Operations Research Letters 50 (2022) 1–7

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

New results on multi-dimensional linear discriminant analysis

Amir Beck *, Raz Sharon

School of Mathematical Sciences, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Received 1 July 2021
Received in revised form 21 October 2021
Accepted 1 November 2021
Available online 15 November 2021

Keywords:
Linear discriminant analysis
Spectral isotonic functions
Generalized eigenvectors
Fixed point methods
Superlinear convergence

ABSTRACT

Fisher linear discriminant analysis is a well-known technique for dimensionality reduction and classification. The method was first formulated in 1936 by Fisher. In this paper we concentrate on three different formulations of the multi-dimensional problem. We provide a mathematical explanation why two of the formulations are equivalent and prove that this equivalency can be extended to a broader class of objective functions. The second contribution is a rate of convergence of a fixed point method for solving the third model.

1. Introduction

1.1. Fisher's paper

The beginning of linear discriminant analysis (LDA) can be traced back to Ronald A. Fisher's 1936 seminal paper [4]. Fisher analyzed data from the flowers of plants, each belonging to one of the two species Iris setosa and Iris versicolor, found growing together in the same colony. Four measurements were given for each of the flowers (sepal length, sepal width, petal length, petal width), and fifty samples were available from each of the two species. Fisher then raises his main research question

> [4, Section II] We shall first consider the question: what linear function of the four measurements
>
> $X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$
>
> will maximize the ratio of the difference between the specific means to the standard deviations within species?

To describe Fisher's criterion in precise mathematical terms, we will first denote the fifty samples of the Iris setosa species by $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{50} \in \mathbb{R}^4$, and the fifty samples of the Iris versicolor species by $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{50} \in \mathbb{R}^4$. Essentially, Fisher was interested in finding a direction $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^4$ for which the fifty scalars $\{\mathbf{y}_1^T\mathbf{v}, \dots, \mathbf{y}_{50}^T\mathbf{v}\}$ are "well separated" from the fifty scalars $\{\mathbf{z}_1^T\mathbf{v}, \dots, \mathbf{z}_{50}^T\mathbf{v}\}$. The measure Fisher suggested (up to a constant) for quantifying the level of separation is the following ratio:

$$R(\mathbf{v}) = \frac{(\bar{\mathbf{y}}^T\mathbf{v} - \bar{\mathbf{z}}^T\mathbf{v})^2}{\sum_{i=1}^{50}(\mathbf{y}_i^T\mathbf{v} - \bar{\mathbf{y}}^T\mathbf{v})^2 + \sum_{i=1}^{50}(\mathbf{z}_i^T\mathbf{v} - \bar{\mathbf{z}}^T\mathbf{v})^2}, \tag{1.1}$$

where here $\bar{\mathbf{y}} = \frac{1}{50}\sum_{i=1}^{50}\mathbf{y}_i, \bar{\mathbf{z}} = \frac{1}{50}\sum_{i=1}^{50}\mathbf{z}_i$ are the means of each of the classes. In this notation, $\bar{\mathbf{y}}^T\mathbf{v}$ and $\bar{\mathbf{z}}^T\mathbf{v}$ are the means of the two projected classes. Therefore, the nominator in (1.1) is the square of the difference between the projected class means, and the denominator is the sum of variances within each of the projected classes, namely the pooled variance. The optimization problem consists of maximizing $R(\mathbf{v})$ over all the possible nonzero vectors $\mathbf{v}$.

Fisher showed that the optimal solution of the above problem can be expressed as a solution to a certain linear system. The problem that Fisher solved was actually the two-classes one-dimensional LDA problem where "one-dimensional" here means that the data is projected onto a one-dimensional space.

Despite being a classical dimensional reduction and classification technique, the literature is surprisingly inconsistent regarding the identity of the "correct" multi-dimensional variant of the LDA problem. The next section describes three of the (probably) most popular formulations of the multi-dimensional case.

