Comput Optim Appl., manuscript No.
(will be inserted by the editor)

On optimizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient on the unit sphere

Lei-Hong Zhang

Abstract Given symmetric matrices $B, D \in \mathbb{R}^{n\times n}$ and a symmetric positive definite matrix $W \in \mathbb{R}^{n\times n}$, maximizing the sum of the Rayleigh quotient $\mathbf{x}^\top D\mathbf{x}$ and the generalized Rayleigh quotient $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ on the unit sphere not only is of mathematical interest in its own right, but also finds applications in practice. In this paper, we first present a real world application arising from the sparse Fisher discriminant analysis. To tackle this problem, our first effort is to characterize the local and global maxima by investigating the optimality conditions. Our results reveal that finding the global solution is closely related with a special extreme nonlinear eigenvalue problem, and in the special case $D = \mu W$ $(\mu > 0)$, the set of the global solutions is essentially an eigenspace corresponding to the largest eigenvalue of a specially-defined matrix. The characterization of the global solution not only sheds some lights on the maximization problem, but motives a starting point strategy to obtain the global maximizer for any monotonically convergent iteration. Our second part then realizes the Riemannian trust-region method of [Absil, Baker and Gallivan, *Found. Comput. Math.*, 7, 303-330 (2007)] into a practical algorithm to solve this problem, which enjoys the nice convergence properties: global convergence and local superlinear convergence. Preliminary numerical tests are carried out and empirical evaluation of its performance is reported.

Keywords Rayleigh quotient · the generalized Rayleigh quotient · the nonlinear eigenvalue problem · trust-region method · linear discriminant analysis

Mathematics Subject Classification (2000) 62H20 · 15A12 · 65F10 · 65K05

This work was supported by the National Natural Science Foundation of China NSFC-11101257.

Lei-Hong Zhang (Corresponding author)
Department of Applied Mathematics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, People's Republic of China.
Tel.: +86-21-65904076
E-mail: longzlh@gmail.com

1 Introduction

Given symmetric matrices B, W and $D \in \mathbb{R}^{n\times n}$ where W is positive definite, in this paper, we are concerned with the solution of the following optimization problem:

$$\max_{\|\mathbf{x}\|_2=1} f(\mathbf{x}) := \frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}} + \mathbf{x}^\top D\mathbf{x}. \tag{1.1}$$

Because of the unit sphere constraint

$$\mathscr{M} := \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\|_2 = 1\},$$

we know that the solution of (1.1) does not change if we replace the matrix D with $D + \xi I_n$ for any $\xi \in \mathbb{R}$, where I_n represents the n-by-n identity matrix; without loss of any generality, therefore, we can further assume D is positive definite. Practical applications of this problem can arise, for example, in the downlink of a multi-user MIMO system [33] and in the sparse Fisher discriminant analysis in pattern recognition. In section 2, we shall briefly discuss how a corresponding problem (1.1) is formulated. On the other hand, as maximizing either $\mathbf{x}^\top D\mathbf{x}$ or $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ on $\mathscr{M}$ is equivalent to solving the extreme eigenvalue problem $D\mathbf{x} = \lambda\mathbf{x}$ or computing an extreme eigenpair of a symmetric-definite matrix pencil (B, W) (see e.g., [22]), respectively, one can expect that problem (1.1) will also be related with some special extreme eigenvalue problem, which makes it attractive in its own right.

Though it is known that optimizing either $\mathbf{x}^\top D\mathbf{x}$ or $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ on $\mathscr{M}$ does not admit local non-global solution (see e.g., [21,20]), difficulty will arise when one attempts to maximize the sum, where multiple local non-global maxima appear. The following simple example where all B, W and D are 2-by-2 and diagonal illustrates the different situation.

Example 1.1 We let $B = \mathrm{diag}\{1,2\}, W = \mathrm{diag}\{10,1\}, D_0 = 0 \in \mathbb{R}^{2\times 2}$ and $D_1 = \mathrm{diag}\{5,1\}$. For the choice $D = D_0$, the global maxima of (1.1) are $[0,\pm 1]^\top$ and the global minima are $[\pm 1, 0]^\top$; whereas for the case $D = D_1$, the global maxima are $[\pm 1, 0]^\top$ and the global minima are[1]

$$[\pm\frac{\sqrt{\frac{\sqrt{19}}{2}-1}}{3}, \pm\frac{\sqrt{10-\frac{\sqrt{19}}{2}}}{3}]^\top.$$

The points $[0,\pm 1]^\top$ now become the local maxima. The value of the cost function $f(\mathbf{x}_1, \mathbf{x}_2)$ on the unit sphere $\mathbf{x}_1^2 + \mathbf{x}_2^2 = 1$ is plotted in Figure 1.

Even in this very simple example, we can see the global maxima for the trivial case $D = D_0$ and $D = D_1$ vary greatly, and thus one cannot expect that a global solution of a general problem (1) could be obtained by simply solving a standard eigenvalue problem or a generalized eigenvalue problem. To clarify our statement, in our first part of this paper, we will investigate the local as well as the global optimality conditions in section 3. It is interesting to point out that any local maximizer $\overline{\mathbf{x}}$ is a unit eigenvector corresponding to either the largest or the second largest eigenvalue of a specially-defined matrix $E(\overline{\mathbf{x}})$ (see section 3); moreover, as a global optimality condition, we can prove that any global solution $\mathbf{x}^*$

[1] We will demonstrate how to compute these points in section 3.

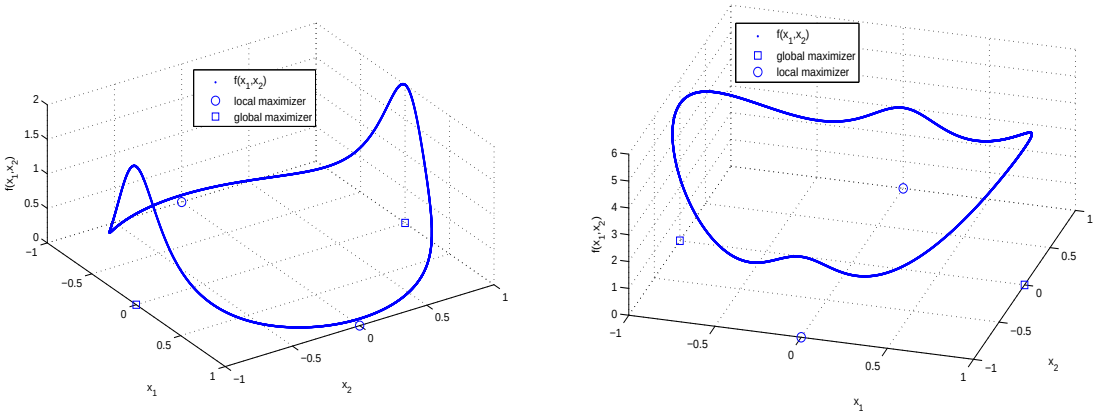

Fig. 1 The left sub-figure plots the value of the cost function $f(\mathbf{x}_1,\mathbf{x}_2)$ on the unit sphere $\mathbf{x}_1^2+\mathbf{x}_2^2=1$ with $D=0$; whereas the right sub-figure plots $f(\mathbf{x}_1,\mathbf{x}_2)$ with $D=\mathrm{diag}\{5,1\}$.

must be a unit dominant eigenvector of $E(\mathbf{x}^*)$; for the special case $D=\mu W$ $(\mu>0)$, further, we are able to completely characterize the global solutions set as the join of an eigenspace corresponding to the largest eigenvalue of $E(\mathbf{x}^*)$ and $\mathscr{M}$. These global optimality conditions are useful in finding a global maximizer because it provides a starting point strategy (section 3.3) for any monotonically convergent iterative algorithm.

As our optimality conditions indicate that problem (1.1) is not simply related with the standard eigenvalue problem or the generalized eigenvalue problem, sophisticated algorithms such as the QR iteration (see e.g., [22,34]) or the implicitly restarted Arnoldi method (see e.g., [22,27,28]) and the QZ algorithm (see e.g., [21]) for the generalized eigenvalue problem could not be straightforwardly applied to solve (1.1). Furthermore, because the cost function $f(\mathbf{x})$ is not concave and the constraint $\mathscr{M}$ is not convex, finding a global solution is a hard problem as the iteration may be trapped by local maxima or saddle points. Knowledge of the perturbation property and the characterization of a solution therefore is important, which could be helpful in increasing the probability for obtaining a global maximizer.

The purpose of our second part of the paper is to suggest, by making use of the structure of the cost function and the constraint, an efficient algorithm for solving (1.1). Viewing $f|_{\mathscr{M}}(\mathbf{x})$ as a smooth function defined on the smooth manifold $\mathscr{M}$, the recently-proposed optimization methods on Riemannian manifolds are appropriate. The Riemannian Newton method [5,7,13] is attractive for its local quadratic convergence; however, the method loses the global convergence and is easily attracted by local minima, local maxima or even saddle points. The Riemannian trust-region (RTR) scheme [5,3,4] is an improvement of the Riemannian Newton method that addresses its two major drawbacks and is particularly suitable for our problem (1.1). In section 5, we realize the general RTR to maximize the cost function $f(\mathbf{x})$ on $\mathscr{M}$. The method basically follows the classical trust-region algorithm [10,30,35] by constructing a so-called Riemannian trust-region subproblem of (1.1) at each iteration. The subproblem could be efficiently solved by, for example, the *truncated conjugate-gradient* (tCG) method of Steihaug [35] and Toint [36], which

only requires the matrix-vector product and therefore, is appropriate for large-scale problems. The RTR algorithm for (1.1) needs the first information for global convergence to a critical point, and requires a "sufficiently good" approximation of second order for fast local convergence. Using the gradient and the Hessian derived in section 3, it is not difficult to implement the RTR algorithm for (1.1). We will show that the resulting RTR algorithm enjoys good global convergence properties with fast local convergence. Furthermore, the values of the cost function are nondecreasing which makes our starting point strategy (section 3) applicable. Our numerical experiment indicates that, though we cannot guarantee that any convergent point is a global solution of (1.1), the framework of RTR and the numerical implementation (the truncated conjugate-gradient) of the Riemannian trust-region subproblem provide a good approach for capturing a global solution.

The structure of the paper is as follows. In section 2, we first present a practical application of problem (1.1) for obtaining the sparse Fisher discriminant vector in pattern recognition. In section 3, we intend to characterize the local and the global solutions by investigating the optimality conditions. These conditions imply that the problem (1.1) is related with some special extreme nonlinear eigenvalue problem. Section 4 then gives some perturbation properties of (1.1) which are useful for designing the stopping criterion for certain iterative algorithms. The Riemannian trust-region method for (1.1) will be presented in section 5, and some preliminary and empirical evaluation of its performance is reported in section 6. We conclude the paper by drawing some final remarks in section 7.

Notation. Throughout the paper, all vectors are column vectors and are typeset in bold, and $\mathbf{x}_i$ represents the ith component of $\mathbf{x}$. For a matrix $A \in \mathbb{R}^{n\times n}$, $A^\top$ denotes its transpose and $\mathrm{tr}(A)$ stands for the trace of A. The symbols $\mathbb{S}_n$ and $\mathbb{S}_n^{++}$ stand for the set of symmetric and symmetric positive definite matrices of size n-by-n, respectively. The range and kernel of A are $\mathscr{R}(A) := \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} = A\mathbf{y} \text{ for some } \mathbf{y} \in \mathbb{R}^n\}$ and $\mathrm{Ker}(A) := \{\mathbf{y} \in \mathbb{R}^n | A\mathbf{y} = \mathbf{0}\}$, respectively. In addition, we use

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$$

to denote the ordered eigenvalues of A when they are all real.

2 An application: the sparse Fisher discriminant analysis

In this section, we will discuss some applications where the problem (1.1) arises. First, we note that (1.1) is in essence equivalent to the following problem (2.1) for matrices $\widehat{B}, \widehat{D} \in \mathbb{S}_n$ and $\widehat{W}, G \in \mathbb{S}_n^{++}$:

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top \widehat{B}\mathbf{x}}{\mathbf{x}^\top \widehat{W}\mathbf{x}} + \frac{\mathbf{x}^\top \widehat{D}\mathbf{x}}{\mathbf{x}^\top G\mathbf{x}}. \tag{2.1}$$

Such kind of problem can arise in the downlink of a multi-user MIMO system [33]. To see the equivalence, by noting that the objective function of (2.1) is invariant with respect to non-zero scaling of $\mathbf{x}$, one can define $\mathbf{y} = G^{\frac{1}{2}}\mathbf{x}$ and impose $\|\mathbf{y}\|_2 = 1$ to equivalently transform (2.1) to the form (1.1).

Besides the above equivalent problem, we will introduce a specific application of problem (1.1) arising from the feature extraction and dimensionality reduction. High dimensional data arises frequently from many modern applications of data mining, machine learning, bioinformatics and others. The linear

discriminant analysis (LDA) is one of the most efficient statistical approaches for supervised dimensionality reduction and classification (see e.g., [8,11,15,16,18,24,38–40]).

Suppose there are c independent classes, and for the ith class $(i = 1,2,\ldots,c)$, we have n_i samples, each of which is represented by a high dimensional vector in $\mathbb{R}^n$. Thus these samples form a data matrix $A \in \mathbb{R}^{n\times d}$, where $d = \sum_{i=1}^{c} n_i$. The basic idea in the Fisher-LDA is to find a matrix $X \in \mathbb{R}^{n\times l}$ (generally $l \ll n$), so that each original sample $\mathbf{a} \in \mathbb{R}^n$ is mapped to a new but reduced 'sample' $\mathbf{y}$ via

$$\mathbf{y} = X^\top \mathbf{a} \in \mathbb{R}^l.$$

The principle in defining an 'optimal' X [18] is to simultaneously maximize the *between-class separation*, measured by $\mathrm{tr}(X^\top S_b X)$, and minimize the *within-class cohesion*, measured by $\mathrm{tr}(X^\top S_w X)$. The matrices $S_b, S_w \in \mathbb{R}^{n\times n}$ are both symmetric and positive semidefinite formed from the data matrix A, which are called the *between-class scatter matrix* and the *within-class scatter matrix* [18], respectively. There are various criteria (see e.g., [11,15,16,18,24,38–40]) for obtaining X, and Foley and Sammon [16] suggest to find X column by column so that $X^\top X = I_l$. Thus, the first Foley-Sammon optimal discriminant vector is just the solution of Fisher criterion [15]:

$$\max_{\|\mathbf{x}\|_2=1} \frac{\mathbf{x}^\top S_b \mathbf{x}}{\mathbf{x}^\top S_w \mathbf{x}}. \tag{2.2}$$

The solution of (2.2) is an extreme eigenvector of the symmetric-definite matrix pencil[2] (S_b, S_w).

