ldots, -\operatorname{sign}(\gamma_\ell))U^{\mathrm{T}}$ to have $\operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}}B) > \operatorname{trace}(Q^{\mathrm{T}}X_{\mathrm{opt}}^{\mathrm{T}}B)$. This then shows $-X_{\mathrm{opt}}^{\mathrm{T}}B \succcurlyeq 0$. Hence, using (2.1), one has that

$$-B^{\mathrm{T}}X = X^{\mathrm{T}}AX + \Lambda \succcurlyeq 0. \tag{2.5}$$

Consequently, the Cauchy interlacing inequalities [37]

$$\theta_i \leq \lambda_i(X^{\mathrm{T}}AX) \leq \theta_{n-\ell+i}$$

and Weyl's theorem [37]

$$0 \leq \lambda_{\ell-r+1}(X^{\mathrm{T}}AX + \Lambda) \leq \lambda_{\ell-r+j}(X^{\mathrm{T}}AX) + \lambda_{\ell-j+1}(\Lambda) \leq \theta_{n-r+j} + \lambda_{\ell-j+1}(\Lambda)$$

$\forall\, 1 \leq j \leq r,\ 1 \leq r \leq \ell$ yield a necessary optimality condition (see, e.g., [14, Theorem 3.8] or [50, Theorem 3.2]).

THEOREM 2.4. *If X is a global minimizer of* (1.2)*, then*

$$X^{\mathrm{T}}AX + \Lambda = -X^{\mathrm{T}}B = -B^{\mathrm{T}}X \succcurlyeq 0, \tag{2.6}$$

which implies

$$\lambda_i(\Lambda) \geq -\theta_{n-i+1} \quad \textit{for } 1 \leq i \leq \ell.$$

There are other types of necessary or sufficient optimality conditions for (1.2), and the reader is referred to [8, 14, 50] for more details.

3. A self-consistent field iteration. In this section, we will propose a novel eigenvalue-based algorithm for solving (1.2). Our method essentially targets the KKT condition given in (2.1) but makes effective use of the necessary condition (2.6) for the global minimizer.

3.1. A related NEPv. Relying on the following well-known von Neumann's trace inequality [39] (see also [34, eq. 6.81]), we will first, equivalently, transform (1.2) into an eigenvalue-related minimization, and then introduce an eigenvector-dependent NEPv for which an SCF iteration can apply.

LEMMA 3.1. *If C_1, C_2 are n-by-n complex matrices with singular values*

$$\sigma_1(C_1) \leq \cdots \leq \sigma_n(C_1), \quad \sigma_1(C_2) \leq \cdots \leq \sigma_n(C_2),$$

respectively, then

$$-\sum_{i=1}^{n} \sigma_i(C_1)\sigma_i(C_2) \leq \mathrm{RE}\left(\mathrm{trace}(C_1^{\mathrm{H}}C_2)\right) \leq \sum_{i=1}^{n} \sigma_i(C_1)\sigma_i(C_2).$$

Equality holds on the right when the singular value decomposition of C_2 is $C_2 = U\Sigma_2 V^{\mathrm{H}}$, and equality on the left holds when $C_2 = -U\Sigma_2 V^{\mathrm{H}}$, where $C_1 = U\Sigma_1 V^{\mathrm{H}}$ is the singular value decomposition of C_1, and $\Sigma_i = \mathrm{diag}(\sigma_1(C_i), \sigma_2(C_i), \ldots, \sigma_n(C_i))$ for $i = 1, 2$.

Define a symmetric matrix function

$$E(X) := A + XB^{\mathrm{T}} + BX^{\mathrm{T}} \in \mathbb{R}^{n\times n}. \tag{3.1}$$

By making use of the necessary optimality condition (2.1) and the orthogonal constraint $X^{\mathrm{T}}X = I_\ell$, we have the following lemma.

LEMMA 3.2. *Let $X \in \mathbb{O}^{n\times\ell}$ satisfying $X^{\mathrm{T}}B = B^{\mathrm{T}}X$. Then X is a KKT point satisfying* (2.1) *if and only if X is an orthonormal eigenbasis matrix of $E(X)$.*

Proof. For a KKT point X, by $X^{\mathrm{T}}X = I_\ell$, we can rewrite (2.1) as

$$(A + XB^{\mathrm{T}} + BX^{\mathrm{T}})X = -X\Lambda + XB^{\mathrm{T}}X \Longleftrightarrow E(X)X = X\Psi_X$$

where $\Psi_X = -\Lambda + B^{\mathrm{T}}X$ is symmetric. The converse also holds by the above relation and $X^{\mathrm{T}}B = B^{\mathrm{T}}X$. ∎

The following theorem further ensures that any global minimizer X_{opt} of (1.2) is an orthonormal eigenbasis matrix of $E(X_{\mathrm{opt}})$ associated with the first ℓ smallest eigenvalues.

THEOREM 3.3. *For the matrix function $E(X)$ in* (3.1)*, we have*

$$\min_{X\in\mathbb{O}^{n\times\ell}} \sum_{i=1}^{\ell} \lambda_i(E(X)) = \min_{X\in\mathbb{O}^{n\times\ell}} f(X). \tag{3.2}$$

That is, the minimization (1.2) *is equivalent to finding $X \in \mathbb{O}^{n\times\ell}$ so that the sum of the first ℓ smallest eigenvalues of $E(X)$ is minimized.*[1] *Moreover, any global minimizer $X_{\rm opt}$ of* (1.2) *is an orthonormal eigenbasis matrix of $E(X_{\rm opt})$, i.e.,*

$$E(X_{\rm opt})X_{\rm opt} = X_{\rm opt}\Psi_{\rm opt}, \tag{3.3}$$

associated with the first ℓ smallest eigenvalues, and $\operatorname{trace}(\Psi_{\rm opt}) = f(X_{\rm opt})$.

Proof. Let $X \in \mathbb{O}^{n\times\ell}$ be arbitrarily given, and let Z be an orthonormal eigenbasis matrix of $E(X)$ corresponding to the ℓ smallest eigenvalues. Note that

$$\sum_{i=1}^{\ell} \lambda_i(E(X)) = \operatorname{trace}(Z^{\rm T}E(X)Z) = \operatorname{trace}(Z^{\rm T}AZ) + 2\operatorname{trace}(Z^{\rm T}BX^{\rm T}Z). \tag{3.4}$$

Moreover, we can choose a suitable orthonormal basis $\widetilde{Z} = ZQ$ for $\mathcal{R}(Z)$ to ensure $\operatorname{trace}(\widetilde{Z}^{\rm T}E(X)\widetilde{Z}) \geq f(X_{\rm opt})$. The orthonormal matrix $Q \in \mathbb{R}^{\ell\times\ell}$ is obtained by minimizing $\operatorname{trace}(Q^{\rm T}Z^{\rm T}B)$ over $\mathbb{O}^{\ell\times\ell}$. In fact,

$$\min_{Q^{\rm T}Q=I_\ell} \operatorname{trace}(Q^{\rm T}Z^{\rm T}B) = -\operatorname{trace}(\Sigma) = -\sum_{i=1}^{\ell} \sigma_i(Z^{\rm T}B),$$

where the minimum is achieved when $Q = -UV^{\rm T}$ and $Z^{\rm T}B = U\Sigma V^{\rm T}$ is the SVD of $Z^{\rm T}B$ and $\Sigma = \operatorname{diag}(\sigma_1(Z^{\rm T}B), \ldots, \sigma_\ell(Z^{\rm T}B))$ with $\sigma_1(Z^{\rm T}B) \leq \cdots \leq \sigma_\ell(Z^{\rm T}B)$. Note that $UV^{\rm T}$ is the orthogonal polar factor [16, section 9.4.3] of $Z^{\rm T}B$. On the other hand, by using Lemma 3.1, we have

$$\operatorname{trace}(Z^{\rm T}BX^{\rm T}Z) \geq -\sum_{i=1}^{\ell} \sigma_i(Z^{\rm T}B)\sigma_i(X^{\rm T}Z),$$

where $\sigma_1(X^{\rm T}Z) \leq \cdots \leq \sigma_\ell(X^{\rm T}Z)$ are the singular values of $X^{\rm T}Z$, which are all less than 1 by the facts $X, Z \in \mathbb{O}^{n\times\ell}$. Setting $\widetilde{Z} = ZQ$, we have

$$\begin{aligned}\operatorname{trace}(\widetilde{Z}^{\rm T}BX^{\rm T}\widetilde{Z}) &= \operatorname{trace}(Z^{\rm T}BX^{\rm T}Z)\\ &\geq -\sum_{i=1}^{\ell} \sigma_i(Z^{\rm T}B)\sigma_i(X^{\rm T}Z) \geq -\sum_{i=1}^{\ell} \sigma_i(Z^{\rm T}B) = \operatorname{trace}(\widetilde{Z}^{\rm T}B),\end{aligned}$$

and by (3.4)

$$\sum_{i=1}^{\ell} \lambda_i(E(X)) = \operatorname{trace}(\widetilde{Z}^{\rm T}E(X)\widetilde{Z}) = \operatorname{trace}(\widetilde{Z}^{\rm T}A\widetilde{Z}) + 2\operatorname{trace}(\widetilde{Z}^{\rm T}BX^{\rm T}\widetilde{Z})$$

[1] We remark that (3.2) is equivalent to

$$\min_{X\in\mathbb{O}^{n\times\ell}} \min_{Y\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(Y^{\rm T}E(X)Y) = \min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(X^{\rm T}E(X)X) = \min_{X\in\mathbb{O}^{n\times\ell}} f(X).$$

$$\begin{aligned}&\geq \operatorname{trace}(\widetilde{Z}^{\mathrm{T}} A \widetilde{Z}) + 2\operatorname{trace}(\widetilde{Z}^{\mathrm{T}} B)\\ &= f(\widetilde{Z}) \geq f(X_{\mathrm{opt}}).\end{aligned}$$

Since $X \in \mathbb{O}^{n\times\ell}$ is arbitrary, the above implies that

$$\min_{X\in\mathbb{O}^{n\times\ell}} \sum_{i=1}^{\ell} \lambda_i(E(X)) \geq f(X_{\mathrm{opt}}).$$

On the other hand, since

$$\min_{X\in\mathbb{O}^{n\times\ell}} \sum_{i=1}^{\ell} \lambda_i(E(X)) \leq \sum_{i=1}^{\ell} \lambda_i(E(X_{\mathrm{opt}})) \leq \operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}} E(X_{\mathrm{opt}}) X_{\mathrm{opt}}) = f(X_{\mathrm{opt}}),$$

we consequently have

$$\min_{X\in\mathbb{O}^{n\times\ell}} \sum_{i=1}^{\ell} \lambda_i(E(X)) = f(X_{\mathrm{opt}}).$$

This proves (3.2).

Furthermore, if X_{opt} is a global solution, then by

$$\begin{aligned}\sum_{i=1}^{\ell} \lambda_i(E(X_{\mathrm{opt}})) = f(X_{\mathrm{opt}}) &= \operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}} A X_{\mathrm{opt}}) + 2\operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}} B)\\ &= \operatorname{trace}(X_{\mathrm{opt}}^{\mathrm{T}} E(X_{\mathrm{opt}}) X_{\mathrm{opt}}),\end{aligned}$$

it follows that X_{opt} is an orthogonal eigenbasis associated with the first ℓ smallest eigenvalues of $E(X_{\mathrm{opt}})$, and thus (3.3) is true. ∎

As a counterpart of (3.3), for a local minimizer X, we can also ensure that X is an orthonormal eigenbasis matrix of $E(X)$.