1.2. Three formulations of the LDA problem

We now consider the natural generalization of Fisher's criterion to the multi-class and multi-dimensional case. Assume we have a

* Corresponding author.
E-mail address: becka@tauex.tau.ac.il (A. Beck).

https://doi.org/10.1016/j.orl.2021.11.003
0167-6377/© 2021 Elsevier B.V. All rights reserved.

dataset in $\mathbb{R}^d$ which contains n samples from c classes labeled as $1, 2, \dots, c$. Each sample $\mathbf{x}_i \in \mathbb{R}^d$, $i \in \{1 \dots n\}$ is associated with one class. We denote by $C^{(k)}$ the set containing all the indices of samples associated with class k. The number of samples in each class is $n_k \equiv \left|C^{(k)}\right|$, so that in particular, $n = \sum_{k=1}^{c} n_k$. In general we wish to project the dataset into a lower dimensional space $\mathbb{R}^p$, $p \le d$, where the different classes can be "easily" separated. We are looking for a matrix, $\mathbf{V} \in \mathbb{R}^{d\times p}$ satisfying $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, and we assume that the new representation of the data set in $\mathbb{R}^p$ is $\mathbf{y}_i \equiv \mathbf{V}^T\mathbf{x}_i$. Note that $\mathbf{V}$ contains in its columns an orthonormal basis of the p-dimensional subspace of $\mathbb{R}^d$ on which the data is projected. We denote by $\boldsymbol{\mu}^{(k)}$ the center of each class, and by $\boldsymbol{\mu}$ the center of the entire dataset:

$$\boldsymbol{\mu}^{(k)} = \tfrac{1}{n_k}\textstyle\sum_{i\in C^{(k)}} \mathbf{x}_i, \quad \boldsymbol{\mu} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{x}_i.$$

Similarly, the class centers, and the entire sample center in the projected space $\mathbb{R}^p$ are denoted by

$$\tilde{\boldsymbol{\mu}}^{(k)} = \tfrac{1}{n_k}\textstyle\sum_{i\in C^{(k)}} \mathbf{y}_i = \mathbf{V}^T\boldsymbol{\mu}^{(k)},$$
$$\tilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = \mathbf{V}^T\boldsymbol{\mu} = \sum_{i=1}^{c}\frac{n_i}{n}\tilde{\boldsymbol{\mu}}^{(k)}.$$

We wish to find a projection of the d-dimensional data to $\mathbb{R}^p$ which in some sense maximizes the separation between classes. To do so, we follow Fisher's idea [4] and define a separation measure which is the ratio of the variance between the classes and the variance within the classes

$$R(\mathbf{V}) = \frac{\phi_B(\mathbf{V})}{\phi_W(\mathbf{V})}. \tag{1.2}$$

Here ϕ_B is the variance of the projected class means and ϕ_W is the sum of variances within the projected classes (multiplicative constants are ignored as they do not change the optimal set of the maximization problem):

$$\phi_B(\mathbf{V}) \equiv \sum_{k=1}^{c} n_k\|\tilde{\boldsymbol{\mu}}^{(k)} - \tilde{\boldsymbol{\mu}}\|_2^2, \quad \phi_W(\mathbf{V}) \equiv \sum_{k=1}^{c}\sum_{i\in C^{(k)}}\left\|\mathbf{y}_i - \tilde{\boldsymbol{\mu}}^{(k)}\right\|_2^2.$$

We can rewrite $\phi_B(\mathbf{V})$ and $\phi_W(\mathbf{V})$ more explicitly as

$$\begin{aligned}
\phi_W(\mathbf{V}) &= \sum_{k=1}^{c}\sum_{i\in C^{(k)}}\left\|\mathbf{y}_i - \tilde{\boldsymbol{\mu}}^{(k)}\right\|_2^2 = \sum_{k=1}^{c}\sum_{i\in C^{(k)}}\left\|\mathbf{V}^T\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)}\right)\right\|_2^2 \\
&= \sum_{k=1}^{c}\sum_{i\in C^{(k)}} \mathrm{Tr}\left(\mathbf{V}^T\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)}\right)^T\mathbf{V}\right) \\
&= \mathrm{Tr}(\mathbf{V}^T\mathbf{W}\mathbf{V}),
\end{aligned}$$
$$\begin{aligned}
\phi_B(\mathbf{V}) &= \sum_{k=1}^{c} n_k\|\mathbf{V}^T\boldsymbol{\mu}^{(k)} - \mathbf{V}^T\boldsymbol{\mu}\|_2^2 \\
&= \mathrm{Tr}\left(\mathbf{V}^T\left(\sum_{k=1}^{c} n_k\left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}\right)^T\right)\mathbf{V}\right) \\
&= \mathrm{Tr}\left(\mathbf{V}^T\mathbf{B}\mathbf{V}\right),
\end{aligned}$$

where $\mathbf{B}$ and $\mathbf{W}$ are the so-called "between-class" "within-class" scatter matrices respectively given by

$$\mathbf{W} \equiv \sum_{k=1}^{c}\sum_{i\in C_k}\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}^{(k)}\right)^T,$$
$$\mathbf{B} \equiv \sum_{k=1}^{c} n_k\left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(k)} - \boldsymbol{\mu}\right)^T.$$