Besides the capability of classification, one always requires a sparse discriminant vector $\mathbf{x}$. This is because in high dimensional data, groups of objects often exist in subspaces rather than in the entire space, and therefore, each group is a set of objects identified by a subset of dimensions and different groups are represented in different subsets of dimensions (see e.g., [12,19,29,37]). This leads to a variable selection problem. Many variable selection techniques are implemented by using various penalty functions (see e.g., [14,19,29,25]). For the Fisher-LDA (2.2), this leads to the following problem:

$$\max_{\|\mathbf{x}\|_2=1} \frac{\mathbf{x}^\top S_b \mathbf{x}}{\mathbf{x}^\top S_w \mathbf{x}} - \sum_{i=1}^{n} p_i(|\mathbf{x}_i|)\lambda_i, \tag{2.3}$$

where $p_i(\cdot)$ for $i = 1,2,\ldots,n$, are given nonnegative penalty functions and $\lambda_i > 0$ are the tuning parameters controlling the strength of the incentive for classification on less dimensions. The penalty functions $p_i(\cdot)$ and the parameters λ_i are not necessarily the same for all i. For ease of presentation, however, we can assume that they are identical for all i, and then we denote $p_i(\cdot)\lambda_i = p_\lambda(\cdot)$. A favorable choice of $p_\lambda(\cdot)$ is the L_1 penalty: $p_\lambda(|\mathbf{x}_i|) = \lambda|\mathbf{x}_i|$ (see e.g., [29,37]). Another popular one is the so-called smoothly clipped absolute deviation penalty (SCAD) proposed in [14], which is proved to be very effective in parametric models such as generalized linear models and robust regression models [14,25]. To solve (2.3), Fan and Li [14] propose a local quadratic approximation for the penalty function $p_\lambda(\cdot)$: Suppose $\mathbf{x}^{(0)}$ is an initial point. If $\mathbf{x}_i^{(0)}$ is very close to zero, then set $\mathbf{x}_i^{(1)} = 0$; otherwise, the penalty function $p_\lambda(|\mathbf{x}_i|)$ is locally approximated by a quadratic function, i.e.,

$$p_\lambda(|\mathbf{x}_i|) \approx p_\lambda(|\mathbf{x}_i^{(0)}|) + \frac{\mathbf{x}_i^2 - (\mathbf{x}_i^{(0)})^2}{2|\mathbf{x}_i^{(0)}|} p'_\lambda(|\mathbf{x}_i^{(0)}|_+),$$

[2] The matrix S_w can be assumed to be positive definite as the singularity in S_w (i.e., the undersampled problem) can be handled by, for example, the regularization [17].

where $p'_\lambda(|\mathbf{x}_i^{(0)}|_+)$ denotes the derivative of $p_\lambda(\cdot)$ at $|\mathbf{x}_i^{(0)}|$ from above. As a result, a local approximation for (2.3) is

$$\max_{\|\mathbf{x}\|_2=1} \frac{\mathbf{x}^\top S_b \mathbf{x}}{\mathbf{x}^\top S_w \mathbf{x}} - \mathbf{x}^\top \Sigma_0 \mathbf{x} + c_0, \tag{2.4}$$

where $\Sigma_0 := \mathrm{diag}\{\frac{p'_\lambda(|\mathbf{x}_1^{(0)}|_+)}{2|\mathbf{x}_1^{(0)}|}, \ldots, \frac{p'_\lambda(|\mathbf{x}_n^{(0)}|_+)}{2|\mathbf{x}_n^{(0)}|}\}$, and c_0 is a constant. The resulting problem (2.4) consequently is an application of (1.1). For the discussion on the convergence of such approximation as well as more variable selection techniques, please refer to [12, 14, 19, 25, 29, 37].

3 Characterization of the solution

Returning to the problem (1.1), in this section, we attempt to shed some lights on the solution of (1.1) by describing the optimality conditions. Let $f_{|\mathscr{M}}(\mathbf{x}) : \mathscr{M} \to \mathbb{R}$ be the restriction of the cost function $f(\mathbf{x})$, and we will provide the local optimality conditions in section 3.1 and present some global optimality conditions in section 3.2. These optimality conditions not only characterize the solutions of (1.1) to some extent, but also motivate a starting point strategy in section 3.3 for certain iterative algorithm.

3.1 Optimality conditions for the local solution

Viewing the unit sphere $\mathscr{M}$ as a Riemannian submanifold of the Euclidean space $\mathbb{R}^n$ endowed with the natural inner product, the tangent space $\mathscr{T}_\mathbf{x}\mathscr{M}$ at any point $\mathbf{x} \in \mathscr{M}$ is given by [5]

$$\mathscr{T}_\mathbf{x}\mathscr{M} = \{\mathbf{z} | \mathbf{z} = P_\mathbf{x}\mathbf{y},\ \forall \mathbf{y} \in \mathbb{R}^n\},$$

where $P_\mathbf{x} = I_n - \mathbf{x}\mathbf{x}^\top$ is the orthogonal projection onto $\mathrm{Ker}(\mathbf{x}^\top) = \mathscr{R}(\mathbf{x})^\perp$. Now, for the smooth real-valued function $f_{|\mathscr{M}}(\mathbf{x})$ defined on $\mathscr{M}$, the gradient at $\mathbf{x} \in \mathscr{M}$ can be expressed by

$$g(\mathbf{x}) := \mathrm{grad}\, f_{|\mathscr{M}}(\mathbf{x}) = P_\mathbf{x} \nabla f(\mathbf{x}), \tag{3.1}$$

and a critical point $\mathbf{x} \in \mathscr{M}$ of $f_{|\mathscr{M}}(\mathbf{x})$, i.e., a KKT point of (1.1), is a point satisfying $g(\mathbf{x}) = \mathbf{0}$. Therefore, we have the following first-order optimality condition:

Theorem 3.1 *Let $B, D \in \mathbb{S}_n$ and $W \in \mathbb{S}_n^{++}$. A point $\mathbf{x} \in \mathscr{M}$ is a critical point of $f_{|\mathscr{M}}(\mathbf{x})$ on $\mathscr{M}$ if and only if it satisfies*

$$E(\mathbf{x})\mathbf{x} = \phi_d(\mathbf{x})\phi_w(\mathbf{x})\mathbf{x}, \tag{3.2}$$

where

$$E(\mathbf{x}) := B - \frac{\phi_b(\mathbf{x})}{\phi_w(\mathbf{x})} W + \phi_w(\mathbf{x}) D, \tag{3.3}$$

and $\phi_b(\mathbf{x}) := \mathbf{x}^\top B \mathbf{x}$, $\phi_w(\mathbf{x}) := \mathbf{x}^\top W \mathbf{x}$ and $\phi_d(\mathbf{x}) := \mathbf{x}^\top D \mathbf{x}$.

Proof From (3.1) and the expression of

$$\nabla f(\mathbf{x}) = 2(\frac{B\phi_w(\mathbf{x}) - W\phi_b(\mathbf{x})}{\phi_w^2(\mathbf{x})} + D)\mathbf{x},$$

we know that $\mathbf{x} \in \mathscr{M}$ is a critical point if and only if $g(\mathbf{x}) = \mathbf{0}$ where

$$g(\mathbf{x}) = \mathrm{grad}\, f_{|\mathscr{M}}(\mathbf{x}) = P_{\mathbf{x}}\nabla f(\mathbf{x}) = 2[\frac{B\phi_w(\mathbf{x}) - W\phi_b(\mathbf{x})}{\phi_w^2(\mathbf{x})} + D - \phi_d(\mathbf{x})I_n]\mathbf{x}. \tag{3.4}$$

By the positive definiteness of W, the conclusion follows.

The specially-defined matrix $E(\mathbf{x})$ plays an important role in characterizing the solutions of (1.1). Theorem 3.1 first implies that any critical point $\overline{\mathbf{x}}$ is an eigenvector of $E(\overline{\mathbf{x}})$ with $\phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})$ as the corresponding eigenvalue. That is, all critical points of $f_{|\mathscr{M}}(\mathbf{x})$ forms the set

$$\mathscr{S} = \{\|\mathbf{x}\|_2 = 1 | E(\mathbf{x})\mathbf{x} = \phi_d(\mathbf{x})\phi_w(\mathbf{x})\mathbf{x}\}. \tag{3.5}$$

This also implies that finding a global maximizer of (1.1) could not simply be obtained by solving a standard eigenvalue or a generalized eigenvalue problem, as the matrix $E(\mathbf{x})$ is itself dependent on $\mathbf{x}$. On the other hand, for very simple cases, Example 1.1 for instant, we can compute all critical points of $f_{|\mathscr{M}}(\mathbf{x})$ directly based on (3.2). For the general case, however, solving a nonlinear system (3.2) is another difficult problem.

There is another interesting question arising from Theorem 3.1. Suppose $\overline{\mathbf{x}}$ is a local or a global maximizer, then is $(\phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}}), \overline{\mathbf{x}})$ a dominant eigenpair of $E(\overline{\mathbf{x}})$? We will make some efforts on this problem. In particular, we will show that if $\overline{\mathbf{x}}$ is a local maximizer of (1.1), $\phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})$ is either *the largest* or *the second largest* eigenvalue of $E(\overline{\mathbf{x}})$, whereas if $\overline{\mathbf{x}}$ is a global maximizer, it must be a dominant eigenvector of $E(\overline{\mathbf{x}})$. To prove these results, we next establish the second-order optimality conditions, where the Hessian of $f_{|\mathscr{M}}(\mathbf{x})$ is required. For $f_{|\mathscr{M}}(\mathbf{x})$, defining the symmetric Hessian operator at $\mathbf{x} \in \mathscr{M}$

$$\mathrm{Hess}\, f_{|\mathscr{M}}(\mathbf{x}) : \mathscr{T}_{\mathbf{x}}\mathscr{M} \to \mathscr{T}_{\mathbf{x}}\mathscr{M} : \mathbf{h} \mapsto \nabla_{\mathbf{h}}\mathrm{grad}\ f_{|\mathscr{M}}(\mathbf{x})$$

involves the so-called *affine connection* ∇ (see [2,5]). A natural and preferable choice of affine connection is the *Riemannian connection*, as it possesses distinctive properties ([5], §5.5) and significantly simplifies the analytical derivations. With the Riemannian connection, the action of Riemannian Hessian of $f_{|\mathscr{M}}(\mathbf{x})$ on a tangent vector $\mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathscr{M}$ can be expressed by (see [5], Definition 5.5.1 and Proposition 5.3.2)

$$\mathrm{Hess}\ f_{|\mathscr{M}}(\mathbf{x})[\mathbf{h}] = P_{\mathbf{x}}(\mathbf{D}\mathrm{grad}\, f_{|\mathscr{M}}(\mathbf{x})[\mathbf{h}]) = P_{\mathbf{x}}(\mathbf{D}g(\mathbf{x})[\mathbf{h}]),\ \ \forall \mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathscr{M}, \tag{3.6}$$

where $\mathbf{D}g(\mathbf{x})[\mathbf{h}]$ represents the derivative at $\mathbf{x} \in \mathscr{M}$ along $\mathbf{h}$. From (3.4), by calculations, one can get that

$$\begin{aligned}\mathbf{D}g(\mathbf{x})[\mathbf{h}] = \ & \frac{2}{\phi_w(\mathbf{x})}[E(\mathbf{x}) - \phi_d(\mathbf{x})\phi_w(\mathbf{x})I_n + \frac{4\phi_b(\mathbf{x})}{\phi_w^2(\mathbf{x})}W\mathbf{x}\mathbf{x}^\top W \\ & - \frac{2}{\phi_w(\mathbf{x})}(B\mathbf{x}\mathbf{x}^\top W + W\mathbf{x}\mathbf{x}^\top B)]\mathbf{h} - 4(\mathbf{x}^\top D\mathbf{h})\mathbf{x}.\end{aligned} \tag{3.7}$$

Based on this expression, we can establish the second-order optimality conditions.