THEOREM 3.4. *Any local minimizer X is an orthonormal eigenbasis matrix of $E(X)$ satisfying $E(X)X = X\Psi_X$ for a symmetric $\Psi_X \in \mathbb{R}^{\ell\times\ell}$. Moreover, for $\ell \geq 2$, it holds $\lambda_1(\Psi_X) \leq \lambda_\ell(E(X))$.*

Proof. According to Lemma 3.2, we have known that any local minimizer X satisfies $E(X)X = X\Psi_X$ and $\operatorname{trace}(\Psi_X) = f(X)$. Thus, all eigenvalues of Ψ_X are eigenvalues of $E(X)$. Let these eigenvalues be

$$\lambda_{\pi_1}(E(X)) \leq \cdots \leq \lambda_{\pi_\ell}(E(X)),$$

where $1 \leq \pi_1 \leq \cdots \leq \pi_\ell \leq n$, and $\lambda_i(\Psi_X) = \lambda_{\pi_i}(E(X))$ for $i = 1, \ldots, \ell$. To show the last part of the theorem, we assume by contradiction that $\lambda_1(\Psi_X) > \lambda_\ell(E(X))$. Let $J \in \mathbb{O}^{n\times\ell}$ be an orthonormal eigenbasis matrix of $E(X)$ corresponding to the first ℓ smallest eigenvalues; then

$$\begin{aligned}\operatorname{trace}(J^{\mathrm{T}} E(X) J) &= \sum_{i=1}^{\ell} \lambda_i(E(X))\\ &< \sum_{i=1}^{\ell} \lambda_i(\Psi_X) = \sum_{i=1}^{\ell} \lambda_{\pi_i}(E(X))\\ &= \operatorname{trace}(X^{\mathrm{T}} E(X) X)\end{aligned}$$

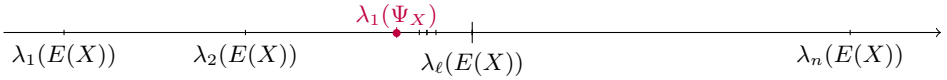
$$= \operatorname{trace}(X^{\mathrm{T}}AX) + 2\operatorname{trace}(X^{\mathrm{T}}B),$$

and moreover, $J^{\mathrm{T}}X = 0$. Now applying such a J to the second order necessary condition (2.4) and by $\operatorname{trace}(X^{\mathrm{T}}B) \le 0$, we have

$$\begin{aligned} 0 &\le \operatorname{trace}(J^{\mathrm{T}}\big[A + X\Lambda X^{\mathrm{T}} + (2I_n - XX^{\mathrm{T}})BX^{\mathrm{T}}\big]J) + \operatorname{trace}(\Lambda) \\ &= \operatorname{trace}(J^{\mathrm{T}}AJ) + \operatorname{trace}(\Lambda) \\ &< \operatorname{trace}(X^{\mathrm{T}}AX) + 2\operatorname{trace}(X^{\mathrm{T}}B) + \operatorname{trace}(\Lambda) \\ &= \operatorname{trace}(X^{\mathrm{T}}B) \le 0, \end{aligned}$$

a contradiction, and the proof is complete. ∎

Comparing with the global minimizer X_{opt} where the associated eigenvalues of $\Psi_{X_{\mathrm{opt}}}$ are just the first ℓ smallest eigenvalues of $E(X_{\mathrm{opt}})$, Theorem 3.4 guarantees that only the smallest eigenvalue $\lambda_1(\Psi_X)$ of Ψ_X must be one of the first ℓ smallest ones of $E(X)$. The following figure illustrates this necessary condition of the location of $\lambda_1(\Psi_X)$ for the local minimizer X.



3.2. An SCF iteration for the associated NEPv. Note that we have characterized the global minimizer X_{opt} of (1.2) as the solution to an eigenvector-dependent NEPv in (3.3). Such a system involves the coefficient matrix $E(X)$ which depends on the pursued eigenvectors X, instead of the eigenvalue in the traditional setting. As has been mentioned in section 1, there are an increasing number of cases and applications of such NEPvs arising from electronic structure calculations (e.g., [25, 27, 33, 43]) and data science (e.g., [4, 24, 40, 44, 45, 48, 46, 47]) recently. For our case, we next propose an SCF iteration for solving the NEPv (3.3). For the kth iteration, the algorithm first finds an arbitrarily orthogonal eigenbasis $\widehat{X}_k$ for the current matrix $E(X_{k-1})$, and refines $\widehat{X}_k$ to obtain a better orthonormal eigenbasis matrix. The following is a technical lemma for this refinement step.

LEMMA 3.5. *For any given $Z \in \mathbb{O}^{n\times\ell}$, we have*

$$\min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(Z^{\mathrm{T}}E(X)Z) = f(-ZP),$$

where the solution is $X = -ZP \in \mathbb{O}^{n\times\ell}$ with P being the orthogonal polar factor of $Z^{\mathrm{T}}B$.

Proof. For the given Z, according to Lemma 3.1, we know that

$$\operatorname{trace}(Z^{\mathrm{T}}BX^{\mathrm{T}}Z) \ge -\sum_{i=1}^{\ell}\sigma_i(Z^{\mathrm{T}}B)\sigma_i(X^{\mathrm{T}}Z) \ge -\sum_{i=1}^{\ell}\sigma_i(Z^{\mathrm{T}}B),$$

where the minimum can be achieved when $X = -ZP$ with P being the orthogonal polar factor of $Z^{\mathrm{T}}B$. Thus, we know by (3.4) that

$$\begin{aligned} \min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(Z^{\mathrm{T}}E(X)Z) &= \operatorname{trace}(Z^{\mathrm{T}}AZ) + 2\operatorname{trace}(-Z^{\mathrm{T}}BP^{\mathrm{T}}) \\ &= \operatorname{trace}(P^{\mathrm{T}}Z^{\mathrm{T}}AZP) + 2\operatorname{trace}((-ZP)^{\mathrm{T}}B) = f(-ZP), \end{aligned}$$

and the proof is complete. ∎

Algorithm 3.1 An SCF iteration for the NEPv (3.2).

Given a symmetric $A \in \mathbb{R}^{n\times n}$, a matrix $B \in \mathbb{R}^{n\times \ell}$, and $X_0 \in \mathbb{O}^{n\times \ell}$, the iteration solves (3.2).

1: **for** $k = 1, \ldots,$ until convergence **do**
2: Find an orthonormal eigenbasis matrix $\widehat{X}_k \in \mathbb{O}^{n\times \ell}$ associated with the ℓ smallest eigenvalues of $E(X_{k-1})$, and we have

$$E(X_{k-1})\widehat{X}_k = \widehat{X}_k\widehat{\Psi}_k. \tag{3.5}$$

3: Compute orthogonal polar factor P of $\widehat{X}_k^{\mathrm{T}}B$ and set $X_k = -\widehat{X}_kP$.
4: **end for**

We now present the simplest form of our SCF iteration for the NEPv (3.2) in Algorithm 3.1.

For Algorithm 3.1, we remark that our SCF iteration can be understood as an alternating scheme. In fact, for the current X_{k-1}, in line 2 of Algorithm 3.1, we fix the variable $X = X_{k-1}$ inside $E(X)$ to solve

$$\min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(X^{\mathrm{T}}E(X_{k-1})X)$$

to have $\widehat{X}_k$; the next update is to fix the $X = \widehat{X}_k$ outside and solve

$$\min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(\widehat{X}_k^{\mathrm{T}}E(X)\widehat{X}_k), \tag{3.6}$$

whose solution by Lemma 3.5 is $X_k = -\widehat{X}_kP$, where P is the orthogonal polar factor of $\widehat{X}_k^{\mathrm{T}}B$. This is just the update in line 3 of Algorithm 3.1, and we therefore have

$$\min_{X\in\mathbb{O}^{n\times\ell}} \operatorname{trace}(\widehat{X}_k^{\mathrm{T}}E(X)\widehat{X}_k) = f(X_k). \tag{3.7}$$

Moreover, it should be noted that X_k is also an orthonormal eigenbasis matrix of $E(X_{k-1})$ satisfying

$$E(X_{k-1})X_k = X_k\Psi_k, \quad \text{where } \Psi_k = P^{\mathrm{T}}\widehat{\Psi}_kP.$$

It is this additional step 3 of Algorithm 3.1 that separates our SCF from the usual SCF iteration [7]. The reason is that the objective function $f(X)$ and the coefficient matrix $E(X)$ are not invariant under orthogonal transformation, i.e., $f(X) \neq f(XP)$ for $P \in \mathbb{O}^{\ell\times\ell}$. Thus, when we have an orthonormal eigenbasis matrix $\widehat{X}_k$ of $E(X_{k-1})$, it is still possible to improve the objective function by refining $\widehat{X}_k$ to a new eigenbasis matrix $X_k = -\widehat{X}_kP$. Indeed, the associated P satisfies

$$P \in \operatorname*{argmin}_{Q\in\mathbb{O}^{\ell\times\ell}} f(\widehat{X}_kQ).$$

It is interesting to note that even if $\widehat{X}_k$ in line 2 of Algorithm 3.1 can be arbitrary, the next iteration X_k in line 3 is uniquely determined under certain conditions.

PROPOSITION 3.6. *For Algorithm* 3.1*, suppose the eigenvalue gap*

$$\eta_{k-1} = \lambda_{\ell+1}(E(X_{k-1})) - \lambda_\ell(E(X_{k-1})) > 0; \tag{3.8}$$

then any two orthogonal eigenbases $\widehat{X}_k$ *and* $\widetilde{X}_k$ *associated with the* ℓ *smallest eigenvalues of* $E(X_{k-1})$ *satisfy* $\widetilde{X}_k = \widehat{X}_k Q$ *for some orthogonal matrix* $Q \in \mathbb{R}^{\ell\times\ell}$. *Furthermore, if additionally* $\operatorname{rank}(B^{\mathrm{T}}\widehat{X}_k) = \ell$, *then* X_k *is uniquely determined.*

Proof. If the eigenvalue gap $\eta_{k-1} > 0$, then the eigenspace associated with the ℓ smallest eigenvalues of $E(X_{k-1})$ is unique (see, e.g., [36, p. 244]), and the first part follows. If $\operatorname{rank}(B^{\mathrm{T}}\widehat{X}_k) = \ell$, we know that the next iteration in line 3 is just $X_k = -\widehat{X}_k P$, where P is the orthogonal polar factor of $\widehat{X}_k^{\mathrm{T}} B$ which is $\widehat{X}_k^{\mathrm{T}} B(B^{\mathrm{T}}\widehat{X}_k\widehat{X}_k^{\mathrm{T}}B)^{-\frac{1}{2}}$. Now since for any other choice $\widetilde{X}_k$ we have $\widetilde{X}_k = \widehat{X}_k Q$, it holds that

$$X_k = \widehat{X}_k\widehat{X}_k^{\mathrm{T}}B(B^{\mathrm{T}}\widehat{X}_k\widehat{X}_k^{\mathrm{T}}B)^{-\frac{1}{2}} = \widetilde{X}_k\widetilde{X}_k^{\mathrm{T}}B(B^{\mathrm{T}}\widetilde{X}_k\widetilde{X}_k^{\mathrm{T}}B)^{-\frac{1}{2}},$$

implying the uniqueness of X_k. ∎

4. Convergence analysis. We next discuss some convergence behaviors of the basic SCF iteration (Algorithm 3.1). The following proposition summarizes some basic facts about the sequence $\{X_k\}_{k=0}^{\infty}$ from Algorithm 3.1.

PROPOSITION 4.1. *Let* $\{X_k\}_{k=0}^{\infty}$ *be the sequence generated by Algorithm* 3.1*; then the following hold:*

(i) *For each* $k \geq 1$, $-B^{\mathrm{T}}X_k = -X_k^{\mathrm{T}}B \succcurlyeq 0$ *and* $\operatorname{trace}(X_k^{\mathrm{T}}B) = -\sum_{i=1}^{\ell}\sigma_i(X_k^{\mathrm{T}}B)$.