Theorem 3.2 *Let $B, D \in \mathbb{S}_n$ and $W \in \mathbb{S}_n^{++}$. Then*

(i) the Hessian operator of $f_{|\mathscr{M}}(\mathbf{x})$ *at point* $\mathbf{x}\in\mathscr{M}$ *acting on* $\forall\mathbf{h}\in\mathscr{T}_{\mathbf{x}}\mathscr{M}$ *is*

$$\mathrm{Hess}\, f_{|\mathscr{M}}(\mathbf{x})[\mathbf{h}] = \frac{2}{\phi_w(\mathbf{x})}H(\mathbf{x})\mathbf{h},$$

where $H(\mathbf{x})\in\mathbb{S}_n$ *is given by*

$$\begin{aligned} H(\mathbf{x}) = P_{\mathbf{x}}[&E(\mathbf{x}) - \phi_d(\mathbf{x})\phi_w(\mathbf{x})I_n + \frac{4\phi_b(\mathbf{x})}{\phi_w^2(\mathbf{x})}W\mathbf{x}\mathbf{x}^\top W \\ &- \frac{2}{\phi_w(\mathbf{x})}(B\mathbf{x}\mathbf{x}^\top W + W\mathbf{x}\mathbf{x}^\top B)]P_{\mathbf{x}}; \end{aligned} \tag{3.8}$$

(ii) if $\overline{\mathbf{x}}$ *is a local maximizer of* (1.1)*, then the matrix*

$$K(\overline{\mathbf{x}}) := E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n + 2P_{\overline{\mathbf{x}}}(D\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W + W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top D)P_{\overline{\mathbf{x}}} \tag{3.9}$$

is negative semidefinite;

(iii) for any $\overline{\mathbf{x}}\in\mathscr{S}$*, if* $K(\overline{\mathbf{x}}):\mathscr{T}_{\mathbf{x}}\mathscr{M}\to\mathscr{T}_{\mathbf{x}}\mathscr{M}$ *is negative definite, where* $\mathscr{S}$ *and* $K(\overline{\mathbf{x}})$ *are given by* (3.5) *and* (3.9)*, respectively, then* $\overline{\mathbf{x}}$ *is a strictly local maximizer of* (1.1)*.*

Proof Based on (3.6), (3.7), $P_{\mathbf{x}}\mathbf{x}=\mathbf{0}$, and $P_{\mathbf{x}}\mathbf{h}=\mathbf{h}$ for any $\mathbf{h}\in\mathscr{T}_{\mathbf{x}}\mathscr{M}$, *(i)* follows. For *(ii)*, we note if $\overline{\mathbf{x}}$ is a local maximizer of (1.1), then

$$E(\overline{\mathbf{x}})\overline{\mathbf{x}} = (B - \frac{\phi_b(\overline{\mathbf{x}})}{\phi_w(\overline{\mathbf{x}})}W + \phi_w(\overline{\mathbf{x}})D)\overline{\mathbf{x}} = \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})\overline{\mathbf{x}}.$$

Therefore,

$$P_{\overline{\mathbf{x}}}[E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n]P_{\overline{\mathbf{x}}} = E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n, \tag{3.10}$$

and

$$\begin{aligned} &\frac{4\phi_b(\overline{\mathbf{x}})}{\phi_w^2(\overline{\mathbf{x}})}W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W - \frac{2}{\phi_w(\overline{\mathbf{x}})}(B\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W + W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B) \\ &= \frac{2}{\phi_w(\overline{\mathbf{x}})}W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top(\frac{\phi_b(\overline{\mathbf{x}})}{\phi_w(\overline{\mathbf{x}})}W - B) + \frac{2}{\phi_w(\overline{\mathbf{x}})}(\frac{\phi_b(\overline{\mathbf{x}})}{\phi_w(\overline{\mathbf{x}})}W - B)\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W \\ &= 2W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top(D - \phi_d(\overline{\mathbf{x}})I_n) + 2(D - \phi_d(\overline{\mathbf{x}})I_n)\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W \\ &= 2(W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top D + D\overline{\mathbf{x}}\overline{\mathbf{x}}^\top W) - 2\phi_d(W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top + \overline{\mathbf{x}}\overline{\mathbf{x}}^\top W). \end{aligned} \tag{3.11}$$

Consequently, from (3.10), (3.11) and

$$P_{\overline{\mathbf{x}}}(W\overline{\mathbf{x}}\overline{\mathbf{x}}^\top + \overline{\mathbf{x}}\overline{\mathbf{x}}^\top W)P_{\overline{\mathbf{x}}} = 0,$$

we have $H(\overline{\mathbf{x}}) = K(\overline{\mathbf{x}})$. Based on the projected Hessian technique developed by [9] (see also [5, 13, 23]), a necessary condition for $\overline{\mathbf{x}}$ to be a local maximizer is that the Hessian $H(\overline{\mathbf{x}})$ is negative semidefinite on $\mathscr{T}_{\overline{\mathbf{x}}}\mathscr{M}$. Now, let $[\overline{\mathbf{x}}, Q]\in\mathbb{R}^{n\times n}$ be an orthogonal basis of $\mathbb{R}^n$. Note that for any $\mathbf{y}\in\mathbb{R}^n$, there are $\tau\in\mathbb{R}$ and $\mathbf{b}\in\mathbb{R}^{n-1}$ so that $\mathbf{y} = \overline{\mathbf{x}}\tau + Q\mathbf{b}$; moreover, from $K(\overline{\mathbf{x}})\overline{\mathbf{x}} = \mathbf{0}$, we have

$$\mathbf{y}^\top K(\overline{\mathbf{x}})\mathbf{y} = \tau^2\overline{\mathbf{x}}^\top K(\overline{\mathbf{x}})\overline{\mathbf{x}} + 2\tau\mathbf{b}^\top Q^\top K(\overline{\mathbf{x}})\overline{\mathbf{x}} + \mathbf{b}^\top Q^\top K(\overline{\mathbf{x}})Q\mathbf{b} = \mathbf{b}^\top Q^\top K(\overline{\mathbf{x}})Q\mathbf{b} \le 0,$$

where the last inequality follows from $Q\mathbf{b}\in\mathscr{T}_{\overline{\mathbf{x}}}\mathscr{M}$. This implies that $K(\overline{\mathbf{x}})$ is a negative semidefinite matrix, and *(ii)* is true. Lastly, the part *(iii)* is a sufficient condition for $\overline{\mathbf{x}}$ to be a strictly local maximizer of (1.1).

Theorem 3.2 reveals some information about the eigenvalue of $E(\overline{\mathbf{x}})$ whenever $\overline{\mathbf{x}}$ is a local or a global maximizer of (1.1). Particularly, we have

Theorem 3.3 *Let $B,D \in \mathbb{S}_n$ and $W \in \mathbb{S}_n^{++}$. If $\overline{\mathbf{x}}$ is a local maximizer of* (1.1)*, then it must be a unit eigenvector corresponding to either the largest or the second largest eigenvalue of $E(\overline{\mathbf{x}})$.*

Proof We prove the result according to *(ii)* of Theorem 3.2. Denote

$$\mathbf{u} = 2P_{\overline{\mathbf{x}}}D\overline{\mathbf{x}} \quad \text{and} \quad \mathbf{v} = P_{\overline{\mathbf{x}}}W\overline{\mathbf{x}},$$

then we have

$$K(\overline{\mathbf{x}}) = E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n + (\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top).$$

Note that the rank of the matrix $\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top \in \mathbb{S}_n$ is at most 2, and it is easy to check that its only possible nonzero eigenvalues are $\mathbf{u}^\top\mathbf{v} \pm \|\mathbf{u}\|_2\|\mathbf{v}\|_2$ with the corresponding eigenvectors $\mathbf{u} \pm \frac{\|\mathbf{u}\|_2}{\|\mathbf{v}\|_2}\mathbf{v}$, respectively. Therefore,

$$\lambda_1(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) \geq 0 = \lambda_2(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) = \cdots = \lambda_{n-1}(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) \geq \lambda_n(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top).$$

On the other hand, according to *(ii)* of Theorem 3.2, since $K(\overline{\mathbf{x}})$ is negative semidefinite, by Weyl's monotonicity principle, we know that

$$\begin{cases} \lambda_1(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) \leq \lambda_1(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top), \\ \lambda_2(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) \leq \lambda_2(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) = 0, \\ \qquad\qquad \vdots \\ \lambda_n(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) \leq \lambda_n(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top). \end{cases} \tag{3.12}$$

Moreover, because $(0,\overline{\mathbf{x}})$ is an eigenpair of $E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n$, the relation (3.12) implies the following three possible cases:

(i) $\lambda_1(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) = 0 > \lambda_2(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n)$,
(ii) $\lambda_1(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) = 0 = \lambda_2(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n)$, and
(iii) $\lambda_1(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n) > 0 = \lambda_2(E(\overline{\mathbf{x}}) - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})I_n)$.

The first two cases imply that $\phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})$ is the largest eigenvalue of $E(\overline{\mathbf{x}})$, whereas the last case implies that $\phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})$ is the second largest eigenvalue of $E(\overline{\mathbf{x}})$. This completes the proof.

It deserves to point out that it is possible that a local non-global maximizer $\overline{\mathbf{x}}$ is indeed an eigenvector corresponding to the second largest eigenvalue of $E(\overline{\mathbf{x}})$. An illustrative example will be presented in section 3.2. However, for some special cases, we can show that, $\overline{\mathbf{x}}$ must be a dominant eigenvector of $E(\overline{\mathbf{x}})$ whenever $\overline{\mathbf{x}}$ is a local maximizer of (1.1).

Theorem 3.4 *Let $B \in \mathbb{S}_n$ and $W \in \mathbb{S}_n^{++}$. Then*

(i) if $D = 0$, then there is no local non-global maximizer for (1.1)*; that is, any local maximizer is a global maximizer;*

(ii) *if* $D=\mu W$ $(\mu>0)$, *then* $\phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})$ *must be the largest eigenvalue of* $E(\bar{\mathbf{x}})$ *for any local maximizer* $\bar{\mathbf{x}}$ *of* (1.1).

Proof The conclusion in part *(i)* is already known (see e.g., [20]). To prove *(ii)*, we suppose $\phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})$ is not the largest eigenvalue of $E(\bar{\mathbf{x}})$, and then assume $\mathbf{y}$ is a dominant unit eigenvector. Therefore, according to Theorem 3.2 again, one has

$$0\geq \mathbf{y}^\top K(\bar{\mathbf{x}})\mathbf{y}=\mathbf{y}^\top(E(\bar{\mathbf{x}})-\phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})I_n)\mathbf{y}+4\mu\|\mathbf{y}^\top P_{\bar{\mathbf{x}}}W\bar{\mathbf{x}}\|_2^2,$$

yielding

$$\mathbf{y}^\top E(\bar{\mathbf{x}})\mathbf{y}-\phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})\leq -4\mu\|\mathbf{y}^\top P_{\bar{\mathbf{x}}}W\bar{\mathbf{x}}\|_2^2\leq 0,$$

which is a contradiction because $\mathbf{y}$ is a dominant unit eigenvector of $E(\bar{\mathbf{x}})$, while $\phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})$ is not the largest eigenvalue of $E(\bar{\mathbf{x}})$. Thus, we conclude that in this case, $\bar{\mathbf{x}}$ must be a dominant eigenvector of $E(\bar{\mathbf{x}})$.

Unlike the case *(i)*, for the case $D=\mu W$ $(\mu>0)$, we point out that we cannot exclude local non-global maximizer. A simple illustrative example is demonstrated.

Example 3.1 We let $B=\mathrm{diag}\{1,9,2\}, W=D=\mathrm{diag}\{5,2,3\}$. Then according to Theorem 3.1, we know that $\bar{\mathbf{x}}=[1,0,0]^\top$ and $\mathbf{x}^*=[0,1,0]^\top$ are critical points. Furthermore, by calculation, one has that

$$E(\bar{\mathbf{x}})=\mathrm{diag}\{25,18.6,16.4\},\quad H(\bar{\mathbf{x}})=\mathrm{diag}\{0,-6.4,-8.6\},$$

and

$$E(\mathbf{x}^*)=\mathrm{diag}\{-11.5,4,-5.5\},\quad H(\mathbf{x}^*)=\mathrm{diag}\{-15.5,0,-9.5\}.$$

According to Theorem 3.2, we know that both $\bar{\mathbf{x}}$ and $\mathbf{x}^*$ are local maxima of (1.1), and they are both the dominant eigenvectors of the related matrices $E(\bar{\mathbf{x}})$ and $E(\mathbf{x}^*)$. However, $f(\bar{\mathbf{x}})=5.2<f(\mathbf{x}^*)=6.5$, which implies that $\bar{\mathbf{x}}$ is a local non-global maximizer. Figure 1 plots the cost function $f(\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta))$ in the 2π-period $\theta-\eta$ plane with $0\leq\theta\leq 2\pi,\ 0\leq\eta\leq 2\pi$, where $\mathbf{x}_1(\theta,\eta)=\cos(\theta)\cos(\eta), \mathbf{x}_2(\theta,\eta)=\cos(\theta)\sin(\eta), \mathbf{x}_3(\theta)=\sin(\theta)$. From Figure 1, we easily find that $\mathbf{x}^*$ is a global maximizer.

3.2 Optimality condition for the global solution

This subsection focuses on some necessary global optimality conditions for (1.1). In particular, we shall prove two results: the first claims that $\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*)$ must be the largest eigenvalue of $E(\mathbf{x}^*)$ if $\mathbf{x}^*$ is a global maximizer; the second one says that, for the case $D=\mu W$ $(\mu>0)$, the global solutions set of (1.1) is essentially an eigenspace corresponding to the largest eigenvalue of $E(\mathbf{x}^*)$. The following lemma is important for this conclusion.

Lemma 3.1 *Let* $B,D\in\mathbb{S}_n$ *and* $W\in\mathbb{S}_n^{++}$. *Then for any two points* $\mathbf{z},\mathbf{x}\in\mathscr{M}$, *we have*

$$f(\mathbf{z})-f(\mathbf{x})=\frac{(\mathbf{z}^\top E(\mathbf{x})\mathbf{z}-\phi_d(\mathbf{x})\phi_w(\mathbf{x}))+(\phi_w(\mathbf{z})-\phi_w(\mathbf{x}))(\phi_d(\mathbf{z})-\phi_d(\mathbf{x}))}{\phi_w(\mathbf{z})}. \tag{3.13}$$

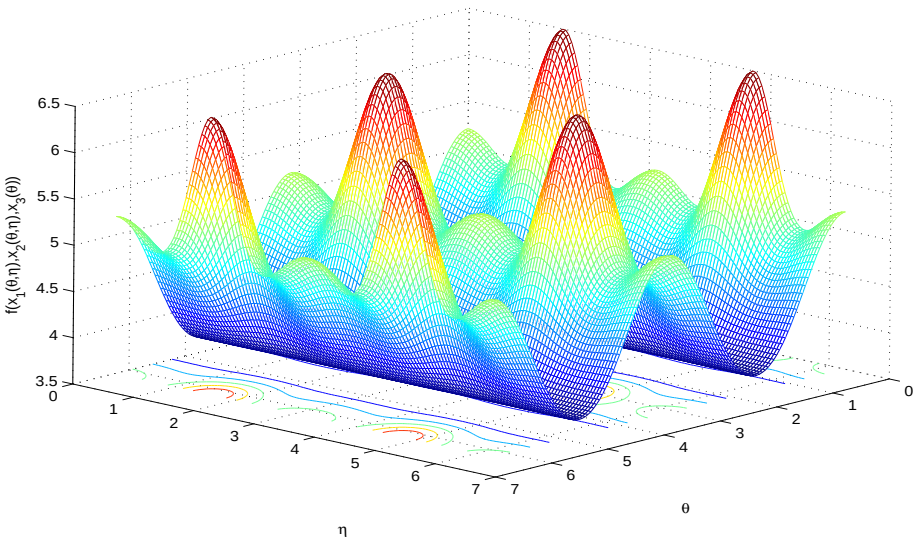

Fig. 1 This figure plots the cost function $f(\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta))$ of Example 3.1 in the 2π-period $\theta-\eta$ plane with $0\le\theta\le 2\pi$, $0\le\eta\le 2\pi$, where $\mathbf{x}_1(\theta,\eta)=\cos(\theta)\cos(\eta),\mathbf{x}_2(\theta,\eta)=\cos(\theta)\sin(\eta),\mathbf{x}_3(\theta)=\sin(\theta)$.