(ii) *The sequence* $\{f(X_k)\}_{k=0}^{\infty}$ *is nonincreasing and convergent.*

(iii) *The sum of the* ℓ *smallest eigenvalues of* $E(X_k)$ *is nonincreasing and convergent; that is,*

$$\sum_{i=1}^{\ell}\lambda_i\big(E(X_{k-1})\big) \geq \sum_{i=1}^{\ell}\lambda_i\big(E(X_k)\big) \to \nu \quad \text{as } k \to \infty. \tag{4.1}$$

Proof. The term (i) is based on our design of SCF iteration in line 3 of Algorithm 3.1, which leads to $X_k^{\mathrm{T}}B = -V\Sigma V^{\mathrm{T}} \preccurlyeq 0$, where $U\Sigma V^{\mathrm{T}}$ is the SVD of $\widehat{X}_k^{\mathrm{T}}B$. Since $X_k = -\widehat{X}_k UV^{\mathrm{T}}$, the singular values of $\widehat{X}_k^{\mathrm{T}}B^{\mathrm{T}}$ and $X_k^{\mathrm{T}}B$ are the same, and (i) is true.

For (ii), since $\widehat{X}_{k-1}$ is the orthogonal eigenbasis of $E(X_{k-1})$ and X_k is the solution to (3.6), it holds that

$$\begin{aligned} f(X_{k-1}) &\geq \operatorname{trace}(\widehat{X}_k^{\mathrm{T}}E(X_{k-1})\widehat{X}_k) \geq \operatorname{trace}(\widehat{X}_k^{\mathrm{T}}E(X_k)\widehat{X}_k) \\ &= \operatorname{trace}(X_k^{\mathrm{T}}E(X_k)X_k) = f(X_k), \end{aligned} \tag{4.2}$$

i.e., $\{f(X_k)\}_{k=0}^{\infty}$ is nonincreasing. The convergence of $\{f(X_k)\}_{k=0}^{\infty}$ is then evident by the boundedness over $\mathbb{O}^{n\times\ell}$.

For (iii), from (3.5) and (4.2), we have

$$\sum_{i=1}^{\ell}\lambda_i\big(E(X_{k-1})\big) = \operatorname{trace}(\Psi_k) = \operatorname{trace}(X_k^{\mathrm{T}}E(X_{k-1})X_k) \geq f(X_k),$$

and also

$$\sum_{i=1}^{\ell}\lambda_i\big(E(X_k)\big) = \operatorname{trace}(X_{k+1}^{\mathrm{T}}E(X_k)X_{k+1}) \leq \operatorname{trace}(X_k^{\mathrm{T}}E(X_k)X_k) = f(X_k).$$

This shows

$$\sum_{i=1}^{\ell}\lambda_i\big(E(X_{k-1})\big) \geq \sum_{i=1}^{\ell}\lambda_i\big(E(X_k)\big),$$

and by the boundedness of $\mathrm{trace}(X^{\mathrm{T}}E(X)X)$ over $\mathbb{O}^{n\times\ell}$, the claim (iii) is then established. ∎

For the convergence of Algorithm 3.1, we next introduce the metric between two subspaces. Given two $X, Y \in \mathbb{O}^{n\times\ell}$, we define the distance between two subspaces $\mathcal{R}(X)$ and $\mathcal{R}(Y)$ of dimension ℓ by

$$\|\sin\Theta(\mathcal{R}(X),\mathcal{R}(Y))\|_2 = \|\mathcal{P}_X - \mathcal{P}_Y\|_2, \tag{4.3}$$

where $\mathcal{P}_X$ denotes the orthogonal projection onto $\mathcal{R}(X)$, and

$$\Theta(\mathcal{R}(X),\mathcal{R}(Y)) = \mathrm{diag}(\theta_1(\mathcal{R}(X),\mathcal{R}(Y)),\ldots,\theta_\ell(\mathcal{R}(X),\mathcal{R}(Y)))$$

stand for the canonical angles between $\mathcal{R}(X)$ and $\mathcal{R}(Y)$, which according to the singular values $\sigma_i(X^{\mathrm{T}}Y)$ are defined by

$$0 \le \theta_i(\mathcal{R}(X),\mathcal{R}(Y)) := \arccos\sigma_i(X^{\mathrm{T}}Y) \le \frac{\pi}{2} \quad \text{for } 1 \le i \le \ell.$$

We remark that $\|\sin\Theta(\mathcal{R}(X),\mathcal{R}(Y))\|_2$ in (4.3) is a particular unitarily invariant metric [32] in the Grassmann manifold $\mathrm{Grass}(n,\ell)$, in which each point $\lfloor X\rfloor \in \mathrm{Grass}(n,\ell)$ can be represented by any basis X of $\mathcal{R}(X)$. In particular, if $X, Y \in \mathbb{O}^{n\times\ell}$, we have $\mathcal{P}_X = XX^{\mathrm{T}}$ and $\mathcal{P}_Y = YY^{\mathrm{T}}$, and thus $\|\sin\Theta(\mathcal{R}(X),\mathcal{R}(Y))\|_2 = \|XX^{\mathrm{T}} - YY^{\mathrm{T}}\|_2$. For simplicity purposes, we denote

$$\|\lfloor X\rfloor - \lfloor Y\rfloor\|_2 := \|\sin\Theta(\mathcal{R}(X),\mathcal{R}(Y))\|_2. \tag{4.4}$$

The following lemma is a modification of the one [29, Lemma 4.10] (see also [21, Proposition 7]) in which the standard Euclidean distance is replaced by (4.4).

LEMMA 4.2. *Assume that $\lfloor W\rfloor \in \mathrm{Grass}(n,\ell)$ is an isolated accumulation point (in the metric (4.4)) of a sequence $\{\lfloor W_k\rfloor\}_{k=0}^{\infty} \subseteq \mathrm{Grass}(n,\ell)$ such that, for every subsequence $\{\lfloor W_k\rfloor\}_{k\in\mathcal{K}}$ converging to $\lfloor W\rfloor$, there is an infinite subset $\widehat{\mathcal{K}} \subseteq \mathcal{K}$ such that $\{\|\lfloor W_{k+1}\rfloor - \lfloor W_k\rfloor\|_2\}_{k\in\widehat{\mathcal{K}}} \to 0$. Then the whole sequence $\{\lfloor W_k\rfloor\}_{k=0}^{\infty}$ converges to $\lfloor W\rfloor$.*

We remark that Lemma 4.2 is a sufficient condition for proving the convergence of $\{\lfloor W_k\rfloor\}_{k=0}^{\infty}$ to $\lfloor W\rfloor$. The condition that $\lfloor W\rfloor \in \mathrm{Grass}(n,\ell)$ is an isolated accumulation point that does not exclude the existence of finitely many or even infinitely many accumulation points. Indeed, when we know that $\{\lfloor W_k\rfloor\}_{k=0}^{\infty}$ only contains finitely many accumulation points, then each is isolated, and the condition of Lemma 4.2 ensures that $\lfloor W_k\rfloor \to \lfloor W\rfloor$. Based on this lemma, we have the following convergence results for Algorithm 3.1.

THEOREM 4.3. *Let $\{X_k\}_{k=0}^{\infty}$ be the sequence generated by Algorithm 3.1. Let $\{X_k\}_{k\in\mathcal{K}}$ be any convergent subsequence with the accumulation point X satisfying*

$$\eta = \lambda_{\ell+1}(E(X)) - \lambda_\ell(E(X)) > 0. \tag{4.5}$$

Then

(i) *X satisfies the first order optimality in Lemma 2.1 and also the necessary condition (2.6) for a global minimizer;*

(ii) *if, additionally, $\lfloor X\rfloor$ is an isolated accumulation point of $\{\lfloor X_k\rfloor\}_{k=0}^{\infty}$, then $\{\lfloor X_k\rfloor\}_{k=0}^{\infty}$ converges to $\lfloor X\rfloor$;*

(iii) *under the conditions of (ii), and if $\mathrm{rank}(B^{\mathrm{T}}X) = \ell$, then $\{X_k\}_{k=0}^{\infty}$ converges to X.*

Proof. For (i), we can choose a subsequence $\{X_k\}_{k\in\widehat{\mathcal{K}}}$ of $\{X_k\}_{k\in\mathcal{K}}$ such that $\{X_{k-1}\}_{k\in\widehat{\mathcal{K}}}$ and $\{X_{k+1}\}_{k\in\widehat{\mathcal{K}}}$ converge to $Y \in \mathbb{O}^{n\times\ell}$ and $Z \in \mathbb{O}^{n\times\ell}$, respectively. Thus, from

$$E(X_{k-1})X_k = X_k\Psi_k \quad \text{and} \quad E(X_k)X_{k+1} = X_{k+1}\Psi_{k+1}, \quad k \in \widehat{\mathcal{K}},$$

we have

$$E(Y)X = X\Psi \quad \text{and} \quad E(X)Z = Z\widetilde{\Psi}. \tag{4.6}$$

Also, by (i) of Proposition 4.1, we know that $-H^{\mathrm{T}}B = -B^{\mathrm{T}}H \succcurlyeq 0$ for $H \in \{X, Y, Z\}$, and

$$\lim_{k\to\infty,\ k\in\widehat{\mathcal{K}}} \sum_{i=1}^{\ell} \lambda_i(E(X_{k-1})) = \operatorname{trace}(\Psi) = \lim_{k\to\infty,\ k\in\widehat{\mathcal{K}}} \sum_{i=1}^{\ell} \lambda_i(E(X_k)) = \operatorname{trace}(\widetilde{\Psi}) = \nu.$$

Note that $0 \le \sigma_i(X_{k-1}^{\mathrm{T}}X_k) \le 1$ and

$$\operatorname{trace}(X_k^{\mathrm{T}}BX_{k-1}^{\mathrm{T}}X_k) - \operatorname{trace}(X_k^{\mathrm{T}}B) \ge \sum_{i=1}^{\ell} \sigma_i(X_k^{\mathrm{T}}B)(1 - \sigma_i(X_{k-1}^{\mathrm{T}}X_k)) \ge 0. \tag{4.7}$$

Applying $k \to \infty,\ k \in \widehat{\mathcal{K}}$ to (4.7), it holds that $\operatorname{trace}(X^{\mathrm{T}}BY^{\mathrm{T}}X) \ge \operatorname{trace}(X^{\mathrm{T}}B)$. Thus,

$$\operatorname{trace}(X^{\mathrm{T}}E(Y)X) - \operatorname{trace}(X^{\mathrm{T}}E(X)X) = 2(\operatorname{trace}(X^{\mathrm{T}}BY^{\mathrm{T}}X) - \operatorname{trace}(X^{\mathrm{T}}B)) \ge 0. \tag{4.8}$$

Moreover, using the fact that Z is an orthogonal eigenbasis of $E(X)$ associated with the first ℓ smallest eigenvalues and (4.6), we conclude that

$$\begin{aligned} \operatorname{trace}(X^{\mathrm{T}}E(Y)X) - \operatorname{trace}(X^{\mathrm{T}}E(X)X) &\le \operatorname{trace}(X^{\mathrm{T}}E(Y)X) - \operatorname{trace}(Z^{\mathrm{T}}E(X)Z) \\ &= \operatorname{trace}(\Psi) - \operatorname{trace}(\widetilde{\Psi}) = 0, \end{aligned} \tag{4.9}$$

which, combined with (4.8), yields $\operatorname{trace}(X^{\mathrm{T}}BY^{\mathrm{T}}X) = \operatorname{trace}(X^{\mathrm{T}}B)$. Therefore, using (4.6) and

$$\operatorname{trace}(X^{\mathrm{T}}BY^{\mathrm{T}}X) = \operatorname{trace}(X^{\mathrm{T}}B),$$

we get

$$\begin{aligned} \nu = \operatorname{trace}(\widetilde{\Psi}) &\le \operatorname{trace}(X^{\mathrm{T}}E(X)X) \\ &= \operatorname{trace}(X^{\mathrm{T}}AX) + 2\operatorname{trace}(X^{\mathrm{T}}B) \\ &= \operatorname{trace}(X^{\mathrm{T}}AX) + 2\operatorname{trace}(X^{\mathrm{T}}BY^{\mathrm{T}}X) \\ &= \operatorname{trace}(X^{\mathrm{T}}E(Y)X) = \operatorname{trace}(\Psi) = \nu. \end{aligned}$$

From $\operatorname{trace}(\widetilde{\Psi}) = \operatorname{trace}(\Psi)$ and $E(X)Z = Z\widetilde{\Psi}$, we know that X is also an orthogonal eigenbasis associated with the first ℓ smallest eigenvalues of $E(X)$. The eigenvalue gap condition (4.5) ensures that $\lfloor X \rfloor = \lfloor Z \rfloor$ (see, e.g., [36, p. 244]). In other words, $Z = XQ$ for an orthonormal $Q \in \mathbb{R}^{\ell\times\ell}$. Thus, the condition $E(X)Z = E(X)XQ = XQ\widetilde{\Psi}$ yields $E(X)X = E(X)X = XQ\widetilde{\Psi}Q^{\mathrm{T}}$, or equivalently,

$$AX + X\Lambda = -B, \quad \text{with} \quad \Lambda = B^{\mathrm{T}}X - Q\widetilde{\Psi}Q^{\mathrm{T}}.$$

This proves (i).