Proof According to the definition of $E(\mathbf{x})$, we know that

$$\begin{aligned}\mathbf{z}^\top E(\mathbf{x})\mathbf{z}-\phi_d(\mathbf{x})\phi_w(\mathbf{x}) &= \phi_b(\mathbf{z})-\frac{\phi_b(\mathbf{x})}{\phi_w(\mathbf{x})}\phi_w(\mathbf{z})+\phi_w(\mathbf{x})\phi_d(\mathbf{z})-\phi_w(\mathbf{x})\phi_d(\mathbf{x})\\ &= \phi_w(\mathbf{z})(\frac{\phi_b(\mathbf{z})}{\phi_w(\mathbf{z})}-\frac{\phi_b(\mathbf{x})}{\phi_w(\mathbf{x})})+\phi_w(\mathbf{x})(\phi_d(\mathbf{z})-\phi_d(\mathbf{x}))\\ &= \phi_w(\mathbf{z})(f(\mathbf{z})-f(\mathbf{x}))-(\phi_w(\mathbf{z})-\phi_w(\mathbf{x}))(\phi_d(\mathbf{z})-\phi_d(\mathbf{x})),\end{aligned}$$

which then leads to our assertion.

Theorem 3.5 *Let $B\in\mathbb{S}_n$ and $W,D\in\mathbb{S}_n^{++}$. Then for any global maximizer $\mathbf{x}^*$ of (1.1), $(\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*),\mathbf{x}^*)$ must be an eigenpair corresponding to the largest eigenvalue of $E(\mathbf{x}^*)$.*

Proof We prove it by contradiction. Suppose it is not true, then we assume that $\mathbf{y}$ is a dominant eigenvector of $E(\mathbf{x}^*)$, and hence it is true that

$$\mathbf{y}^\top E(\mathbf{x}^*)\mathbf{y}-\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*)>0,\quad \mathbf{y}^\top\mathbf{x}^*=0,\quad \text{and}\quad P_{\mathbf{x}^*}\mathbf{y}=\mathbf{y}. \tag{3.14}$$

Moreover, by *(ii)* of Theorem 3.2, one has

$$0\ge\mathbf{y}^\top K(\mathbf{x}^*)\mathbf{y}=\mathbf{y}^\top E(\mathbf{x}^*)\mathbf{y}-\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*)+4(\mathbf{y}^\top D\mathbf{x}^*)(\mathbf{y}^\top W\mathbf{x}^*),$$

or equivalently,

$$4(\mathbf{y}^\top D\mathbf{x}^*)(\mathbf{y}^\top W\mathbf{x}^*)\le-\mathbf{y}^\top E(\mathbf{x}^*)\mathbf{y}+\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*)<0. \tag{3.15}$$

On the other hand, according to Lemma 3.1 with $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{z} = \mathbf{y}$, and the fact that $f(\mathbf{y}) - f(\mathbf{x}^*) \leq 0$ and (3.14), it must follow that

$$\delta(\mathbf{x}^*, \mathbf{y}) := (\phi_w(\mathbf{y}) - \phi_w(\mathbf{x}^*))(\phi_d(\mathbf{y}) - \phi_d(\mathbf{x}^*)) < 0, \tag{3.16}$$

which leads to the following two possible cases:

Case I.

$$\phi_w(\mathbf{y}) > \phi_w(\mathbf{x}^*) \quad \text{and} \quad \phi_d(\mathbf{y}) < \phi_d(\mathbf{x}^*),$$

Case II.

$$\phi_w(\mathbf{y}) < \phi_w(\mathbf{x}^*) \quad \text{and} \quad \phi_d(\mathbf{y}) > \phi_d(\mathbf{x}^*).$$

Now for Case I, we will show the contradiction by constructing a vector of the form $\mathbf{z} = \alpha\mathbf{x}^* + \beta\mathbf{y}$ so that

$$\|\mathbf{z}\|_2 = 1 \quad \text{and} \quad f(\mathbf{z}) > f(\mathbf{x}^*), \tag{3.17}$$

which implies that $\mathbf{x}^*$ is not a global maximizer.

For this purpose, we note first that, because $\mathbf{x}^*$ and $\mathbf{y}$ are both the eigenvectors of $E(\mathbf{x}^*)$, from (3.14), $\|\mathbf{z}\|_2 = 1$ is true if $\alpha^2 + \beta^2 = 1$; moreover, for any scalars α, β satisfying $\alpha^2 + \beta^2 = 1$, one has that

$$\mathbf{z}^\top E(\mathbf{x}^*)\mathbf{z} = \alpha^2(\mathbf{x}^*)^\top E(\mathbf{x}^*)\mathbf{x}^* + \beta^2\mathbf{y}^\top E(\mathbf{x}^*)\mathbf{y} > (\mathbf{x}^*)^\top E(\mathbf{x}^*)\mathbf{x}^* = \phi_w(\mathbf{x}^*)\phi_d(\mathbf{x}^*).$$

Therefore, based on (3.13), we know that (3.17) is fulfilled if we can construct a vector $\mathbf{z} = \alpha\mathbf{x}^* + \beta\mathbf{y} \in \mathscr{M}$ with $\phi_d(\mathbf{z}) = \phi_d(\mathbf{x}^*)$. To this end, we note that the condition $\phi_d(\mathbf{z}) = \phi_d(\mathbf{x}^*)$ yields a quadratic with respect to α :

$$\alpha^2\phi_d(\mathbf{x}^*) + 2\alpha\beta(\mathbf{y}^\top D\mathbf{x}^*) + \beta^2\phi_d(\mathbf{y}) - \phi_d(\mathbf{x}^*) = 0. \tag{3.18}$$

Denote $\gamma = \mathbf{y}^\top D\mathbf{x}^*$ which is nonzero by (3.15). Equation (3.18) has two distinctive roots because according to Case I and $\beta^2 \leq 1$,

$$\begin{aligned}\Delta &= 4\beta^2\gamma^2 - 4\phi_d(\mathbf{x}^*)(\beta^2\phi_d(\mathbf{y}) - \phi_d(\mathbf{x}^*)) \\ &> 4\beta^2\gamma^2 - 4\phi_d(\mathbf{x}^*)(\phi_d(\mathbf{y}) - \phi_d(\mathbf{x}^*)) > 0.\end{aligned}$$

Consider the root of (3.18)

$$\alpha(\beta) = \frac{-\beta\gamma + \sqrt{(\beta\gamma)^2 - \phi_d(\mathbf{x}^*)(\beta^2\phi_d(\mathbf{y}) - \phi_d(\mathbf{x}^*))}}{\phi_d(\mathbf{x}^*)}. \tag{3.19}$$

As we can freely choose the sign of β, we assume $\beta\gamma > 0$, and thus it is easy to check that

$$0 < \alpha(\beta) < \frac{-\beta\gamma + \sqrt{(\beta\gamma)^2 + \phi_d^2(\mathbf{x}^*) + 2\beta\gamma\phi_d(\mathbf{x}^*)}}{\phi_d(\mathbf{x}^*)} = 1.$$

We next seek a nonzero $|\beta| < 1$ so that

$$\alpha(\beta)^2 + \beta^2 = 1. \tag{3.20}$$

Plugging (3.19) into (3.20) yields

$$2\beta\gamma^2+\beta\phi_d(\mathbf{x}^*)(\phi_d(\mathbf{x}^*)-\phi_d(\mathbf{y}))=2\gamma\sqrt{(\beta\gamma)^2-\phi_d(\mathbf{x}^*)(\beta^2\phi_d(\mathbf{y})-\phi_d(\mathbf{x}^*))}. \tag{3.21}$$

Because $\beta\gamma>0$, (3.21) is satisfied if

$$\begin{aligned}&4\gamma^4\beta^2+\beta^2\phi_d^2(\mathbf{x}^*)(\phi_d(\mathbf{x}^*)-\phi_d(\mathbf{y}))^2+4\beta^2\gamma^2\phi_d(\mathbf{x}^*)(\phi_d(\mathbf{x}^*)-\phi_d(\mathbf{y}))\\=\;&4\gamma^2(\beta^2\gamma^2-\beta^2\phi_d(\mathbf{x}^*)\phi_d(\mathbf{z})+\phi_d^2(\mathbf{x}^*))\end{aligned}$$

is met. This leads to

$$\beta^2=\frac{4\gamma^2\phi_d^2(\mathbf{x}^*)}{\phi_d^2(\mathbf{x}^*)(\phi_d(\mathbf{x}^*)-\phi_d(\mathbf{y}))^2+4\phi_d^2(\mathbf{x}^*)\gamma^2}\in(0,1).$$

Therefore, we have constructed a vector $\mathbf{z}=\alpha(\beta)\mathbf{x}^*+\beta\mathbf{y}$ where

$$\beta=\frac{2\mathrm{sgn}(\gamma)|\gamma|\phi_d(\mathbf{x}^*)}{\sqrt{\phi_d^2(\mathbf{x}^*)(\phi_d(\mathbf{x}^*)-\phi_d(\mathbf{y}))^2+4\phi_d^2(\mathbf{x}^*)\gamma^2}}$$

and $\alpha(\beta)$ is given by (3.19) so that (3.17) is fulfilled, which is a contradiction.

An analogous argument with D replaced by W applies to the Case II, and we can again construct a new vector $\mathbf{z}$ so that (3.17) is fulfilled. This completes the proof.

To illustrate the difference between the optimality conditions for the global maximizer (Theorem 3.5) and for the local maximizer (Theorem 3.3), we present a numerical example below. In this example, we employ the Riemannian trust-region method (RTR) proposed in section 5 to find a local maximizer $\hat{\mathbf{x}}$, which is shown to be an eigenvector corresponding to the second largest eigenvalue of $E(\hat{\mathbf{x}})$, and hence is not a global solution for (1.1) according to Theorem 3.5. This serves as an illustration for Theorem 3.5 and Theorem 3.3 as well.

Example 3.2 In this example,

$$B=\begin{pmatrix}2.3969 & 0.4651 & 4.6392\\ 0.4651 & 5.4401 & 0.7838\\ 4.6392 & 0.7838 & 10.1741\end{pmatrix},W=\begin{pmatrix}0.8077 & 0.8163 & 1.0970\\ 0.8163 & 4.1942 & 0.8457\\ 1.0970 & 0.8457 & 1.8810\end{pmatrix},\quad\text{and}$$

$$D=\begin{pmatrix}3.9104 & -0.9011 & -2.0128\\ -0.9011 & 0.9636 & 0.6102\\ -2.0128 & 0.6102 & 1.0908\end{pmatrix}.$$

By parameterizing $\mathbf{x}=[\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta)]^\top$ via $\mathbf{x}_1(\theta,\eta)=\cos(\theta)\cos(\eta),\mathbf{x}_2(\theta,\eta)=\cos(\theta)\sin(\eta)$, and $\mathbf{x}_3(\theta)=\sin(\theta)$, the value $f(\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta))$ versus the parameters (θ,η) in the 2π-period $\theta-\eta$ plane with $0\le\theta\le2\pi,\ 0\le\eta\le2\pi$, is plotted in Figure 2.

It is observed that the global optimal value $f^*\approx11.2008$, and a local optimal value is near 7.9664. Now, starting from the initial point $\mathbf{x}^{(0)}=[0.9755,-0.0025,-0.2198]^\top$, the RTR method (Algorithm 2-1) proposed in section 5 converges to a point $\hat{\mathbf{x}}\approx[0.9692,-0.2410,-0.0507]^\top$ with residual $\|E(\hat{\mathbf{x}})\hat{\mathbf{x}}-$

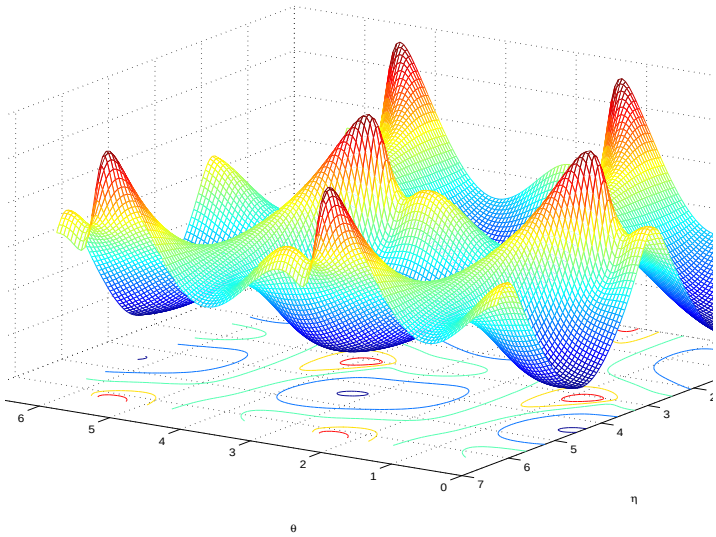

Fig. 2 This figure plots the cost function $f(\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta))$ of Example 3.2 in the 2π-period $\theta-\eta$ plane with $0\le\theta\le 2\pi$, $0\le\eta\le 2\pi$, where $\mathbf{x}_1(\theta,\eta)=\cos(\theta)\cos(\eta),\mathbf{x}_2(\theta,\eta)=\cos(\theta)\sin(\eta),\mathbf{x}_3(\theta)=\sin(\theta)$.

$\phi_d(\hat{\mathbf{x}})\phi_w(\hat{\mathbf{x}})\hat{\mathbf{x}}\|_2\approx 3.7\times 10^{-13}$ implying that $\hat{\mathbf{x}}$ is a critical point; moreover, we observe that the Hessian matrix $H(\hat{\mathbf{x}})$ at $\hat{\mathbf{x}}$ has eigenvalues $-12.5406,-4.1470$ and zero, which by *(iii)* of Theorem 3.2, implies that $\hat{\mathbf{x}}$ is a strictly local maximizer. On the other hand, we find that the eigenvalues of $E(\hat{\mathbf{x}})$ are $-10.1784,\ 2.3516,\ 4.2678$ and $\phi_d(\hat{\mathbf{x}})\phi_w(\hat{\mathbf{x}})\approx 2.3516$. This indicates that $\hat{\mathbf{x}}$ is only an eigenvector corresponding to the second largest eigenvalue of $E(\hat{\mathbf{x}})$, and by Theorem 3.5, we claim $\hat{\mathbf{x}}$ is not a global maximizer; indeed, $f(\hat{\mathbf{x}})\approx 7.9664<f^*$.