For (ii), as we have shown, there is a subsequence $\widehat{\mathcal{K}} \subseteq \mathcal{K}$ such that $\{\|\lfloor X_{k+1}\rfloor - \lfloor X_k\rfloor\|_2\}_{k\in\widehat{\mathcal{K}}} \to 0$ as $k \to \infty$. According to Lemma 4.2, the condition that $\lfloor X\rfloor$ is an isolated accumulation point of $\{\lfloor X_k\rfloor\}_{k=0}^{\infty}$ leads to (ii).

For (iii), according to (ii), the limit of any convergent subsequence of $\{X_k\}_{k=0}^{\infty}$ can be expressed as XQ for some $Q \in \mathbb{O}^{\ell\times\ell}$. Moreover, Proposition 4.1(ii) implies that $f(X) = f(XQ)$, leading to $\mathrm{trace}(B^{\mathrm{T}}XQ) = \mathrm{trace}(B^{\mathrm{T}}X)$. Note also that Proposition 4.1(i) implies that $B^{\mathrm{T}}X$ is symmetric and $\mathrm{trace}(B^{\mathrm{T}}X) = -\sum_{i=1}^{\ell}\sigma_i(B^{\mathrm{T}}X)$. Let the eigenvalue decomposition of $B^{\mathrm{T}}X$ be $-V\Sigma V^{\mathrm{T}}$ with $\Sigma \succcurlyeq 0$ and $\widehat{Q} = V^{\mathrm{T}}QV \in \mathbb{O}^{\ell\times\ell}$. Thus, we have

$$-\sum_{i=1}^{\ell}\sigma_i(B^{\mathrm{T}}X) = \mathrm{trace}(B^{\mathrm{T}}X) = \mathrm{trace}(B^{\mathrm{T}}XQ) = \mathrm{trace}(-\Sigma\widehat{Q}) = -\sum_{i=1}^{\ell}\sigma_i(B^{\mathrm{T}}X)\widehat{Q}_{(i,i)}$$

implying $\sigma_i(B^{\mathrm{T}}X)(1-\widehat{Q}_{(i,i)}) = 0\ \forall i = 1, 2, \ldots, \ell$, where $\widehat{Q}_{(i,i)}$ are the diagonal entries of $\widehat{Q}$. Thus, if $\mathrm{rank}(B^{\mathrm{T}}X) = \ell$, we have $\widehat{Q}_{(i,i)} = 1\ \forall i = 1, \ldots, \ell$, which implies $\widehat{Q} = Q = I_\ell$. This proves (iii). ∎

We next discuss the local convergence behavior for the SCF iteration. As Theorem 4.3(iii) implies that the sequence $\{X_k\}$ converges to a solution X, we will establish an estimate of the local convergence rate, which reflects the factors for the speed of the SCF iteration. Our convergence analysis for this part basically follows from the general local convergence analysis [7] for the SCF iteration in solving the standard eigenvector-dependent NEPv. The following lemma [45, Lemma 3.5] is obvious.

LEMMA 4.4. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be of unit norm and $\|\boldsymbol{x} - \boldsymbol{y}\|_2 < 1$. Then*

$$\frac{\sqrt{2}}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2 \le \sin\theta(\boldsymbol{x}, \boldsymbol{y}) = \|\lfloor\boldsymbol{x}\rfloor - \lfloor\boldsymbol{y}\rfloor\|_2 \le \|\boldsymbol{x} - \boldsymbol{y}\|_2, \tag{4.10}$$

where $\sin\theta(\boldsymbol{x}, \boldsymbol{y})$ is the sine of angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

THEOREM 4.5. *Under the assumptions of Theorem 4.3(iii), then $\{X_k\}$ converges to X, and moreover, for sufficiently large k, we have*

$$\|\lfloor X\rfloor - \lfloor X_{k+1}\rfloor\|_2 \le \frac{3\|B\|_2}{\eta}\|X_k - X\|_2, \tag{4.11}$$

where η is given by (4.5). If $\ell = 1$, then

$$\|\lfloor X\rfloor - \lfloor X_{k+1}\rfloor\|_2 \le \frac{3\sqrt{2}\|B\|_2}{\eta}\|\lfloor X\rfloor - \lfloor X_k\rfloor\|_2, \tag{4.12}$$

$$\|X - X_{k+1}\|_2 \le \frac{3\sqrt{2}\|B\|_2}{\eta}\|X - X_k\|_2. \tag{4.13}$$

Proof. For k, we denote $Z_k = X_k - X$ and $\Delta_k = BZ_k^{\mathrm{T}} + Z_kB^{\mathrm{T}}$, and thus, $E(X_k) = E(X) + \Delta_k$. Also, let

$$X_{k+1} = XC + X_\perp S, \quad C \in \mathbb{R}^{\ell\times\ell}, \quad S \in \mathbb{R}^{(n-\ell)\times(n-\ell)}. \tag{4.14}$$

It is then known that for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$, it follows that

$$\|\sin\Theta(\mathcal{R}(X_{k+1}, \mathcal{R}(X)))\|_{\mathrm{ui}} = \|S\|_{\mathrm{ui}}. \tag{4.15}$$

By premultiplying $X_\perp^{\mathrm{T}}$ on both sides of $E(X_k)X_{k+1} = (E(X)+\Delta_k)X_{k+1} = X_{k+1}\Psi_{k+1}$, we have

$$X_\perp^{\mathrm{T}}E(X)X_{k+1} - X_\perp^{\mathrm{T}}X_{k+1}\Psi_{k+1} = \Psi_\perp S - S\Psi_{k+1} = -X_\perp^{\mathrm{T}}\Delta_k X_{k+1}, \tag{4.16}$$

where $\Psi_\perp = X_\perp^{\mathrm{T}}E(X)X_\perp \in \mathbb{R}^{(n-\ell)\times(n-\ell)}$ whose eigenvalues are $\mathrm{eig}(E(X))$ with $\mathrm{eig}(\Psi) = \mathrm{eig}(X^{\mathrm{T}}E(X)X)$ excluded. Note that the solution S in the Sylvester equation (4.16) satisfies (see [9])

$$\|S\|_{\mathrm{ui}} = \|\sin\Theta(\mathcal{R}(X_{k+1},\mathcal{R}(X)))\|_{\mathrm{ui}} \le \frac{\|X_\perp^{\mathrm{T}}\Delta_k X_{k+1}\|_{\mathrm{ui}}}{\eta_k} \le \frac{3\|X_\perp^{\mathrm{T}}\Delta_k X_{k+1}\|_{\mathrm{ui}}}{2\eta}, \tag{4.17}$$

where $\eta_k = \lambda_{\ell+1}(E(X_k)) - \lambda_\ell(E(X_k))$ which can be assumed to be positive and $\eta_k \ge 2\eta/3$ for sufficiently large k. Therefore, (4.11) follows by $\|X_\perp^{\mathrm{T}}\Delta_k X_{k+1}\|_2 \le 2\|B\|_2\|Z_k\|_2$, and (4.12)–(4.13) follow by (4.10). ∎

Remark 4.6. As an implication of Theorem 4.5, we know that the convergence of SCF is fast whenever $\|B\|_2$ is small and the eigenvalue gap η in (4.5) is large. Indeed, the result (4.11) reflects the nature of the problem (1.2): When B is small in the norm (relative to that of A), then (1.2) is much like an eigenvalue problem of finding the eigenpairs associated with the ℓ smallest eigenvalues of A, while when $\|B\|$ gets bigger (relative to that of A), the problem (1.2) gets closer to maximizing $\mathrm{trace}(B^{\mathrm{T}}X)$ over $\mathbb{O}^{n\times\ell}$, whose solution is just the orthogonal polar factor of B. This useful observation will be used in our implementation for the eigensolver in line 2 of Algorithm 3.1; furthermore, we will illustrate the convergence behavior with a different norm $\|B\|_2$ in our numerical testing.

5. Efficient implementation of the SCF iteration.

In this section, we provide several techniques to make the simple SCF iteration more efficient.

5.1. The inexact SCF iteration.

Note that the main computational effort of the pure SCF iteration lies in computing an orthonormal eigenbasis matrix of $E(X_{k-1})$ associated with the ℓ smallest eigenvalues (line 2 in Algorithm 3.1); we will discuss efficient iterative eigensolvers for this task, instead of resorting the full eigendecomposition of $E(X_{k-1})$. The following proposition first justifies the monotonic convergence of the SCF iteration even with an approximate orthonormal eigenbasis matrix $\widehat{X}_k$ of $E(X_{k-1})$ in line 2 in Algorithm 3.1.

PROPOSITION 5.1. *For the $(k-1)$th iterate $X_{k-1} \in \mathbb{O}^{n\times\ell}$ in Algorithm* 3.1*, let $\widehat{X}_k \in \mathbb{O}^{n\times\ell}$ be an approximate orthonormal eigenbasis matrix associated with the first ℓ smallest eigenvalues of $E(X_{k-1})$ satisfying*

$$\mathrm{trace}(\widehat{X}_k^{\mathrm{T}}E(X_{k-1})\widehat{X}_k) \le \mathrm{trace}(X_{k-1}^{\mathrm{T}}E(X_{k-1})X_{k-1}) = f(X_{k-1}); \tag{5.1}$$

then the next iterate X_k of the SCF iteration Algorithm 3.1 *satisfies $f(X_k) \le f(X_{k-1})$.*

Proof. Recall from Algorithm 3.1 that $X_k = -\widehat{X}_k P$, where P is the orthogonal polar factor of $\widehat{X}_k^{\mathrm{T}}B$. By Lemma 3.5, X_k minimizes $\mathrm{trace}(\widehat{X}_k^{\mathrm{T}}E(X)\widehat{X}_k)$ over all $X \in \mathbb{O}^{n\times\ell}$. Thus,

$$\begin{aligned} f(X_k) &= \mathrm{trace}(X_k^{\mathrm{T}}E(X_k)X_k) = \mathrm{trace}(\widehat{X}_k^{\mathrm{T}}E(X_k)\widehat{X}_k) \\ &\le \mathrm{trace}(\widehat{X}_k^{\mathrm{T}}E(X_{k-1})\widehat{X}_k) \le f(X_{k-1}), \end{aligned} \tag{5.2}$$

and the proof is complete. ∎

We remark, by (5.2), that as long as the inequality (5.1) is strict, we have $f(X_k) < f(X_{k-1})$. If X_{k-1} is not the solution of the NEPv (3.3), then (5.1) holds strictly as long as the approximate $\widehat{X}_k$ is better than X_{k-1} in the sense of Rayleigh quotient minimization. The latter could be generically true for any efficient eigensolver. Our choice here is the locally optimal block preconditioned conjugate gradient (LOBPCG) method[2] (see [22, 23]). It is known that the $(j+1)$th iterate $\widehat{X}_k^{(j+1)}$ of LOBPCG for an approximate eigenbasis corresponding to the first ℓ smallest eigenvalues of $E(X_{k-1})$ is obtained by solving

$$\widehat{X}_k^{(j+1)} := \arg \min_{X \in \mathcal{R}\left([\widehat{X}_k^{(j-1)},\ \widehat{X}_k^{(j)},\ NR^{(j)}]\right),\ X \in \mathbb{O}^{n\times \ell}} \operatorname{trace}(X^{\mathrm{T}} E(X_{k-1}) X), \tag{5.3}$$

where $N \in \mathbb{R}^{n\times n}$ is a symmetric preconditioner for the matrix $E(X_{k-1})$, and $R^{(j)} = E(X_{k-1})\widehat{X}_k^{(j)} - \widehat{X}_k^{(j)}\Theta_j$ is the residual associated with the Ritz eigenpair $(\widehat{X}_k^{(j)}, \Theta_j)$. Thus, if we choose the initial $\widehat{X}_k^{(0)} = X_{k-1}$, by (5.3), for the approximate $\widehat{X}_k = \widehat{X}_k^{(j+1)}$ we have

$$\begin{aligned}\operatorname{trace}(\widehat{X}_k^{\mathrm{T}} E(X_{k-1}) \widehat{X}_k) &= \operatorname{trace}((\widehat{X}_k^{(j+1)})^{\mathrm{T}} E(X_{k-1}) \widehat{X}_k^{(j+1)}) \\ &\le \operatorname{trace}((\widehat{X}_k^{(0)})^{\mathrm{T}} E(X_{k-1}) \widehat{X}_k^{(0)}) \\ &= \operatorname{trace}(X_{k-1}^{\mathrm{T}} E(X_{k-1}) X_{k-1}) = f(X_{k-1}),\end{aligned}$$

validating the condition (5.1).

The termination of the LOBPCG for the eigenpairs of $E(X_{k-1})$ is controlled by the residual $\|R^{(j)}\| \le$ tol_lobpcg and the maximal number maxit_lobpcg of iterations. By Remark 4.6, we know that when $\|B\|$ is small relative to $\|A\|$ (implying that the problem is close to an eigenvalue problem associated with A), we attempt to choose a relatively tight stopping criterion or use a preconditioner for LOBPCG. Detailed parameters of tol_lobpcg and maxit_lobpcg will be specified in section 6.

5.2. Inexact SCF iteration with subspace acceleration. In the traditional eigenvalue problem computation, the Rayleigh–Ritz (RR) procedure [11, 16] is one of most widely used and efficient approaches. It serves as a basic framework for most state-of-the-art eigensolvers for large scale and sparse eigenvalue problems. In general, the RR procedure seeks a good subspace (usually the Krylov subspace) containing sufficient information of desired eigenpairs, onto which the original eigenproblem is projected to compute the Ritz pairs as approximate eigenpairs of the original problem.

For the NEPv (3.3), we observe that the iterates $\{X_k\}$ gradually converge to X_{opt} and naturally collect increasingly useful information of X_{opt}. Suppose $\mathcal{S}$ is a subspace which can be formed by the previous t iterates $X_{k-t}, \ldots, X_{k-1}$ at the current $(k-1)$th iteration. In the framework of RR, we attempt to refine the current X_{k-1} by seeking a better approximate $\breve{X}_{k-1}$ than X_{k-1} in the next SCF iteration: $E(\breve{X}_{k-1})X_k = X_k\Lambda_k$. In particular, we let Q_k be the orthogonal basis of $\mathcal{S}$, and we choose to find $\breve{X}_{k-1} \in \mathcal{R}(Q_k)$. Assume $\operatorname{rank}(\mathcal{S}) = r_k$, i.e., $Q_k \in \mathbb{O}^{n\times r_k}$, and thus, $\breve{X}_{k-1} = Q_k Y_k$ for some $Y_k \in \mathbb{O}^{r_k\times \ell}$. In the sprit of RR, the parameter matrix Y_k can be naturally obtained via solving

$$\min_{\breve{X}\in\mathcal{R}(Q_k)} f(\breve{X}) \Longleftrightarrow \min_{Y\in\mathbb{O}^{r_k\times\ell}} \operatorname{trace}(Y^{\mathrm{T}} A_k Y) + 2\operatorname{trace}(B_k^{\mathrm{T}} Y), \tag{5.4}$$

[2]LOBPCG (in MATLAB) is available at http://cn.mathworks.com/matlabcentral/fileexchange/48-lobpcg-m.

where $A_k = Q_k^{\mathrm{T}} A Q_k \in \mathbb{R}^{r_k \times r_k}$ and $B_k = Q_k^{\mathrm{T}} B \in \mathbb{R}^{r_k \times \ell}$. It can be seen by Theorem 3.3 that (5.4) is equivalent to

$$\min_{Y \in \mathbb{O}^{r_k \times \ell}} \sum_{i=1}^{\ell} \lambda_i (A_k + B_k Y^{\mathrm{T}} + Y B_k^{\mathrm{T}}),$$

and the solution Y_k is an orthogonal eigenbasis of $A_k + B_k Y^{\mathrm{T}} + Y B_k^{\mathrm{T}}$ associated with the first ℓ smallest eigenvalues.

Notice that (5.4) is again a minimization of the quadratic function over the Stiefel manifold, but the dimension r_k is generally small. The Riemannian Trust-Region (RTR) method[3] in [1, 2] is suitable for such a task, where we mainly need to specify the gradient and Hessian operator of the objective function $f(X)$ restricted on $\mathbb{O}^{n\times\ell}$ for its implementation. By calculation, we have the following proposition.

PROPOSITION 5.2. *Let $f_{|\mathbb{O}^{n\times\ell}(X)}$ be the restriction of the objective function $f(X)$ of* (1.2) *on the Stiefel manifold $\mathbb{O}^{n\times\ell}$. Then for any $X \in \mathbb{O}^{n\times\ell}$, the gradient $g(X)$ of $f_{|\mathbb{O}^{n\times\ell}(X)}$ is $g(X) = AX + X\Lambda + B$, where $\Lambda = -X^{\mathrm{T}}AX - \frac{X^{\mathrm{T}}B + B^{\mathrm{T}}X}{2}$. Moreover, the Hessian operation $h(X) : \mathcal{T}_X\mathbb{O}^{n\times\ell} \to \mathcal{T}_X\mathbb{O}^{n\times\ell}$ of $f_{|\mathbb{O}^{n\times\ell}(X)}$ at $X \in \mathbb{O}^{n\times\ell}$ acting on an arbitrary tangent vector $H \in \mathcal{T}_X\mathbb{O}^{n\times\ell}$ is given by*

$$h(X)[H] = \Pi_X(AH + H\Lambda),$$

where $\Pi_X(\bullet)$ is the projection defined by (2.3).

Similar to the monotonic property in Proposition 5.1, we next show that the solution $\breve{X}_{k-1}$ to (5.4) obtained from RTR can be computed inexactly.

PROPOSITION 5.3. *For $t \geq 1$, suppose the approximate solution $\breve{X}_{k-1}$ to* (5.4) *obtained from RTR satisfies $f(\breve{X}_{k-1}) \leq f(X_{k-1})$ and $\widehat{X}_k \in \mathbb{O}^{n\times\ell}$ is an approximate orthonormal eigenbasis matrix associated with the ℓ smallest eigenvalues of $E(\breve{X}_{k-1})$ satisfying*

$$\mathrm{trace}(\widehat{X}_k^{\mathrm{T}} E(\breve{X}_{k-1}) \widehat{X}_k) \leq \mathrm{trace}(\breve{X}_{k-1}^{\mathrm{T}} E(\breve{X}_{k-1}) \breve{X}_{k-1}) = f(\breve{X}_{k-1}); \tag{5.5}$$

then the next iterate $X_k = -\widehat{X}_k P$ satisfies $f(X_k) \leq f(X_{k-1})$, where P is the orthogonal polar factor of $\widehat{X}_k^{\mathrm{T}} B$.

Proof. We note by assumptions and Lemma 3.5 that

$$\begin{aligned} f(X_{k-1}) \geq f(\breve{X}_{k-1}) &\geq \mathrm{trace}(\widehat{X}_k^{\mathrm{T}} E(\breve{X}_{k-1}) \widehat{X}_k) \geq \min_{X \in \mathbb{O}^{n\times\ell}} \mathrm{trace}(\widehat{X}_k^{\mathrm{T}} E(X) \widehat{X}_k) \\ &= f(-\widehat{X}_k P) = f(X_k), \end{aligned}$$

and the assertion follows. ∎

To efficiently apply this subspace acceleration strategy, in our numerical testing, we restart the subspace acceleration by introducing the maximal number of blocks, say ς_m, of the subspace $\mathcal{S} = \mathcal{R}([X_{k-t}, \ldots, X_{k-1}])$; that is, we require $t \leq \varsigma_m$, and whenever $t > \varsigma_m$, we discard the first block by setting $\mathcal{S}(:, 1 : \ell) = [\ ]$ and then restart expanding $\mathcal{S}$ (in our numerical testing, we set $\varsigma_m = 5$). Incorporating such subspace acceleration strategy, we arrive at the following practical SCF iteration, named SCFRTR (Algorithm 5.1).

[3] The MATLAB solver of the generic RTR method is available at https://www.manopt.org.

Algorithm 5.1 SCFRTR: The SCF iteration with subspace acceleration for (1.2).

Given a symmetric $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times \ell}$, and $X_0 \in \mathbb{O}^{n\times \ell}$, SCFRTR returns an approximate solution of (1.2).

1: Set $\mathcal{S} = \emptyset$, $t = 0$;
2: **for** $k = 1, \ldots,$ until convergence **do**
3: **if** $t \geq 1$ **then**
4: Find an orthonormal basis matrix $Q_k \in \mathbb{R}^{n\times r_k}$ of $\mathcal{S}$.
5: Solve (5.4) by RTR for Y_k and set $X_{k-1} = Q_k Y_k$.
6: **end if**
7: Call LOBPCG (or other eigensolver) to find an approximate orthonormal eigenbasis matrix $\widehat{X}_k \in \mathbb{R}^{n\times \ell}$ associated with the first ℓ smallest eigenvalues of $E(X_{k-1})$.
8: Compute the orthogonal polar factor P of $\widehat{X}_k^{\mathrm{T}} B$, and set $X_k = -\widehat{X}_k P$.
9: **if** $t = \varsigma_m$ **then**
10: $\mathcal{S}(:, 1:\ell) = [\,]$ and $t = t - 1$
11: **else**
12: $\mathcal{S} = [\mathcal{S}, X_k]$ and $t = t + 1$
13: **end if**
14: **end for**

6. Numerical experiments. In this section, we will conduct numerical testing to show basic behavior of our method; also we will evaluate the performance on a preliminary test problems and report the numerical experiments together with those from other generic optimization methods designed upon the Stiefel manifold. All of these experiments were conducted on a PC using Windows 7 (64bit) system with Intel Core i5-3230M CPU (2.6GHz) and 4GB memory on MATLAB (2010b).