Theorem 3.5 serves as a necessary global optimality condition for (1.1), which only partially characterizes the set of global solutions to (1.1). In the next theorem, with the assumption $D=\mu W$ $(\mu>0)$, we are able to make the global solutions set $\mathscr{S}^*$ clear.

Theorem 3.6 *Let $B\in\mathbb{S}_n$ and $W\in\mathbb{S}_n^{++}$. If $D=\mu W$ $(\mu>0)$, then the set of all global maxima of* (1.1) *is*

$$\mathscr{S}^*=\mathscr{E}_1(E(\mathbf{x}^*))\cap\mathscr{M}, \tag{3.22}$$

where $\mathbf{x}^*$ is an arbitrary global maximizer and $\mathscr{E}_1(E(\mathbf{x}^*))$ stands for the eigenspace associated with the largest eigenvalue of $E(\mathbf{x}^*)$. Moreover, for any two global maxima $\mathbf{x}^*$ and $\mathbf{y}^*$, it follows that $\phi_b(\mathbf{x}^*)=\phi_b(\mathbf{y}^*)$ and $\phi_w(\mathbf{x}^*)=\phi_w(\mathbf{y}^*)$.

Proof Based on Theorem 3.5, we only need to show that any $\mathbf{z}\in\mathscr{S}^*$ is a global maximizer. To this end, with the assumption $D=\mu W$ $(\mu>0)$ and the fact $\mathbf{z}^\top E(\mathbf{x}^*)\mathbf{z}=\phi_d(\mathbf{x}^*)\phi_w(\mathbf{x}^*)$, we have from (3.13) that

$$f(\mathbf{z})-f(\mathbf{x}^*)=\mu\frac{(\phi_w(\mathbf{z})-\phi_w(\mathbf{x}^*))^2}{\phi_w(\mathbf{z})}\ge 0.$$

Since $\mathbf{x}^*$ is a global maximizer, the above implies that $f(\mathbf{z}) = f(\mathbf{x}^*)$ and $\phi_w(\mathbf{z}) = \phi_w(\mathbf{x}^*)$. Therefore, $\mathbf{z}$ is also a global maximizer, and $\phi_w(\mathbf{z}) = \phi_w(\mathbf{x}^*)$ consequently leads to $\phi_b(\mathbf{z}) = \phi_b(\mathbf{x}^*)$. The proof is completed.

To conclude this subsection, we draw some remarks for Theorem 3.6. First, this theorem says that in case of $D = \mu W$ ($\mu > 0$), for all global maxima, the values $\xi(\mathbf{x}) = \frac{\phi_b(\mathbf{x})}{\phi_w(\mathbf{x})} - \mu\phi_w(\mathbf{x})$ are equal to ξ^*, which is another identity of the global solution rather than the global optimal value f^*. Therefore, we can say that finding the global solution to (1.1) in this case is closely related with a special extreme eigenvalue problem of the associated matrix $B - \xi^* W$, and if we can successfully fix the value ξ^*, all global maxima can be obtained. Moreover, as we shall see in Section 4, the characterization of the global maxima in Theorem 3.6 is also crucial in deriving the perturbation property of (1.1) when the matrices B and W are perturbed locally.

3.3 A starting point strategy

As we have seen that local non-global maxima of (1.1) could not be generally ruled out, and therefore, we cannot guarantee that the computed solution of certain iterative algorithm is a global maximizer (see Example 3.2). However, since we have already established a necessary optimality condition for the global solution, one may expect that a reasonable method should be able to reach a solution satisfying at least Theorem 3.5. In this subsection, we will propose a starting point strategy for any iterative algorithm that is able to generate a sequence $\{\mathbf{x}^{(k)}\}$ converging to a critical point of $f_{|_{\mathscr{M}}}(\mathbf{x})$ with $\{f(\mathbf{x}^{(k)})\}$ monotonically increasing. By repeatedly applying our starting point strategy, we can *always* obtain a critical point $\overline{\mathbf{x}}$ so that it is a dominant eigenvector of $E(\overline{\mathbf{x}})$. The starting point strategy is based on our constructive proof for Theorem 3.5 and we state it below.

Recall that (3.13) of Lemma 3.1 reveals the following relation

$$f(\mathbf{z}) - f(\mathbf{x}) = \frac{\widehat{\delta}(\mathbf{x},\mathbf{z}) + \delta(\mathbf{x},\mathbf{z})}{\phi_w(\mathbf{z})}, \tag{3.23}$$

for any two points $\mathbf{x}, \mathbf{z} \in \mathscr{M}$, where $\delta(\mathbf{x},\mathbf{z})$ is given by (3.16) and

$$\widehat{\delta}(\mathbf{x},\mathbf{z}) := \mathbf{z}^\top E(\mathbf{x})\mathbf{z} - \phi_d(\mathbf{x})\phi_w(\mathbf{x}).$$

Suppose that $\hat{\mathbf{x}}$ is a critical point of $f_{|_{\mathscr{M}}}(\mathbf{x})$ obtained by a monotonically convergent algorithm, and that $\hat{\mathbf{x}}$ is not a dominant eigenvector of $E(\hat{\mathbf{x}})$. The underlying principle of our starting point strategy is to construct a new point $\mathbf{z} \in \mathscr{M}$ (as the next initial point) so that $\widehat{\delta}(\hat{\mathbf{x}},\mathbf{z}) > 0$ and $\delta(\hat{\mathbf{x}},\mathbf{z}) \geq 0$, which by (3.23) then implies $f(\mathbf{z}) > f(\hat{\mathbf{x}})$. Assume $\mathbf{y}$ is a dominant unit eigenvector of $E(\hat{\mathbf{x}})$. Then there are two mutually exclusive scenarios for the value of $\delta(\hat{\mathbf{x}},\mathbf{y})$:

1). $\delta(\hat{\mathbf{x}},\mathbf{y}) \geq 0$.
In this case, by (3.23) and $\widehat{\delta}(\hat{\mathbf{x}},\mathbf{y}) = \mathbf{y}^\top E(\hat{\mathbf{x}})\mathbf{y} - \phi_d(\hat{\mathbf{x}})\phi_w(\hat{\mathbf{x}}) > 0$, one can easily take $\mathbf{y}$ as a starting point of the iterative algorithm;

2). $\delta(\hat{\mathbf{x}},\mathbf{y}) < 0$.

In this case, we still have two possible cases. For

$$\text{Case I.} \qquad \phi_w(\mathbf{y}) > \phi_w(\hat{\mathbf{x}}) \quad \text{and} \quad \phi_d(\mathbf{y}) < \phi_d(\hat{\mathbf{x}}),$$

if $\gamma = \mathbf{y}^\top D\hat{\mathbf{x}} \neq 0$ which is satisfied whenever $\hat{\mathbf{x}}$ is a local maximizer (see (3.15)), then according to the proof of Theorem 3.5, the starting point could be $\mathbf{z} = \alpha(\beta)\hat{\mathbf{x}} + \beta\mathbf{y}$, where $\alpha(\beta)$ and β are given respectively by

$$\begin{aligned} \alpha(\beta) &= \frac{-\beta\gamma + \sqrt{(\beta\gamma)^2 - \phi_d(\hat{\mathbf{x}})(\beta^2\phi_d(\mathbf{y}) - \phi_d(\hat{\mathbf{x}}))}}{\phi_d(\hat{\mathbf{x}})}, \\ \beta &= \frac{2\text{sgn}(\gamma)|\gamma|\phi_d(\hat{\mathbf{x}})}{\sqrt{\phi_d^2(\hat{\mathbf{x}})(\phi_d(\hat{\mathbf{x}}) - \phi_d(\mathbf{y}))^2 + 4\phi_d^2(\hat{\mathbf{x}})\gamma^2}}; \end{aligned}$$

otherwise for

$$\text{Case II.} \qquad \phi_w(\mathbf{y}) < \phi_w(\hat{\mathbf{x}}) \quad \text{and} \quad \phi_d(\mathbf{y}) > \phi_d(\hat{\mathbf{x}}),$$

if $\tilde{\gamma} = \mathbf{y}^\top W\hat{\mathbf{x}} \neq 0$ which is satisfied whenever $\hat{\mathbf{x}}$ is a local maximizer (see (3.15)), then according to the proof of Theorem 3.5, the starting point could be $\tilde{\mathbf{z}} = \tilde{\alpha}(\tilde{\beta})\hat{\mathbf{x}} + \tilde{\beta}\mathbf{y}$, where $\tilde{\alpha}(\tilde{\beta})$ and $\tilde{\beta}$ are given respectively by

$$\begin{aligned} \tilde{\alpha}(\tilde{\beta}) &= \frac{-\tilde{\beta}\tilde{\gamma} + \sqrt{(\tilde{\beta}\tilde{\gamma})^2 - \phi_w(\hat{\mathbf{x}})(\tilde{\beta}^2\phi_w(\mathbf{y}) - \phi_w(\hat{\mathbf{x}}))}}{\phi_w(\hat{\mathbf{x}})}, \\ \tilde{\beta} &= \frac{2\text{sgn}(\tilde{\gamma})|\tilde{\gamma}|\phi_w(\hat{\mathbf{x}})}{\sqrt{\phi_w^2(\hat{\mathbf{x}})(\phi_w(\hat{\mathbf{x}}) - \phi_w(\mathbf{y}))^2 + 4\phi_w^2(\hat{\mathbf{x}})\tilde{\gamma}^2}}. \end{aligned}$$

We make a remark for this starting point strategy when the algorithm only produces an approximation, say $\check{\mathbf{x}} \in \mathscr{M}$, of the critical point $\hat{\mathbf{x}}$. Let $\check{\mathbf{y}}$ be the unit dominant eigenvector of $E(\check{\mathbf{x}})$ and hence $\widehat{\delta}(\check{\mathbf{x}},\check{\mathbf{y}}) > 0$. According to our starting strategy, for the first case $\delta(\check{\mathbf{x}},\check{\mathbf{y}}) \geq 0$, the new point is $\mathbf{z} = \check{\mathbf{y}}$, which by (3.23) satisfies $f(\check{\mathbf{y}}) > f(\check{\mathbf{x}})$. This implies that our starting point strategy in the first case does not depend on how accurately $\check{\mathbf{x}}$ approximates to the critical point $\hat{\mathbf{x}}$. For the second case when $\delta(\check{\mathbf{x}},\check{\mathbf{y}}) < 0$, according to our constructive procedure for the new point $\mathbf{z}$ (refer to the proof of Theorem 3.5) and by continuity, we know that there is a neighborhood $\mathscr{N}(\hat{\mathbf{x}})$ at $\hat{\mathbf{x}}$, such that for any $\check{\mathbf{x}} \in \mathscr{N}(\hat{\mathbf{x}})$, the constructed point $\mathbf{z}$ satisfies $\delta(\check{\mathbf{x}},\mathbf{z}) + \widehat{\delta}(\check{\mathbf{x}},\mathbf{z}) > 0$, which by (3.23) again implies that $f(\mathbf{z}) > f(\check{\mathbf{x}})$.

To show the effectiveness of this starting point strategy, we present the following numerical example.

Example 3.3 With the definition of B, W and D given in Example 3.2, we have known that the local maximizer $\hat{\mathbf{x}} = [0.9692, -0.2410, -0.0507]^\top$ is an eigenvector associated with the second largest eigenvalue of $E(\hat{\mathbf{x}})$. Therefore, based on our starting point strategy, we are able to improve the point $\hat{\mathbf{x}}$ to be new starting point for the RTR method. Indeed, by computation, we find that the dominant eigenvector of $E(\hat{\mathbf{x}})$ is $\mathbf{y} \approx [-0.0155, 0.1460, -0.9892]^\top$, and moreover, $\phi_w(\mathbf{y}) < \phi_w(\hat{\mathbf{x}})$ and $\phi_d(\mathbf{y}) > \phi_d(\hat{\mathbf{x}})$. This implies that Case II in the second scenario occurs. According to our formulation, we get

$$\tilde{\gamma} \approx 1.9004, \quad \tilde{\beta} \approx 0.7346, \quad \text{and} \quad \tilde{\alpha}(\tilde{\beta}) \approx 0.6785, \quad \text{with} \quad (\tilde{\alpha}(\tilde{\beta}))^2 + \tilde{\beta}^2 - 1 \approx 2.2 \times 10^{-16},$$

and thus a new point $\tilde{\mathbf{z}} \approx [0.6462, -0.0563, -0.7610]^\top$ is constructed. We note that $f(\tilde{\mathbf{z}}) \approx 10.7305 > f(\overline{\mathbf{x}}) \approx 7.9664$, implying $\tilde{\mathbf{z}}$ is an improved point in terms of the value of $f(\mathbf{x})$. Using $\tilde{\mathbf{z}}$ as the new starting point for the RTR method (Algorithm 2-1), we finally get a convergent point $\mathbf{x}^* \approx [0.7199, -0.0200, -0.6938]^\top$ with $f(\mathbf{x}^*) \approx 11.2008$ and $\phi_d(\mathbf{x}^*)\phi_d(\mathbf{x}^*) \approx 1.0585$; moreover, the eigenvalues of $E(\mathbf{x}^*)$ and $H(\mathbf{x}^*)$ are $-24.4690, -2.6019, 1.0585$ and $-25.0447, -6.6080, 0$, respectively. This implies $\mathbf{x}^*$ is a dominant eigenvector of $E(\mathbf{x}^*)$ and is also a global maximizer for (1.1).

4 Perturbation analysis

In this section, we will establish some perturbation properties for the objective function $f(\mathbf{x})$ of (1.1). Whenever a real-world problem of (1.1) is solved numerically, two types of perturbations should be considered. First, since we have known from Section 3 that the global maximizer is a dominant eigenvector of a related matrix, (1.1) can only be solved by means of iteration, in which the roundoff error cannot be avoided. Secondly, in real-world application, the original data matrices in (1.1) are frequently corrupted by noise (this is always the case in the sparse Fisher discriminant analysis for face recognition, where B and W are formed from sample images), and thereby, the formulated problem (1.1) can only be thought as a perturbed model of the original one. For these two kinds of perturbation, we will establish upper perturbation bounds for the objective function $f(\mathbf{x})$.