There are two goals in our numerical report: (i) to demonstrate the improvements of the inexact eigenvalue computation and our subspace acceleration strategy used in Algorithm 5.1 over the pure SCF iteration Algorithm 3.1, and (ii) to show the efficiency of the algorithm SCFRTR by comparing it with four other solvers (the RTR method of Absil, Baker, and Gallivan [1], the WYBB algorithm by Wen and Yin [42], the JDCP algorithm of Jiang and Dai [20], and the GPI iteration of Nie, Zhang, and Li [30]). Following the stopping criterion used in [42, 20], for our method SCFRTR we terminate the iteration in step 2 of Algorithm 5.1 when $k > 60$ or if any of the following three conditions is satisfied:

(6.1)
$$\frac{f(X_k) - f(X_{k+1})}{|f(X_k)| + 1} \leq \mathtt{tol}_f, \ \frac{\|X_k - X_{k+1}\|_F}{\sqrt{n}} \leq \mathtt{tol}_x, \ \|AX_k + B + X_k\Lambda_k\|_F \leq \mathtt{tol}_g.$$

Note that the last condition of (6.1) is the absolute norm of the gradient at the computed point X_k, which might be affected by the scaling of the data matrices A and B. Therefore, in order to eliminate the impaction from the scaling and in order to have a relative norm of the gradient at the computed point, for a given problem with data matrices A and B in our testing, we conduct a procedure

$$A \leftarrow \frac{A}{\|A\|_1 + \|B\|_1}, \quad B \leftarrow \frac{B}{\|A\|_1 + \|B\|_1}.$$

We remark that, without modifying the stopping criteria used in [42, 20], the absolute norm of the gradient (6.1) on the scaled problem now serves as an approximation of

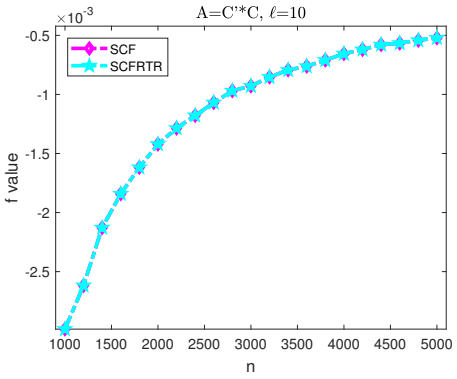


FIG. 1. *Optimal values.*

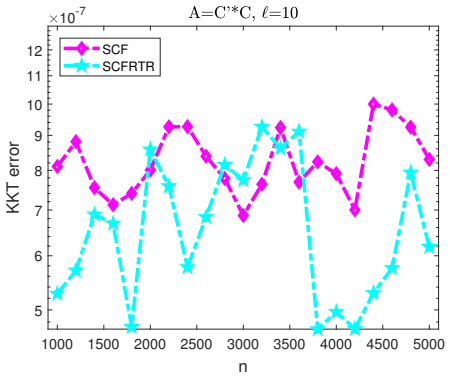


FIG. 2. *KKT error.*

the relative norm of the gradient at X_k for the unscaled one. In our testing, we set $\texttt{tol}_f = 10^{-12}$ and $\texttt{tol}_x = \texttt{tol}_g = 10^{-6}$; moreover, unless otherwise specified, we choose X_0 as the orthonormal basis of $\mathcal{R}(B)$ by the modified Gram–Schmidt process if $\text{rank}(B) = \ell$ and $X_0 = [I_\ell, 0]^{\mathrm{T}}$ otherwise, for all methods in the tests of sections 6.1, 6.2, and 6.3.

6.1. The pure SCF versus the practical SCF. In this subsection, we will demonstrate the behavior of our practical SCFRTR (Algorithm 5.1) by comparing that of the pure SCF algorithm (Algorithm 3.1). For this purpose, we choose the data matrices A and B generated by MATLAB command `randn`: `C=randn(n)`; `A=C'*C`; `B=randn(n,`ℓ`)` for (1.2). With an increment 200, we choose various n in the interval $[1000, 5000]$, and compare the results of optimal values, KKT errors $\|AX_k + B + X_k\Lambda_k\|_F$, the number of iterations, and running CPU time (in seconds) from SCFRTR (Algorithm 5.1) and SCF (Algorithm 3.1), in Figures 1–4, respectively. In our implementation of the baseline SCF in Algorithm 3.1, we use the MATLAB function `eigs` to compute an orthonormal eigenbasis matrix $\widehat{X}_k$ in step 2. For SCFRTR, we call LOBPCG by setting the inner tolerance $\mathsf{tol\_lobpcg} = 10^{-7}$ and activate the Cholesky factor of A (or the incomplete Cholesky for sparse case) as the preconditioner whenever $\|A\|_1/\|B\|_1 \geq 100$ and the estimated condition number of A is larger than 10^6; moreover, we set the maximal inner iterations $\mathsf{maxit\_lobpcg} = 200$ in the first three outer iterations when no preconditioner is used, and set $\mathsf{maxit\_lobpcg} \leq 20$ for others.

Figures 1 and 2 show that, in terms of the optimal values and KKT errors, SCFRTR and SCF nearly are of the same performance. However, for the efficiency, we observed from Figures 3 and 4 that SCFRTR saves about 2/3 fraction of the number of outer-loop iterations, and reduces the running times in CPU significantly from those used in the pure SCF iteration. This demonstrates the improvement of our practical SCF iteration over the pure SCF.

6.2. Comparison with other solvers. We now compare our algorithm with the RTR algorithm of Absil, Baker, and Gallivan [1], the WYBB algorithm of Wen and Yin [42], the JDCP algorithm of Jiang and Dai [20], and the GPI iteration of Nie, Zhang, and Li [30]. Here, RTR [1] is a generic trust-region method for minimization over a general Riemannian manifold (we terminate RTR if the norm of Riemannian gradient is no greater than 10^{-4}), while WYBB and JDCP are two generic optimization methods built on the Stiefel manifold. In particular, the WYBB algorithm

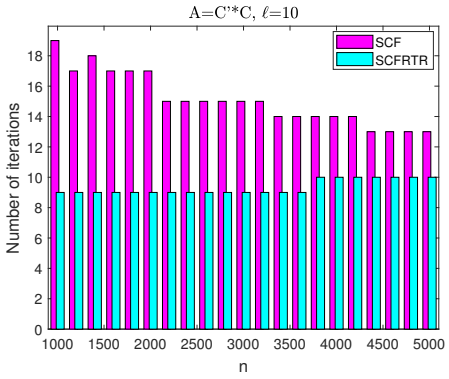


FIG. 3. *Number of iterations.*

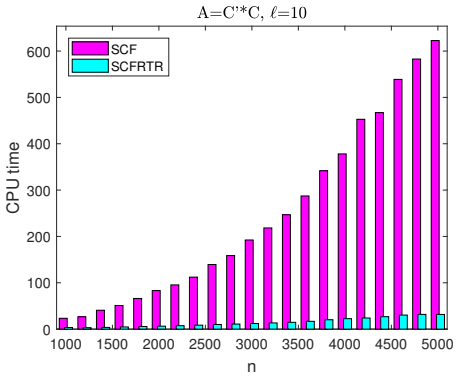


FIG. 4. *CPU time.*

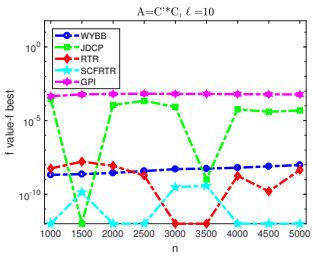


FIG. 5. *Difference of f.*

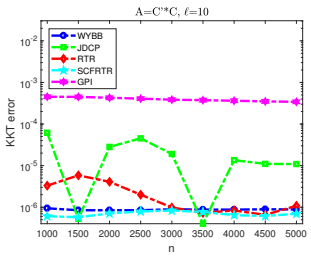


FIG. 6. *KKT error.*

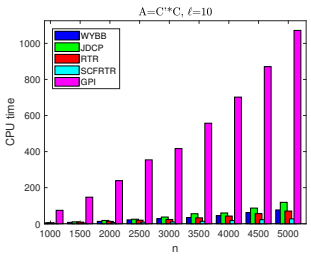


FIG. 7. *CPU time.*

is a feasible gradient method which is based on the Cayley transform and Barzilai–Borwein stepsizes, and JDCP is a uniform framework of feasible gradient methods which decompose each feasible point X into the range space of X and the null space of X^{T}. GPI is the generalized power iteration especially designed for the unbalanced Procrustes problem (1.1), and the same stopping criteria (6.1) as SCFRTR are set for GPI in our testing.

6.2.1. Test on synthetic problems. We first report the numerical results of (1.2) obtained by solving the randomly generated problems in section 6.1 but with $1000 \le n \le 5000$. In Figures 5–7 we plot the averaged quantities of the difference of optimal values, KKT errors, and CPU time over ten random tests for each n. The notion of "f best" in Figure 5 stands for the smallest computed optimal values among all solvers. These preliminary numerical results demonstrate the efficiency of SCFRTR for our generated minimization problems (1.2).

We next extend our numerical evaluation. As has been pointed out in Remark 4.6, the factor of $\frac{\|B\|2}{\eta}$ has an impact on the convergence of the SCF iteration. Note that the eigenvalue gap $\eta = \lambda_{\ell+1}(E(X_{\mathrm{opt}})) - \lambda_{\ell}(E(X_{\mathrm{opt}}))$ cannot be estimated a priori; in order to evaluate behavior of the iteration with respect to that factor, we therefore only choose to vary the relative norm $\|A\|/\|B\|$ between A and B. In particular, we test both the minimization (1.2) and the Procrustes problem (1.1) by selecting various parameters $a > 0$ in

$$\min_{X \in \mathbb{O}^{n\times\ell}} \operatorname{trace}(X^{\mathrm{T}} A_a X) + 2\operatorname{trace}(X^{\mathrm{T}} B), \quad A_a = aA, \tag{6.2}$$

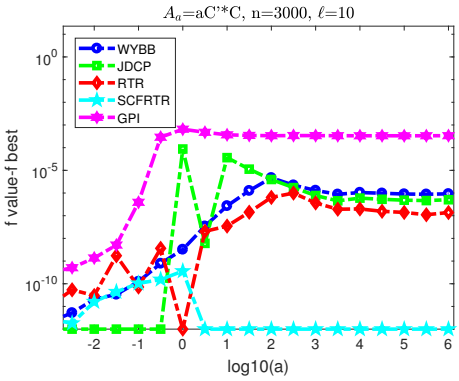


FIG. 8. *Difference of optimal values for* (6.2).

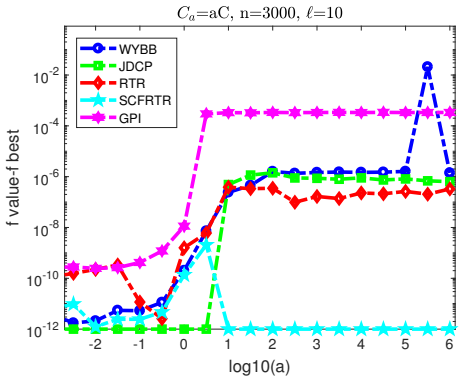


FIG. 9. *Difference of optimal values for* (6.3).

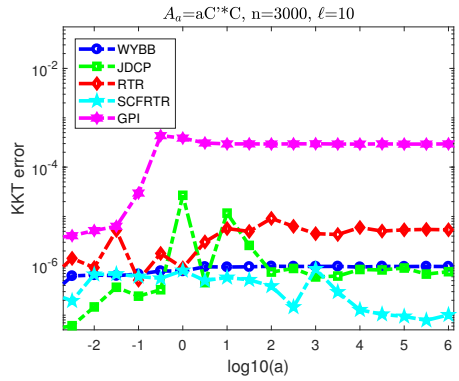


FIG. 10. *KKT error for* (6.2).

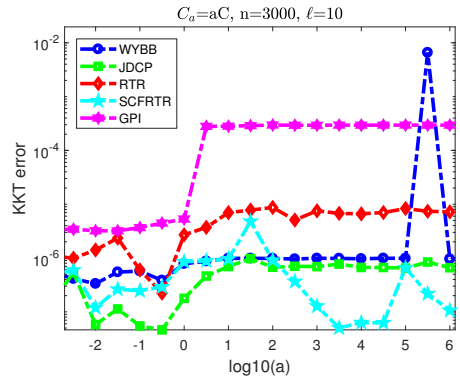


FIG. 11. *KKT error for* (6.3).

and

$$\min_{X\in\mathbb{O}^{n\times\ell}} \|C_aX - D\|_F^2, \quad C_a = aC, \tag{6.3}$$

where B, C, and D are generated with entries randomly drawn from standard normal distribution, and $A = C^{\mathrm{T}}C$ in (6.2). By fixing $n = 3000$ and $\ell = 10$, in Figures 8–13, we plot the averaged quantities of the difference of optimal values, KKT errors, and CPU time, for the minimization problems (6.2) and (6.3), over ten random tests for each a.