First, let us assume the matrices in (1.1) are exact but $\mathbf{x} \in \mathscr{M}$ is a point near a solution of (1.1). Taking another look at the cost function $f(\mathbf{x})$, we find that either the standard Rayleigh quotient $\mathbf{x}^\top D\mathbf{x}$ or the generalized Rayleigh quotient $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ possesses a distinguished property: around any critical point, say $\overline{\mathbf{x}}$, of $\mathbf{x}^\top D\mathbf{x}$ or $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ on $\mathscr{M}$, a perturbation $\Delta\mathbf{x}$ on $\overline{\mathbf{x}}$ will yield a quadratic perturbation $\mathscr{O}(\|\Delta\mathbf{x}\|_2^2)$ on the $\mathbf{x}^\top D\mathbf{x}$ or $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$. Such perturbation property has been proved to be important for analyzing the behavior, especially the local quadratic convergence, of some iterations (see e.g., [22, 31, 32, 39]) and designing the stopping criterion as well. As the cost function $f(\mathbf{x})$ is the sum of $\mathbf{x}^\top D\mathbf{x}$ and $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$, we also expect such property could also be inherited at any critical point of $f_{|_{\mathscr{M}}}(\mathbf{x})$. Theorem 4.1 gives the positive answer, which also indicates that a reasonably good feasible approximate to the solution (1.1) is of higher order accuracy in the objective value. The underlying principle for this conclusion is that mappings $f(\mathbf{x})$ and $\mathbf{x} \mapsto \mathbf{x}^\top E(\overline{\mathbf{x}})\mathbf{x}$ are smooth, and $\overline{\mathbf{x}}$ is a critical point for both.

Theorem 4.1 *Let $B, D \in \mathbb{S}_n$ and $W \in \mathbb{S}_n^{++}$. Suppose $\overline{\mathbf{x}}$ is any critical point of $f_{|_{\mathscr{M}}}(\mathbf{x})$, i.e., $\overline{\mathbf{x}} \in \mathscr{S}$ defined by* (3.5)*, then there exist positive constants c_1 and c_2 such that for any $\mathbf{x} \in \mathscr{M}$,*

$$|\mathbf{x}^\top E(\overline{\mathbf{x}})\mathbf{x} - \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})| \le c_1\|\Delta\mathbf{x}\|_2^2, \quad \text{and} \tag{4.1}$$

$$|f(\mathbf{x}) - f(\overline{\mathbf{x}})| \le c_2\|\Delta\mathbf{x}\|_2^2, \tag{4.2}$$

where $\Delta\mathbf{x} = \mathbf{x} - \overline{\mathbf{x}}$.

Proof For (4.1), we first note that $E(\overline{\mathbf{x}})\overline{\mathbf{x}} = \phi_d(\overline{\mathbf{x}})\phi_w(\overline{\mathbf{x}})\overline{\mathbf{x}}$ and

$$1 = (\overline{\mathbf{x}} + \Delta\mathbf{x})^\top(\overline{\mathbf{x}} + \Delta\mathbf{x}) = 1 + 2\overline{\mathbf{x}}^\top\Delta\mathbf{x} + \|\Delta\mathbf{x}\|_2^2, \quad \text{implying} \quad 2\overline{\mathbf{x}}^\top\Delta\mathbf{x} = -\|\Delta\mathbf{x}\|_2^2.$$

Thus

$$\mathbf{x}^\top E(\bar{\mathbf{x}})\mathbf{x} = (\bar{\mathbf{x}}+\Delta\mathbf{x})^\top E(\bar{\mathbf{x}})(\bar{\mathbf{x}}+\Delta\mathbf{x}) = \phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}}) - \phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})\|\Delta\mathbf{x}\|_2^2 + \Delta\mathbf{x}^\top E(\bar{\mathbf{x}})\Delta\mathbf{x},$$

which leads to (4.1). For (4.2), by observing the relation

$$|\delta(\mathbf{x},\bar{\mathbf{x}})| = |(\phi_w(\mathbf{x}) - \phi_w(\bar{\mathbf{x}}))(\phi_d(\mathbf{x}) - \phi_d(\bar{\mathbf{x}}))| \le c_3\|\Delta\mathbf{x}\|_2^2,$$

for some constant c_3, we know that, with the aid of (3.13) and (4.1), the assertion is true.

Now, we will investigate how the global optimal value of (1.1) changes whenever the matrices B, W and D are perturbed slightly. For this purpose, we introduce some new notation. Denoting $Z = [B,W,D] \in \mathbb{S}_n \times \mathbb{S}_n^{++} \times \mathbb{S}_n$, we use $v(Z)$ to represent the global optimal value of (1.1) that is dependent on the variable Z, and use $\mathscr{G}(Z) \subseteq \mathscr{M}$ to denote the set of all global optimal solutions of (1.1) for the given Z. Our next question is then whether $v(Z)$ is continuous or differentiable with respect to Z, and if it is, what is the derivative $\mathbf{D}v(Z)$. In the following theorem, we will first show that $v(Z)$ is Fréchet directionally differentiable at Z, and thereby is continuous with respect to Z. Furthermore, for the special case where $D = \mu W$ for a constant $\mu > 0$, we will show that the global optimal value is differentiable. In the latter case, since (1.1) is only dependent on $Y := [B,W] \in \mathbb{S}_n \times \mathbb{S}_n^{++}$, we will denote the global optimal value by $v(Y)$ and the set of global optimal solutions by $\mathscr{G}(Y)$ instead.

Theorem 4.2 *Let $Z = [B,W,D] \in \mathbb{S}_n \times \mathbb{S}_n^{++} \times \mathbb{S}_n$. Then*

(i) $v(Z)$ is Fréchet directionally differentiable at Z, and the directional derivative at Z in the direction $\Theta = [\Theta_B, \Theta_W, \Theta_D] \in \mathbb{S}_n \times \mathbb{S}_n \times \mathbb{S}_n$ is given by

$$v'(Z,\Theta) = \max_{\mathbf{x}\in\mathscr{G}(Z)} \{\frac{\mathbf{x}^\top\Theta_B\mathbf{x}}{\phi_w(\mathbf{x})} - \frac{\phi_b(\mathbf{x})(\mathbf{x}^\top\Theta_W\mathbf{x})}{\phi_w^2(\mathbf{x})} + \mathbf{x}^\top\Theta_D\mathbf{x}\}; \tag{4.3}$$

(ii) If $D = \mu W \in \mathbb{S}_n^{++}$ for a constant μ and

$$\lambda_1(E(\mathbf{x}^*)) > \lambda_2(E(\mathbf{x}^*)),$$

where $\mathbf{x}^* \in \mathscr{G}(Y)$ is arbitrary, then $v(Y)$ is differentiable at $Y = [B,W] \in \mathbb{S}_n \times \mathbb{S}_n^{++}$ with the derivative given by

$$\mathbf{D}v(Y) = [\frac{\mathbf{x}^*(\mathbf{x}^*)^\top}{\phi_w(\mathbf{x}^*)}, (\mu - \frac{\phi_b(\mathbf{x}^*)}{\phi_w^2(\mathbf{x}^*)})\mathbf{x}^*(\mathbf{x}^*)^\top] \in \mathbb{S}_n \times \mathbb{S}_n. \tag{4.4}$$

Proof In our proof, we use $f(\mathbf{x},Z)$ to denote the objective function of (1.1), which indicates that it is also dependent on Z, and thus $\mathscr{G}(Z) = \arg\max_{\mathbf{x}\in\mathscr{M}} f(\mathbf{x},Z)$. For any sufficiently small $\Theta = [\Theta_B,\Theta_W,\Theta_D] \in \mathbb{S}_n \times \mathbb{S}_n \times \mathbb{S}_n$ so that $W + \Theta_W \in \mathbb{S}_n^{++}$, we have

$$\begin{aligned} f(\mathbf{x},Z+\Theta) &= f(\mathbf{x},Z) + \frac{\mathbf{x}^\top\Theta_B\mathbf{x}}{\phi_w(\mathbf{x})} - \frac{\phi_b(\mathbf{x})(\mathbf{x}^\top\Theta_W\mathbf{x})}{\phi_w^2(\mathbf{x})} + \mathbf{x}^\top\Theta_D\mathbf{x} + o(\|\Theta\|_2), \\ &= f(\mathbf{x},Z) + \frac{\mathrm{tr}(\Theta_B\mathbf{x}\mathbf{x}^\top)}{\phi_w(\mathbf{x})} - \frac{\phi_b(\mathbf{x})\mathrm{tr}(\Theta_W\mathbf{x}\mathbf{x}^\top)}{\phi_w^2(\mathbf{x})} + \mathrm{tr}(\Theta_D\mathbf{x}\mathbf{x}^\top) + o(\|\Theta\|_2), \end{aligned}$$

which implies that the Fréchet derivative of $f(\mathbf{x},Z)$ with respect to Z is given by

$$\mathbf{D}_Z f(\mathbf{x},Z) = [\frac{\mathbf{x}\mathbf{x}^\top}{\phi_w(\mathbf{x})}, -\frac{\phi_b(\mathbf{x})\mathbf{x}\mathbf{x}^\top}{\phi_w^2(\mathbf{x})}, \mathbf{x}\mathbf{x}^\top] \in \mathbb{S}_n \times \mathbb{S}_n \times \mathbb{S}_n.$$

By [1, Theorem 4.13], we then know that $v(Z)$ is Fréchet directionally differentiable at $Z \in \mathbb{S}_n \times \mathbb{S}_n^{++} \times \mathbb{S}_n$ and the Fréchet directional derivative $v'(Z,\Theta)$ in the direction Θ is given by (4.3).

For the second part (ii) where $D = \mu W$, we have known from Theorem 3.6 that $\mathscr{G}(Y) = \mathscr{S}^*$, where $\mathscr{S}^*$ is given by (3.22). Therefore, it is clear that under the condition $\lambda_1(E(\mathbf{x}^*)) > \lambda_2(E(\mathbf{x}^*))$, we have $\mathscr{G}(Y) = \mathscr{S}^* = \{\mathbf{x}^*, -\mathbf{x}^*\}$, and from the first part of this theorem, it follows that for any $\Psi = [\Psi_B, \Psi_W] \in \mathbb{S}_n \times \mathbb{S}_n$,

$$v'(Y,\Psi) = \frac{\mathrm{tr}(\Psi_B \mathbf{x}^*(\mathbf{x}^*)^\top)}{\phi_w(\mathbf{x}^*)} - \frac{\phi_b(\mathbf{x}^*)\mathrm{tr}(\Psi_W \mathbf{x}^*(\mathbf{x}^*)^\top)}{\phi_w^2(\mathbf{x}^*)} + \mu\,\mathrm{tr}(\Psi_W \mathbf{x}^*(\mathbf{x}^*)^\top),$$

which implies the expression (4.4). This completes this proof.

As a final remark of this section, we provide an upper perturbation bound for the global optimal value $v(Z)$. Suppose $Z = [B,W,D] \in \mathbb{S}_n \times \mathbb{S}_n^{++} \times \mathbb{S}_n$ is perturbed by a sufficiently small $\Theta = [\Theta_B, \Theta_W, \Theta_D] \in \mathbb{S}_n \times \mathbb{S}_n \times \mathbb{S}_n$ such that $W + \Theta_W \in \mathbb{S}_n^{++}$, then according to Theorem 4.2, we have

$$v(Z+\Theta) = v(Z) + v'(Z,\Theta) + o(\|\Theta\|_2);$$

on the other hand, according to (4.3), it is true that

$$|v'(Z,\Theta)| \le \frac{\|\Theta_B\|_2}{\|W^{-1}\|_2} + \frac{\|B\|_2}{\|W^{-1}\|_2^2}\|\Theta_W\|_2 + \|\Theta_D\|_2.$$

Consequently, we have the following upper perturbation bound for $v(Z)$

$$|\Delta v| = |v(Z+\Theta) - v(Z)| \le \frac{1}{\|W^{-1}\|_2}\|\Theta_B\|_2 + \frac{\|B\|_2}{\|W^{-1}\|_2^2}\|\Theta_W\|_2 + \|\Theta_D\|_2 + o(\|\Theta\|_2),$$

which explicitly reveals how the matrices B,W,D and their corresponding perturbations $\Theta_B, \Theta_W, \Theta_D$ determine the perturbation on the global optimal value of (1.1).

5 The Riemannian trust-region method for problem (1.1)

The purpose of this section is to realize the generic Riemannian trust-region algorithm (RTR) [3] of [5,4] to solve our discussed problem (1.1). By taking advantages of the smooth function $f_{|\mathscr{M}}(\mathbf{x})$, and the smooth Riemannian manifold $\mathscr{M}$, we find that the RTR method is particularly appropriate for the following reasons: *(i)* RTR utilizes the first and the second order information of $f_{|\mathscr{M}}(\mathbf{x})$, which have been established in section 3, and the local superlinear convergence is achievable, *(ii)* the RTR method converges to a set of critical points of $f_{|\mathscr{M}}(\mathbf{x})$ from all starting points, *(iii)* the sequence $\{f(\mathbf{x}^{(k)})\}$ is nondecreasing for the generated iterates $\{\mathbf{x}^{(k)}\}$, which not only favors convergence to a local maximizer, but makes our starting

[3] In Matlab environment, the generic Riemannian trust-region package for the optimization of functions defined on Riemannian manifolds is available at: http://www.math.fsu.edu/∼cbaker/GenRTR/.

point strategy in section 3.3 applicable, and *(iv)* the method is suitable for large-scale problems because only the matrix-vector product is required. For more detailed discussions on the RTR method as well as its applications in the numerical linear algebra, please refer to [3–6] and many references cited therein.

To state the details of the RTR method in solving (1.1), we note that the RTR method consists of three basic steps: *(i)* building a quadratic model, i.e., a *trust-region subproblem*, of $f_{|\mathscr{M}}(\mathbf{x})$ on the tangent space at each iterate $\mathbf{x}^{(k)} \in \mathscr{M}$, *(ii)* solving the trust-region subproblem by some sophisticated solver, and *(iii)* mapping the solution of the trust-region subproblem onto $\mathscr{M}$ via the so-called *retraction* to complete a single iterate.