We observed that as a gets large, the quadratic term $\mathrm{trace}(X^{\mathrm{T}}AX)$ dominates the function $f(X)$ and thus (6.2) gradually becomes an eigenvalue problem, for which our eigenvalue-based solver SCFRTR manifests its efficiency. This can be seen by the consuming CPU times (we also noticed that the number of outer-loop iterations of SCFRTR does not change much when $a \geq 10^3$ and $a \geq 10^2$ for (6.2) and (6.3), respectively). On the other hand, for a small $a \leq 1$, even though all solvers efficiently handle (6.2) and (6.3), our method SCFRTR generally needs a little bit more CPU time to reach the stopping rule than RTR, WYBB, and JDCP. We think that in this case, the linear term $\mathrm{trace}(X^{\mathrm{T}}B)$ dominates $f(X)$, and gradient-based methods are able to solve it efficiently; the consuming CPU time of our SCFRTR is due to the calling of the eigensolver LOBPCG. Overall, this type of numerical testing also illustrates that our eigenvalue-based method SCFRTR is an efficient candidate to

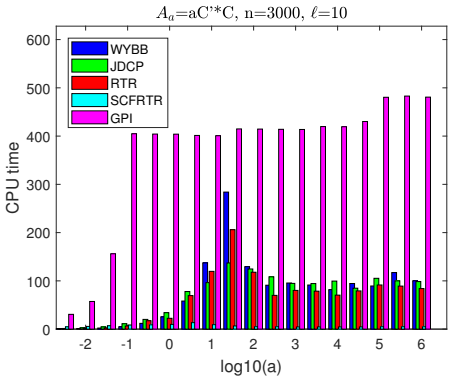


FIG. 12. *CPU time for* (6.2).

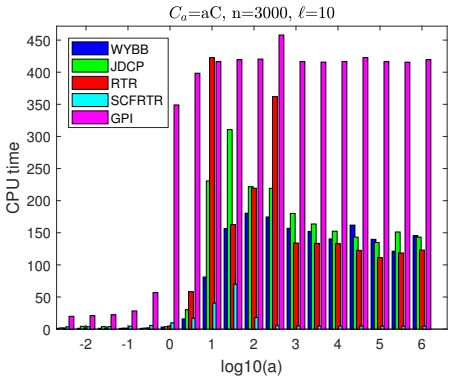


FIG. 13. *CPU time for* (6.3).

solve the minimization (1.2) or the Procrustes problem (1.1).

6.2.2. Test on the orthogonal least squares regression for feature extraction. We now evaluate the efficiency of our method on a practical model in data mining. The orthogonal least squares regression (OLSR) proposed in [51] (see also [30]) is a supervised learning method in linear discriminant analysis. Suppose we are given a dataset $S = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m] \in \mathbb{R}^{n\times m}$ containing m samples with n features drawn from ℓ classes. Let $K = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_m] \in \mathbb{R}^{\ell\times m}$ be the corresponding class indicator matrix; in particular, if the sample $\boldsymbol{s}_i$ belongs to the jth class, then $\boldsymbol{k}_i = \boldsymbol{e}_j = [0, \ldots, 0, 1, 0, \ldots, 0]^{\mathrm{T}} \in \mathbb{R}^{\ell}$. OLSR is based on the least squares regression to establish a dimension reduction procedure, and the model is to find a transformation matrix $X \in \mathbb{O}^{n\times\ell}$ with usually $\ell \ll n$ and an associated bias $\boldsymbol{b} \in \mathbb{R}^{\ell}$ by solving

$$\min_{X\in\mathbb{O}^{n\times\ell},\ \boldsymbol{b}\in\mathbb{R}^{\ell}} \left\{ J(X,\boldsymbol{b}) := \|S^{\mathrm{T}}X + \boldsymbol{e}\boldsymbol{b}^{\mathrm{T}} - K^{\mathrm{T}}\|_F^2 \right\}, \tag{6.4}$$

where $\boldsymbol{e} = [1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^m$. It is claimed that the orthogonal constraint of the transformation matrix X is able to preserve more local structure and discriminant information [51]. Take the partial derivative of $J(X,\boldsymbol{b})$ with respect to $\boldsymbol{b}$ and set it to zero to have $\boldsymbol{b} = (K\boldsymbol{e} - X^{\mathrm{T}}S\boldsymbol{e})/m$, (6.4) can be solved by the following unbalanced Procrustes problem:

$$\min_{X\in\mathbb{O}^{n\times\ell}} \|CX - D\|_F^2, \quad \text{where } C = \left(I_m - \frac{\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}}{m}\right) S^{\mathrm{T}},\ D = \left(I_m - \frac{\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}}{m}\right) K^{\mathrm{T}}. \tag{6.5}$$

For our testing in this case, we choose image datasets and text datasets[4] summarized in Table 1. With the same stopping criterion used in subsection 6.2.1, we report the numerical results of the five methods in Table 2. Our experiments on the nine datasets in Table 1 demonstrate the efficiency of SCFRTR for OLSR model (6.5).

TABLE 1
Summary of test datasets.

	Dataset	Feature (n)	Number of samples (m)	Number of classes (ℓ)
Image	ORL	10304	400	40
	Mpeg-7	6000	1400	70
Text	Text-1	7511	1946	2
	20Newsgroups-4	8014	3970	4
	Cora-DS	6234	751	9
	Cora-HA	3989	400	7
	Cora-ML	8329	1617	4
	Cora-OS	6737	1246	4
	Cora-PL	7949	1575	9

TABLE 2
Numerical results on OLSR model for the datasets in Table 1.

Dataset	KKT error					CPU time (in seconds)				
	RTR	WYBB	JDCP	GPI	SCFRTR	RTR	WYBB	JDCP	GPI	SCFRTR
ORL	4.96e-6	9.66e-7	2.25e-6	3.98e-5	3.12e-7	5.95e+2	7.34e+2	8.26e+2	1.42e+4	9.28e+1
Mpeg-7	8.28e-5	9.02e-7	8.57e-7	3.47e-4	9.53e-7	7.55e+2	8.54e+2	8.05e+2	5.54e+3	3.82e+2
Text-1	2.97e-5	9.32e-7	3.44e-7	1.17e-5	8.17e-7	2.94e+1	3.48e+1	4.39e+1	9.28e+2	1.02e+1
20Newsgroups-4	7.52e-5	9.79e-7	8.76e-7	5.61e-5	9.82e-7	6.62e+1	7.40e+2	5.00e+2	2.08e+3	7.94e+1
Cora-DS	9.11e-5	9.38e-7	3.21e-7	3.64e-5	5.26e-7	4.29e+1	7.20e+1	1.09e+2	1.93e+3	5.45e+1
Cora-HA	6.88e-5	7.00e-7	4.14e-7	3.14e-5	4.56e-7	3.06e+1	2.75e+1	3.49e+1	7.41e+2	2.07e+1
Cora-ML	2.23e-5	9.14e-7	1.20e-6	9.20e-6	9.29e-7	1.46e+2	8.83e+1	1.18e+2	2.99e+3	7.35e+1
Cora-OS	6.40e-6	8.83e-7	2.69e-7	7.85e-6	4.40e-7	6.80e+1	4.61e+1	6.90e+1	1.51e+3	3.28e+1
Cora-PL	3.21e-5	9.60e-7	4.53e-7	3.03e-5	4.80e-7	1.00e+2	8.95e+1	1.03e+2	3.31e+3	8.39e+1

6.3. The effectiveness in finding a better KKT point. Besides the computation efficiency, we finally remark that our SCF iteration, which is based on a necessary optimality condition (2.4) for the global minimizer, seems more prone to find the global minimizer than a generic optimization method. Using the same stopping criterion for each method as in section 6.2, we have two types of numerical evaluation for this claim.

In the first test, for a given pair (n, ℓ), we generate 10^4 problems (1.2) each with `C=randn(n)`, `A=C'*C`, and `B=randn(n,`ℓ`)`. For each particular test problem, we compare the final computed objective function values from RTR, WYBB, JDCP, GPI,

[4]All datasets are available at http://www.escience.cn/people/fpnie/papers.html, except for the dataset ORL. The ORL database (http://featureselection.asu.edu/datasets.php) consists of 400 face images taken from 40 distinct subjects, each with 10 images taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling), and facial details (glasses/no glasses). The description of other datasets in Table 1 can be found in [13, Chapter 5].

and SCF[5] starting from the same random initial guess. In order to evaluate whether a method $\mathtt{M}_1$ returns a better computed solution than the other method $\mathtt{M}_2$ (denoted by $\mathtt{M}_1 \rhd \mathtt{M}_2$), we count the numbers of cases for $\mathtt{M}_1$ when $f_{\mathtt{M}_2} - f_{\mathtt{M}_1} \geq 0.1 \cdot |f_{\mathtt{M}_1}|$ (such a criterion, in general, protects the effect from the rounding-off errors), where $f_{\mathtt{M}_1}$ and $f_{\mathtt{M}_2}$ represent the computed objective function values from $\mathtt{M}_1$ and $\mathtt{M}_2$, respectively. Table 3 summarizes the number of problems for which the objective function values from SCF are superior (denoted by SCF $\rhd$) and inferior to (denoted by SCF $\lhd$) others for various pairs (n, ℓ). Also reported are the mean and standard deviation of the objective function values for each method. As the computed solution from the SCF iteration satisfies a global necessary condition (2.4), it can be seen that, in many cases, our customized SCF iteration can achieve a KKT point with a smaller objective value than others.

TABLE 3

Number of cases (among 10^4) in which SCF is superior and inferior to others, and the mean and standard deviation of objective values for each method.

	Number of SCF $\rhd$ M (SCF $\lhd$ M) Mean (std) of the computed $f_{\mathtt{M}}$				
(n, ℓ)	RTR	WYBB	JDCP	GPI	SCF
$(5, 2)$	1765 (134) -2.053 (1.1)	1717 (168) -2.064 (1.1)	1719 (142) -2.060 (1.1)	1815 (178) -2.060 (1.1)	- -2.214 (1.0)
$(5, 3)$	3239 (169) -2.128 (1.5)	3197 (170) -2.135 (1.5)	3076 (166) -2.146 (1.5)	3334 (217) -2.120 (1.5)	- -2.416 (1.5)
$(10, 3)$	954 (98) -4.386 (1.3)	930 (117) -4.390 (1.3)	893 (103) -4.393 (1.3)	1025 (164) -4.384 (1.3)	- -4.471 (1.2)
$(10, 5)$	1940 (99) -5.009 (2.0)	1945 (104) -5.009 (2.0)	1827 (100) -5.021 (2.0)	1972 (126) -5.008 (2.0)	- -5.218 (2.0)
$(10, 7)$	3031 (175) -1.170 (3.4)	3065 (204) -1.170 (3.4)	2894 (190) -1.183 (3.4)	2983 (180) -1.174 (3.4)	- -1.395 (3.4)

In our second numerical evaluation, the data matrices $F, M \in \mathbb{R}^{10\times 4}$ (see [6, Table 19.1]) given in (6.6) are from the real application of the Procrustes problem (1.1) in the multidimensional similarity structure analysis (SSA) [6].

(6.6)
$$F = \begin{bmatrix} 0.08 & 0.59 & 2.54 & 0.23 \\ 2.05 & 0.31 & 0.20 & 0.03 \\ 1.22 & 0.73 & 0.33 & 1.63 \\ 2.11 & 0.25 & 0.79 & 1.15 \\ 1.75 & -0.14 & 1.56 & -0.93 \\ 1.84 & -0.12 & 1.17 & 1.09 \\ 0.27 & 1.31 & 1.82 & -0.68 \\ 0.41 & 2.17 & 0.04 & 0.91 \\ 0.49 & 2.46 & -0.22 & -0.07 \\ 0.13 & 1.96 & 0.14 & -0.39 \end{bmatrix}, \quad M = \begin{bmatrix} 1.89 & -0.76 & 1.50 & -0.76 \\ 2.30 & -1.30 & 0.27 & 1.41 \\ 2.26 & -0.34 & -1.15 & -1.38 \\ 2.53 & -1.21 & -0.92 & 0.29 \\ 1.65 & -1.05 & 1.33 & -0.38 \\ 2.20 & -1.42 & -0.71 & 0.28 \\ 2.06 & 0.88 & 1.46 & -0.39 \\ 2.54 & 1.63 & -1.22 & 0.10 \\ 1.83 & 2.01 & -0.23 & -0.34 \\ 1.74 & 2.01 & 0.77 & 1.00 \end{bmatrix}.$$

The matrix F represents the coordinates (configuration) of a vector representation of ten emotions, and M is the corresponding solution for a replication study [6, Chapter 19]. In SSA, two configurations are similar if they can be brought to a complete match by rigid motions and dilations [6]. As pointed out in [6, section 19.8], two configurations generally have the same dimensionality, but sometimes it is useful for

[5]Due to the small dimension n, in this subsection, we inactivated our subspace acceleration procedure in subsection 5.2 and just used the basic SCF iteration.

them not to have the same dimensionality. A technical treatment [6, section 19.8] for the latter situation is to append zero columns on one configuration until it matches the other, and then solve the balanced Procrustes problem. Another possible way is just to solve the unbalanced Procrustes problem (1.1). Therefore, in our experiments, we choose either F or M with certain columns deleted as the target matrix D to be used in (1.1), and the other is for C. Note that $F, M \in \mathbb{R}^{10\times 4}$ and we can delete one and two columns to have $\ell = 3$ and $\ell = 2$, respectively. For each problem's result, we randomly generate 10^4 initial points for the five methods. Excluding cases where all methods return the same objective function value, in Tables 4 and 5, we report the numbers of problems for which the objective values from SCF are superior and inferior to others among these 10^4 tests. The results again indicate that, in many cases, SCF can converge to a KKT point with a smaller objective value than others.

TABLE 4

Number of cases (among 10^4) in which SCF is superior and inferior to others, and the mean and standard deviation of objective values. M (with certain columns deleted) is the D in (1.1).

		Number of SCF $\triangleright$ M (SCF $\triangleleft$ M) Mean (std) of the computed $f_{\mathtt{M}}$				
	$D = M$ with	RTR	WYBB	JDCP	GPI	SCF
$\ell = 3$	$M(:,1) = [\ ]$	3819 (0) 5.403 (3.0)	3676 (0) 5.315 (3.0)	3722 (0) 5.344 (3.0)	4573 (0) 5.866 (3.0)	- 3.057 (0.0)
$\ell = 3$	$M(:,2) = [\ ]$	3106 (0) 5.699 (2.9)	3135 (0) 5.716 (2.9)	3193 (0) 5.752 (2.9)	2601 (0) 5.388 (2.7)	- 3.786 (0.0)
$\ell = 2$	$M(:,[1,2]) = [\ ]$	3337 (0) 4.661 (2.9)	3234 (0) 4.598 (2.9)	3308 (0) 4.643 (2.9)	3890 (0) 5.001 (3.0)	- 2.609 (0.0)

TABLE 5

Number of cases (among 10^4) in which SCF is superior and inferior to others, and the mean and standard deviation of objective values. F (with certain columns deleted) is the target D in (1.1).

		Number of SCF $\triangleright$ M (SCF $\triangleleft$ M) Mean (std) of the computed $f_{\mathtt{M}}$				
	$D = F$ with	RTR	WYBB	JDCP	GPI	SCF
$\ell = 3$	$F(:,1) = [\ ]$	2562 (2068) 4.078 (2.1)	2171 (2173) 3.862 (2.1)	2366 (2224) 3.929 (2.2)	1822 (2507) 3.565 (2.0)	- 3.862 (2.0)
$\ell = 3$	$F(:,2) = [\ ]$	4100 (0) 5.712 (3.0)	4091 (0) 5.706 (3.0)	4226 (0) 5.789 (3.0)	4348 (0) 5.864 (3.1)	- 3.185 (0.0)

7. Conclusions. In this paper, we treat the numerical solution of the unbalanced orthogonal Procrustes problem or the minimization of a quadratic function over a Stiefel manifold from the eigenvalue point of view. With a motivation of exploiting the modern techniques in eigenvalue computations, we first connect the problem (1.2) with an eigenvalue minimization (3.2), for which an eigenvector-dependent nonlinear eigenvalue problem (3.3) can be established. Our proposed algorithm is a succinct SCF iteration in which each iteration involves computing a particular eigenbasis matrix of a symmetric matrix. The convergence behavior of this basic SCF iteration is analyzed and two practical implementation treatments are further built in to make the pure SCF iteration more efficient. Preliminary numerical testings are conducted to indicate that such an eigenvalue-based approach is an efficient way to solve the unbalanced Procrustes problem or the minimization (1.2), particularly for $\ell \ll n$.

Acknowledgments. The authors are grateful to the associate editor and two anonymous referees for their helpful comments and suggestions which significantly improved the presentation of this paper.

REFERENCES

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.
[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
[3] L.-E. Andersson and T. Elfving, *A constrained Procrustes problem*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 124–139, https://doi.org/10.1137/S0895479894277545.
[4] Z. Bai, D. Lu, and B. Vandereycken, *Robust Rayleigh quotient minimization and nonlinear eigenvalue problems*, SIAM J. Sci. Comput., 40 (2018), pp. A3495–A3522, https://doi.org/10.1137/18M1167681.
[5] A. W. Bojanczyk and A. Lutoborski, *The Procrustes problem for orthogonal Stiefel matrices*, SIAM J. Sci. Comput., 21 (1999), pp. 1291–1304, https://doi.org/10.1137/S106482759630992X.
[6] I. Borg and J. Lingoes, *Multidimensional Similarity Structure Analysis*, Springer-Verlag, New York, 1987.
[7] Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li, *On an eigenvector-dependent nonlinear eigenvalue problem*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1360–1382, https://doi.org/10.1137/17M115935X.
[8] M. T. Chu and N. T. Trendafilov, *The orthogonally constrained regression revisited*, J. Comput. Graph. Stat., 10 (2001), pp. 746–771.
[9] C. Davis and W. Kahan, *The rotation of eigenvectors by a perturbation,* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46, https://doi.org/10.1137/0707001.
[10] J. P. Van de Geer, *Linear relations among k sets of variables*, Psychometrika, 49 (1984), pp. 70–94.
[11] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997, https://doi.org/10.1137/1.9781611971446.
[12] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353, https://doi.org/10.1137/S0895479895290954.
[13] T. Hokimoto, ed., *Advances in Statistical Methodologies and Their Application to Real Problems*, InTech, Rijeka, Croatia, 2017.
[14] L. Eldén and H. Park, *A Procrustes problem on the Stiefel manifold*, Numer. Math., 82 (1999), pp. 599–619.
[15] D. M. Gay, *Computing optimal locally constrained steps*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 186–197, https://doi.org/10.1137/0902016.
[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.
[17] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*, Oxford University Press, New York, 2004.
[18] N. J. Higham, *The symmetric Procrustes problem*, BIT, 28 (1998), pp. 133–143.
[19] J. R. Hurley and R. B. Cattell, *The Procrustes program: Producing direct rotation to test a hypothesized factor structure*, Beh. Sci., 7 (1962), pp. 258–262.
[20] B. Jiang and Y.-H. Dai, *A framework of constraint preserving update schemes for optimization on Stiefel manifold*, Math. Program., 153 (2015), pp. 535–575.
[21] C. Kanzow and H.-D. Qi, *A QP-free constrained Newton-type method for variational inequality problems*, Math. Program., 85 (1999), pp. 81–106.
[22] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541, https://doi.org/10.1137/S1064827500366124.
[23] A. V. Knyazev and K. Neymeyr, *Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method*, Electron. Trans. Numer. Anal., 15 (2003), pp. 38–55.
[24] L. Li and Z. Zhang, *Semi-supervised domain adaptation by covariance matching*, IEEE Trans. Pattern Anal. Mach. Intell., 41 (2019), pp. 2724–2739.
[25] X. Liu, X. Wang, Z. Wen, and Y. Yuan, *On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Matrix Anal. Appl., 35 (2014),

pp. 546–558, https://doi.org/10.1137/130911032.
[26] X.-G. Liu, X.-F. Wang, and W.-G. Wang, *Maximization of matrix trace function of product Stiefel manifolds*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1489–1506, https://doi.org/10.1137/15M100883X.
[27] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, UK, 2004.
[28] J. M. Martínez, *Local minimizers of quadratic functions on Euclidean balls and spheres*, SIAM J. Optim., 4 (1994), pp. 159–176, https://doi.org/10.1137/0804009.
[29] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572, https://doi.org/10.1137/0904038.
[30] F. Nie, R. Zhang, and X. Li, *A generalized power iteration method for solving quadratic problem on the Stiefel manifold*, Sci. China Inf. Sci., 60 (2017), 112101.
[31] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.
[32] L. Qiu, Y. Zhang, and C.-K. Li, *Unitarily invariant metrics on the Grassmann space*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 507–531, https://doi.org/10.1137/040607605.
[33] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, *Numerical methods for electronic structure calculations of materials*, SIAM Rev., 52 (2010), pp. 3–54, https://doi.org/10.1137/060651653.
[34] G. A. F. Seber, *A Matrix Handbook for Statisticians*, John Wiley & Sons, Hoboken, NJ, 2007.
[35] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426, https://doi.org/10.1137/0719026.
[36] G. W. Stewart, *Matrix Algorithms*: Volume II: Eigensystems, SIAM, Philadelphia, 2001, https://doi.org/10.1137/1.9780898718058.
[37] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[38] T. Viklands, *Algorithms for the Weighted Orthogonal Procrustes Problem and Other Least Squares Problems*, Ph.D. thesis, Umeå University, Umeå, Sweden, 2006.
[39] J. von Neumann, *Some matrix-inequalities and metrization of matrix-space*, Tomck. Univ. Rev., 1 (1937), pp. 286–300.
[40] Z. Wang, Q. Ruan, and G. An, *Projection-optimal local Fisher discriminant analysis for feature extraction*, Neural Comput. & Applic., 26 (2015), pp. 589–601.
[41] G. A. Watson, *The solution of orthogonal Procrustes problems for a family of orthogonally invariant norms*, Adv. Comput. Math., 2 (1994), pp. 393–405.
[42] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Math. Program., 142 (2013), pp. 397–434.
[43] C. Yang, W. Gao, and J. C. Meza, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1773–1788, https://doi.org/10.1137/080716293.
[44] L.-H. Zhang, *Uncorrelated trace ratio LDA for undersampled problems*, Patt. Recog. Lett., 32 (2011), pp. 476–484.
[45] L.-H. Zhang, *On a self-consistent-field-like iteration for maximizing the sum of the Rayleigh quotients*, J. Comput. Appl. Math., 257 (2014), pp. 14–28.
[46] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold,* I*: Theory*, Sci. China Math., 57 (2014), pp. 2495–2508.
[47] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold,* II*: Computation*, Sci. China Math., 58 (2015), pp. 1549–1566.
[48] L.-H. Zhang, L.-Z. Liao, and M. K. Ng, *Fast algorithms for the generalized Foley–Sammon discriminant analysis*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1584–1605, https://doi.org/10.1137/080720863.
[49] Z. Y. Zhang and K. Q. Du, *Successive projection method for solving the unbalanced procrustes problem*, Sci. China Math., 49 (2006), pp. 971–986.
[50] Z. Y. Zhang, Y. Qiu, and K. Q. Du, *Conditions for optimal solutions of unbalanced Procrustes problem on Stiefel manifold*, J. Comput. Math., 25 (2007), pp. 661–671.
[51] H. Zhao, Z. Wang, and F. Nie, *Orthogonal least squares regression for feature extraction*, Neurocomputing, 216 (2016), pp. 200–207.