For the step *(i)*, suppose $\mathbf{x}^{(k)} \in \mathscr{M}$ is the current iterate and we attempt to build a quadratic model which should be a sufficient good approximation of $f_{|\mathscr{M}}(\mathbf{x})$. This purpose, with the aid of our discussions on the first and the second order information of $f_{|\mathscr{M}}(\mathbf{x})$ in section 3, can be realized by the following quadratic form [5] for $\mathbf{h} \in \mathscr{T}_{\mathbf{x}^{(k)}}\mathscr{M}$:

$$m_{\mathbf{x}^{(k)}}(\mathbf{h}) := f(\mathbf{x}^{(k)}) + \mathbf{h}^\top g(\mathbf{x}^{(k)}) + \frac{\mathbf{h}^\top H(\mathbf{x}^{(k)})\mathbf{h}}{\phi_w(\mathbf{x}^{(k)})}, \tag{5.1}$$

where $g(\mathbf{x}^{(k)})$ and $H(\mathbf{x}^{(k)})$ are defined by (3.4) and (3.8), respectively. Notice that this quadratic form is built on the tangent space $\mathscr{T}_{\mathbf{x}^{(k)}}\mathscr{M}$, and the trust-region subproblem could then be simply expressed as[4]

$$\min_{\mathbf{h}\in\mathscr{T}_{\mathbf{x}^{(k)}}\mathscr{M}, \|\mathbf{h}\|_2 \le \Delta_k} \{-\mathbf{h}^\top g(\mathbf{x}^{(k)}) - \frac{\mathbf{h}^\top H(\mathbf{x}^{(k)})\mathbf{h}}{\phi_w(\mathbf{x}^{(k)})}\}, \tag{5.2}$$

where $\Delta_k > 0$ is the *trust-region radius* which will be updated according to how good is the trust-region subproblem in approximating the problem (1.1).

Now for the step *(ii)*, where the trust-region subproblem (5.2) should be solved, we can employ several classical approaches [10,30]. This forms the inner iteration of the RTR Algorithm 2. Among various methods for (5.2), the *truncated conjugate-gradient* (tCG) method of Steihaug [35] and Toint [36], is particularly efficient and appealing. For completeness, the pesudo-code of tCG is presented in Algorithm 1. It is worth mentioning that as long as the initial guess $\mathbf{h}^{(0)} \in \mathscr{T}_{\mathbf{x}^{(k)}}\mathscr{M}$, the tCG iteration guarantees that the sequence $\{\mathbf{h}^{(j)}\}$, and hence the (approximate) solution $\bar{\mathbf{h}}$ are all in $\mathscr{T}_{\mathbf{x}^{(k)}}\mathscr{M}$, and moreover, preconditioning techniques can be easily incorporated in the tCG iteration. For the stopping criterion in (A), as suggested by Absil *et al.* [4], we can terminate either after a fixed number of iterations (for example, we can truncate if $j > \frac{n(n-1)}{2}$), or by the criterion:

$$\|\mathbf{g}^{(j+1)}\|_2 \le \|\mathbf{g}^{(0)}\|_2 \min\{\|\mathbf{g}^{(0)}\|_2^{\sigma}, \kappa\}, \tag{5.3}$$

where $\kappa, \sigma > 0$ are real parameters. The latter is similar to the stopping criterion used in the inexact Newton iteration (see e.g., [26,30]), and with this stopping criterion, the RTR-tCG algorithm (Algorithm 2-1) is shown to possess local superlinear convergence (see [4], Theorems 4.13 and 4.14). In our numerical testing, we set $\kappa = 0.1$ and $\sigma = 1$.

[4] Because the general RTR method proposed in [4,5] is stated to *minimize* a cost function on a general manifold, we will solve $\min_{\mathbf{x}\in\mathscr{M}} -f(\mathbf{x})$, instead of $\max_{\mathbf{x}\in\mathscr{M}} f(\mathbf{x})$.

Algorithm 1: The truncated CG (tCG) method [35] for the trust-region subproblem.

Initialization: set $\mathbf{h}^{(0)}=\mathbf{0}$, $\mathbf{g}^{(0)}=-g(\mathbf{x}^{(k)})$, $\mathbf{q}^{(0)}=-\mathbf{g}^{(0)}$, where $g(\mathbf{x}^{(k)})$ is given by (3.4);
for $j=0,1,\cdots$, **do**
  if $(\mathbf{q}^{(j)})^\top H(\mathbf{x}^{(k)})\mathbf{q}^{(j)}\geq 0$ **then**
    compute $\tau\in\mathbb{R}$ such that $\mathbf{h}(\tau)=\mathbf{h}^{(j)}+\tau\mathbf{q}^{(j)}$ minimizes $-m_{\mathbf{x}^{(k)}}(\mathbf{h}(\tau))$ of (5.1) with $\|\mathbf{h}(\tau)\|_2=\Delta_k$;
    return $\bar{\mathbf{h}}=\mathbf{h}(\tau)$;
  set $\alpha_j=-\frac{\phi_w(\mathbf{x}^{(k)})(\mathbf{g}^{(j)})^\top\mathbf{g}^{(j)}}{2(\mathbf{q}^{(j)})^\top H(\mathbf{x}^{(k)})\mathbf{q}^{(j)}}$, where $H(\mathbf{x}^{(k)})$ is given by (3.8);
  set $\mathbf{h}^{(j+1)}=\mathbf{h}^{(j)}+\alpha_j\mathbf{q}^{(j)}$;
  if $\|\mathbf{h}^{(j+1)}\|_2\geq\Delta_k$ **then**
    compute $\tau\geq 0$ such that $\mathbf{h}(\tau)=\mathbf{h}^{(j)}+\tau\mathbf{q}^{(j)}$ satisfies $\|\mathbf{h}(\tau)\|_2=\Delta_k$;
    return $\bar{\mathbf{h}}=\mathbf{h}(\tau)$;
  set $\mathbf{g}^{(j+1)}=\mathbf{g}^{(j)}-\frac{2\alpha_j}{\phi_w(\mathbf{x}^{(k)})}H(\mathbf{x}^{(k)})\mathbf{q}^{(j)}$;
(A) **if** *the given stopping criterion is met* **then**
    return $\bar{\mathbf{h}}=\mathbf{h}^{(j+1)}$;
  set $\beta_{j+1}=\frac{(\mathbf{g}^{(j+1)})^\top\mathbf{g}^{(j+1)}}{(\mathbf{g}^{(j)})^\top\mathbf{g}^{(j)}}$;
  set $\mathbf{q}^{(j+1)}=-\mathbf{g}^{(j+1)}+\beta_{j+1}\mathbf{q}^{(j)}$;

If we have a (approximate) solution $\bar{\mathbf{h}}$ to (5.2), then for the last step *(iii)* of RTR, we can use a natural *retraction* [5]

$$\Pi_{\mathbf{x}}(\mathbf{h}):=\frac{\mathbf{x}+\mathbf{h}}{\|\mathbf{x}+\mathbf{h}\|_2}\in\mathscr{M},\quad\forall\mathbf{x}\in\mathscr{M}\quad\text{and}\quad\forall\mathbf{h}\in\mathscr{T}_{\mathbf{x}}\mathscr{M},\tag{5.4}$$

to get a corresponding point $\bar{\mathbf{x}}^{(k+1)}=\Pi_{\mathbf{x}^{(k)}}(\bar{\mathbf{h}})\in\mathscr{M}$. The quality of (5.2) and the (approximate) solution $\bar{\mathbf{h}}$ can be measured by the quotient

$$\rho_k:=\frac{f(\mathbf{x}^{(k)})-f(\bar{\mathbf{x}}^{(k+1)})}{m_{\mathbf{x}^{(k)}}(\mathbf{0})-m_{\mathbf{x}^{(k)}}(\bar{\mathbf{h}})}.\tag{5.5}$$

The rule of accepting or rejecting the candidate $\bar{\mathbf{x}}^{(k+1)}$ as the next iterate, as well as updating the radius Δ_k, basically follows the classical trust-region method (see [10,30]), and details of the rule are formalized in outer loop iteration of Algorithm 2. In our numerical testing presented in this paper, we choose $\bar{\Delta}=0.1n$, $\Delta_0=0.2\bar{\Delta}$, $\varepsilon=10^{-6}$ and $\rho'=0.1$.

The convergence of the general RTR method has been well established in [4,5] under reasonable assumptions. For our problem (1.1) in particular, since the cost function $f_{|\mathscr{M}}(\mathbf{x})$ and the retraction (5.4) are smooth, and $\mathscr{M}$ is a smooth and compact Riemannian manifold, the nice convergence properties are preserved. Precisely, from Theorem 4.4 and Corollary 4.6 of [4], for any starting point $\mathbf{x}^{(0)}\in\mathscr{M}$, the sequence $\{\mathbf{x}^{(k)}\}$ generated by the RTR-tCG method satisfies

$$\lim_{k\to+\infty}[E(\mathbf{x}^{(k)})\mathbf{x}^{(k)}-\phi_d(\mathbf{x}^{(k)})\phi_w(\mathbf{x}^{(k)})\mathbf{x}^{(k)}]=\mathbf{0},$$

implying any accumulation of $\{\mathbf{x}^{(k)}\}$ is a critical point of $f_{|\mathscr{M}}(\mathbf{x})$; moreover, based on Theorems 4.12 and 4.13 of [4], for any point $\bar{\mathbf{x}}\in\mathscr{M}$ satisfying the condition *(iii)* of Theorem 3.2, there is a neighborhood $\mathscr{U}_{\bar{\mathbf{x}}}$

Algorithm 2: The Riemannian trust-region algorithm for (1.1).

Initialization: choose parameters $\bar{\Delta} > 0, \Delta_0 \in (0, \bar{\Delta}), \rho' \in (0, \frac{1}{4})$, the tolerance $\varepsilon > 0$ and $\mathbf{x}^{(0)} \in \mathscr{M}$; set $k := 0$;
while $\|g(\mathbf{x}^{(k)})\|_2 > \varepsilon$, **do**
  solving (approximately by Algorithm 1) (5.2) to obtain $\bar{\mathbf{h}}$;
  compute the quotient ρ_k by (5.5);
  if $\rho_k < \frac{1}{4}$ **then**
    $\Delta_{k+1} = \frac{\Delta_k}{4}$
  else
    if $\rho_k > \frac{3}{4}$ *and* $\|\bar{\mathbf{h}}\|_2 = \Delta_k$ **then**
      $\Delta_{k+1} = \min\{2\Delta_k, \bar{\Delta}\}$
    else
      $\Delta_{k+1} = \Delta_k$
  if $\rho_k > \rho'$ **then**
    $\mathbf{x}^{(k+1)} = \frac{\mathbf{x}^{(k)} + \bar{\mathbf{h}}}{\|\mathbf{x}^{(k)} + \bar{\mathbf{h}}\|_2}$
  else
    $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$
  set $k := k+1$;

of $\bar{\mathbf{x}}$ in $\mathscr{M}$ such that for all $\mathbf{x}^{(0)} \in \mathscr{U}_{\bar{\mathbf{x}}}$, the RTR-tCG algorithm with the stopping criterion (5.3) converges to $\bar{\mathbf{x}}$ superlinearly.

To conclude this section, we point out that even though the RTR-tCG method favors the local maximizer due to the nondecreasing property of $\{f(\mathbf{x}^{(k)})\}$, finding a global solution to (1.1) is not guaranteed. For instance, for Example 1.1 with $D = D_1 = \mathrm{diag}\{5,1\}$, the left sub-figure in Figure 1 demonstrates starting points $\mathbf{x}^{(0)} \in \mathbb{R}^2$ converging to a global maximizer (such $\mathbf{x}^{(0)}$ is marked by '$\cdot$') and to a local maximizer (such $\mathbf{x}^{(0)}$ is marked by '$*$'), respectively; in the right sub-figure, demonstrated are the parameters (θ, η) in the 2π-period $\theta - \eta$ plane of the starting point $\mathbf{x}^{(0)} = [\mathbf{x}_1(\theta,\eta), \mathbf{x}_2(\theta,\eta), \mathbf{x}_3(\theta)]^\top$ for Example 3.1, from which the RTR-tCG algorithm converges to a global maximizer (corresponding to '$\cdot$') and to a local maximizer (corresponding to '$*$'), respectively. Indeed, as long as the local non-global solution exists, due to the monotonic property of $\{f(\mathbf{x}^{(k)})\}$, the sequence $\{\mathbf{x}^{(k)}\}$ would be trapped by a local solution near $\mathbf{x}^{(0)}$. From this point of view, some improvement on the starting point, such as that in section 3.3, can be helpful for the RTR algorithm to escape from a local solution. This is preliminarily demonstrated by our numerical experiments in next section.

6 Numerical experiments

In this section, we report on some preliminary numerical experiments on the problem (1.1). All our numerical tests in this paper are carried out in MATLAB platform on a PC with Intel(R) Core(R)i3 CPU 550@3.20GHz, 3.20GHz. In solving our tested problems, we also employed the MATLAB function fmincon, which uses the sequential quadratic programming (SQP) method [30] with quasi-Newton and line search techniques (for medium-scale optimization). Unfortunately, we observed that fmincon fails for most of our tested cases, mainly due to the violation of the nonlinear constraint, and therefore, these numerical results will not be presented.

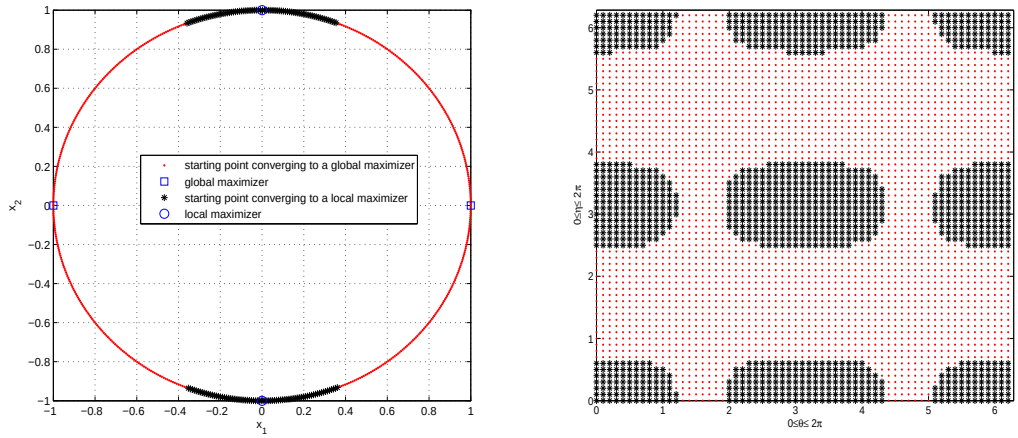

Fig. 1 For Example 1.1 with $D = D_1 = \mathrm{diag}\{5,1\}$, the left sub-figure demonstrates starting points $\mathbf{x}^{(0)} \in \mathbb{R}^2$ converging to a global maximizer (such $\mathbf{x}^{(0)}$ is marked by '$\cdot$') and to a local maximizer (such $\mathbf{x}^{(0)}$ is marked by '$*$'), respectively; the right sub-figure demonstrates the parameters (θ,η) of the starting point $\mathbf{x}^{(0)} = [\mathbf{x}_1(\theta,\eta),\mathbf{x}_2(\theta,\eta),\mathbf{x}_3(\theta)]^\top$ for Example 3.1, from which the RTR-tCG algorithm converges to a global maximizer (corresponding to '$\cdot$') and to a local maximizer (corresponding to '$*$'), respectively.

As our starting point strategy is applicable for the RTR-tCG algorithm, we first test its capability in finding a global solution. For this purpose, we generated 10^4 starting points with elements normally distributed for Example 3.2, and counted the cases for which the RTR-tCG algorithm converges to a global maximizer. It is observed that there are almost 41.88% cases fail to converge to a global solution. By contrast, using the same starting points but activating our starting point strategy, we find that for *all* the cases, it converges to the global maximizer. This implies that our starting point strategy is effective in finding a global solution on this example. On the other hand, however, we should keep in mind that additional computational costs are required to implement our starting point strategy (mainly on computation of the dominant eigenvector), and furthermore, we observed that for randomly generated problem (1.1), it has high probability that the convergent point $\bar{\mathbf{x}}$ of the RTR-tCG method is a dominant eigenvector of $E(\bar{\mathbf{x}})$. Therefore, we claim that the RTR-tCG algorithm, without the starting point strategy, is still effective. In our following numerical reports, we inactivate the starting point strategy and focus on the global and local convergence of the RTR-tCG algorithm.

To investigate the global convergence of the RTR-tCG algorithm, we tested the convergence behavior on randomly generated problems. For a fixed dimension n, we produced $B,W,D \in \mathbb{S}_n^{++}$ and the starting point $\mathbf{x}^{(0)}$ with elements chosen from normal distribution. We set $n = 5,10,50,100,200,500$ and 1000. For each given n, we generated 10^4 testing problems and recorded the convergence information. The global convergence is observed because for all these cases, the RTR-tCG algorithm converges to a solution $\bar{\mathbf{x}}$ with the residual satisfying

$$\|\delta\bar{\mathbf{x}}\|_2 = \|E(\bar{\mathbf{x}})\bar{\mathbf{x}} - \phi_d(\bar{\mathbf{x}})\phi_w(\bar{\mathbf{x}})\bar{\mathbf{x}}\|_2 \leq 10^{-5}.$$

Our last numerical testing was conducted on some ill-conditioned matrices B, W and D. As maximizing the term $\frac{\mathbf{x}^\top B\mathbf{x}}{\mathbf{x}^\top W\mathbf{x}}$ is equivalent to computing an extreme eigenpair of a generalized eigenvalue problem: $W\mathbf{x} = \lambda B\mathbf{x}$, we choose the pair (B,W) from the set BCSSTRUC1 (from the Harwell-Boeing collection[5]). In this set, there are 13 pairs named: (BCSSTM01, BCSSTK01),..., (BCSSTM13, BCSSTK13), where BCSSTM01, ..., BCSSTM13 are symmetric positive semidefinite, and BCSSTK01, ..., BCSSTK13 are symmetric positive definite. Most of these matrices are sparse and ill-conditioned. In our testing, the first 5 pairs are selected as (B,W) and the statistics of these matrices are summarized in Table 1. For the choice of D, on the other hand, as it does not lose the generality to assume D to be diagonal, we therefore set

$$D = \mathrm{diag}\{10^{-4}, 10^{-4}, 1, \ldots, n-4, 10^2, 10^2\},$$

which is ill-conditioned and contains repeated extreme eigenvalues.

Table 1 Summary of the pair (B,W).

B	n	Condition number	W	Condition number
BCSSTM01	48	Inf	BCSSTK01	$1.6E+06$
BCSSTM02	66	8.8	BCSSTK02	$1.3E+04$
BCSSTM03	112	Inf	BCSSTK03	$9.5E+06$
BCSSTM04	132	Inf	BCSSTK04	$5.6E+06$
BCSSTM05	153	Inf	BCSSTK05	$3.5E+04$

For this set of testing problems, we summarized the convergence results of the RTR-tCG algorithm starting from a randomly generated $\mathbf{x}^{(0)}$. In Table 2, reported are the number of outer iterations, the number of overall inner iterations in tCG (Algorithm 1), and the CPU time. These results demonstrate the efficiency of the RTR-tCG algorithm. For all these cases furthermore, we observed that each convergent point $\overline{\mathbf{x}}$ not only satisfies the second-order sufficient optimality condition (i.e., *(iii)* of Theorem 3.2), but is a dominant eigenvector of the corresponding matrix $E(\overline{\mathbf{x}})$. To illustrate the superlinear convergence, we plot the sequences $\{f(\mathbf{x}^{(k)})\}$ and $\{log_{10}\|\delta\mathbf{x}^{(k)}\|_2\}$ for the last tested case (BCSSTM05, BCSSTK05) in the left and the right sub-figures of Figure 1, respectively, where the superlinear convergence is clearly observed.

7 Concluding remarks and future work

In this paper, we have made some efforts in optimizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient on the unit sphere. A specific application arising from the sparse Fisher discriminant analysis is introduced. As the related optimization problem is not convex and consists of local non-global maxima, it becomes much harder to find the global solution. We therefore first investigated some characterizations for the local and the global maxima, which are described as a special extreme

[5] http://math.nist.gov/MatrixMarket/

Table 2 Summary of the convergence results.

(B,W)	outer iter. #.	inner iter. #.	CPU(s)
(BCSSTM01, BCSSTK01)	9	20	0.0156
(BCSSTM02, BCSSTK02)	15	85	0.0156
(BCSSTM03, BCSSTK03)	16	191	0.0624
(BCSSTM04, BCSSTK04)	17	192	0.0468
(BCSSTM05, BCSSTK05)	14	196	0.0624

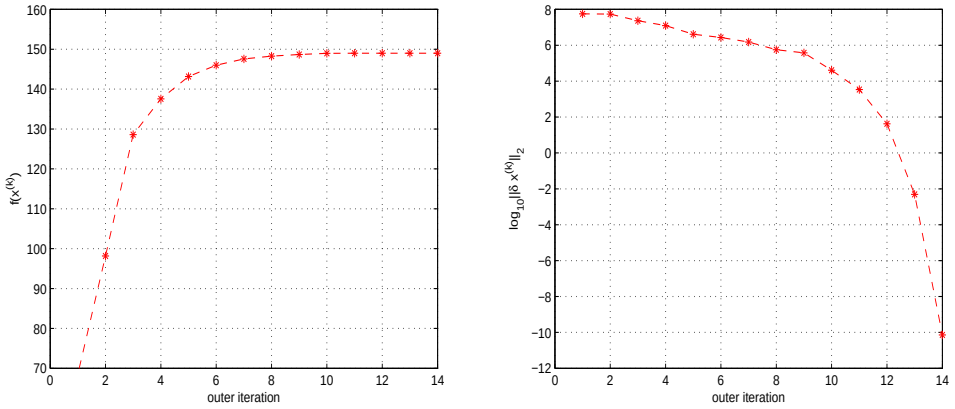

Fig. 1 For the case where (B,W)=(BCSSTM05, BCSSTK05), the left and the right sub-figures plot the sequences $\{f(\mathbf{x}^{(k)})\}$ and $\{log_{10}\|\delta\mathbf{x}^{(k)}\|_2\}$, respectively.

eigenvalue problem. A necessary optimality for the global maximizer sheds some lights on the problem and also leads us to a starting point strategy for any monotonically convergent algorithm. An analogous perturbation property for the Rayleigh quotient and the generalized Rayleigh quotient is also proved to be true for our discussed problem. Taking advantages of the cost function and the constraint, the general Riemannian trust-region algorithm in [4] is applied and empirical evaluation of its performance is reported. By activating our starting point strategy, the RTR-tCG algorithm is effective in solving the problem (1.1), and our numerical experiments demonstrate the global convergence and local superlinear convergence.

Acknowledgement

The author would like to thank the Editor and two anonymous referees for careful reading, helpful comments and suggestions that have improved the presentation of the paper.

References

1. Bonnans, J. F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York (2000)

2. Abraham, R., Marsden, J. E., Ratiu, T.: *Manifolds, tensor analysis, and applications*, vol. 75 of Applied Mathematical Sciences, Springer-Verlag, New York, second edition (1988)
3. Absil P.-A., Baker, C. G., Gallivan, K. A.: *A truncated-CG style method for symmetric generalized eigenvalue problems*, J. Comput. Appl. Math., 189, 274-285 (2006)
4. Absil P.-A., Baker, C. G., Gallivan, K. A.: *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7, 303-330 (2007)
5. Absil P.-A., Mahony, R., Sepulchre, R.: *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ (2008)
6. Absil P.-A., Gallivan, K. A.: *Accelerated line-search and trust-region methods*, SIAM Journal on Numerical Analysis, 47, 997-1018 (2009)
7. Adler, R. L., Dedieu, J.-P., Margulies, J. Y., Martens M., Shub, M.: *Newton's method on Riemannian maniflods and a geometric model for the human spine*, IMA J. Numer. Anal., 22, 359-390 (2002)
8. Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer (2006)
9. Chu M. T., Driessel, K. R.: *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27, 1050-1060 (1990)
10. Conn, A. R., Gould, N. I. M., Toint, P. L.: *Trust-region methods*, SIAM, Philadelphia (2000)
11. Duchene L., Leclerq, S.: *An optimal transformation for discriminant and principal component analysis*, IEEE Trans. Pattern Analysis and Machine Intelligence, 10, 978-983 (1988)
12. Dundar, M. M., Fung, G., Bi J., Sandilya S., Rao, B.: *Sparse Fisher discriminant analisys for computer aided detection*, Proceedings of SIAM International Conference on Data Mining (2005)
13. Edelman, A., Arias T. A., Smith, S. T.: *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20, 303-353 (1998)
14. Fan J., Li, R.: *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Statistical Assoc., 96, 1348-1360 (2001)
15. Fisher R. A.: *The use of multiple measurements in taxonomic problems*, Annual of Eugenics, 7, 179-188 (1936)
16. Foley D., Sammon J.: *An optimal set of discriminant vectors*, IEEE Trans Computers, 24, 281-289 (1975)
17. Friedman J.: *Regularized discriminant analysis*, J. Am. Statistical Assoc., 84, 165-175 (1989)
18. Fukunaga K.: *Introduction to statistical pattern classification*, Academic Press (1990)
19. Fung E., Ng, M.: *On sparse Fisher discriminant method for microarray data analysis*, Bioinformation, 2, 230-234 (2007)
20. Gao X. B., Golub G. H., Liao, L.-Z.: *Continuous methods for symmetric generalized eigenvalue problems*, Linear Alg. Appl., 428, 676-696 (2008)
21. Golub G. H., Liao, L.-Z.: *Continuous methods for extreme and interior eigenvalue problems*, Linear Alg. Appl., 415, 31-51 (2006)
22. Golub G. H., Van Loan, C. F.: *Matrix computations, 3rd ed.*, Johns Hopkins University Press, Baltimore, MD (1996)
23. Helmke U., Moore, J. B.: *Optimization and dynamical systems*, Springer-Verlag, London, UK (1994)
24. Howland, P., Jeon, M., Park, H.: *Structure Preserving Dimension Reduction for Clustered Text Data Based on the Generalized Singular Value Decomposition*, SIAM J. Matrix Analysis and Applications, 25, 165-179 (2003)
25. Hunter D. R., Li, R.: *Variable selection using MM algorithms*, The Annals of Statistics, 33, 1617-1642 (2005)
26. Kelley, C. T.: *Iterative methods for linear and nonlinear equations*, SIAM, Philadelphia, PA (1995)
27. Lehoucq R. B., Sorensen, D. C.: *Deflation techniques for an implicitly re-started Arnoldi iteration*, SIAM J. Matrix Anal. Appl., 17, 789-821 (1996)
28. Lehoucq, R. B., Sorensen, D. C., Yang, C.: *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, SIAM Publications, Philadelphia (1998)
29. Ng, M. K., Liao, L.-Z., Zhang, L.-H.: *On sparse linear discriminant analysis for high-dimensional data*, Numerical Linear Algebra with Applications, 18, 223-235 (2011)
30. Nocedal J., Wright, S. J.: *Numerical optimization, second edition*, Springer Verlag, New York (2006)
31. Parlett, B. N.: *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28, 679-693 (1974)
32. Parlett, B. N.: *The symmetric eigenvalue problem*, Classics Appl. Math. 20, SIAM, Philadelphia (1998)
33. Primolevo, G., Simeone O., Spagnolini U.: *Towards a joint optimization of scheduling and beamforming for MIMO downlink*, IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications, 493-497 (2006)

34. Saad, Y.: *Numerical methods for large eigenvalue problems*, Algorithms and Architectures for advanced scientific computing, Manchester University Press (1992)
35. Steihaug, T.: *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal. 20, 626-637 (1983)
36. Toint, P. L.: *Towards an efficient sparsity exploiting Newton method for minimization*, Sparse Matrices and Their Uses (I. S. Duff, ed.), Academic Press, London, 57-88 (1981)
37. Wu, M. C., Zhang, L. S., Wang, Z. X., Christiani, D. C., Lin, X. H.: *Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection*, Bioinformatics, 25, 1145-1151 (2009)
38. Ye, J.-P., Janardan, R., Park, C., Park, H.: *An Optimization Criterion for Generalized Discriminant Analysis on Undersampled Problems*, IEEE Trans. Pattern Analysis and Machine Intelligence, 26, 982-994 (2004)
39. Zhang, L.-H., Liao, L.-Z., Ng, M. K.: *Fast Algorithms for the generalized Foley-Sammon discriminant analysis*, SIAM J. Matrix Anal. Appl., 31, 1584-1605 (2010)
40. Zhang, L.-H.: *Uncorrected trace ratio LDA for undersampled problems*, Pattern Recognition Letters, 32, 476-484 (2011)