To summarize, Fisher's LDA problem is naturally generalized to the maximization problem

$$\textbf{(FLDA)} \quad \begin{array}{ll} \max & \left\{\frac{\phi_B(\mathbf{V})}{\phi_W(\mathbf{V})} = \dfrac{\mathrm{Tr}(\mathbf{V}^T\mathbf{B}\mathbf{V})}{\mathrm{Tr}(\mathbf{V}^T\mathbf{W}\mathbf{V})}\right\} \\ \text{s.t.} & \mathbf{V}^T\mathbf{V} = \mathbf{I}_p, \mathbf{V} \in \mathbb{R}^{d\times p}. \end{array} \tag{1.3}$$

The specific formulation above is also called *the trace ratio problem* and its solution will be discussed later on in Section 3. Note that by its definition, $\mathbf{W}$ is positive semidefinite, and we will assume throughout the paper that $\mathbf{W} \succ \mathbf{0}$.

Perhaps surprisingly, while in the one-dimensional case ($p = 1$), the formulation (FLDA) is well agreed, different formulations for the multidimensional case ($p > 1$) exist in the literature [2,3,5]. In fact, the trace ratio formulation is not considered to be the "standard" formulation since it was often thought-of as too hard to handle, see for example the discussion in [8]. However, it is now well known that the problem can be efficiently solved [8]. One of the popular formulations for the LDA problem (see e.g., [2,5]) is given by

$$\textbf{(LDA-T)} \quad \begin{array}{ll} \max_{\mathbf{V}} & \mathrm{Tr}((\mathbf{V}^T\mathbf{W}\mathbf{V})^{-1}(\mathbf{V}^T\mathbf{B}\mathbf{V})) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}, \mathbf{V} \in \mathbb{R}^{d\times p}. \end{array} \tag{1.4}$$

Another possibility is to exchange the trace by a determinant in the above problem, leading to the formulation

$$\textbf{(LDA-D)} \quad \begin{array}{ll} \max_{\mathbf{V}} & \det((\mathbf{V}^T\mathbf{W}\mathbf{V})^{-1}(\mathbf{V}^T\mathbf{B}\mathbf{V})) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}, \mathbf{V} \in \mathbb{R}^{d\times p}. \end{array} \tag{1.5}$$

1.3. Contributions

The natural question that arises is what are the connections between the optimal solutions of the three models defined above. It is known [5] that the two most common formulations (LDA-T) and (LDA-D) can be solved via a matrix whose columns consist of p leading $\mathbf{W}$-orthogonal generalized eigenvalues of the matrix pair $(\mathbf{B}, \mathbf{W})$. This solution will be referred to from now on as "the generalized eigenvalue solution". The fact that the trace and determinant-based problems share common optimal solutions raises the natural question whether this phenomenon holds for a larger class of functions, and not just the trace and the determinant. Our first contribution will be to show that this is indeed the case.

- **Contribution 1.** We will show in Section 2 that the generalized eigenvalue solution is an optimal solution of the problem

  $$\textbf{(LDA-G)} \quad \begin{array}{ll} \max_{\mathbf{V}} & \mathcal{F}((\mathbf{V}^T\mathbf{W}\mathbf{V})^{-1}(\mathbf{V}^T\mathbf{B}\mathbf{V})) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}, \mathbf{V} \in \mathbb{R}^{d\times p}, \end{array}$$

  whenever $\mathcal{F}$ belongs to the class of *symmetric spectral functions*, a notion that will be defined later in Section 2.

The second contribution of the paper relates to the solution of the trace ratio problem (FLDA). In Section 3 we will study the rate of convergence of the fixed point scheme

$$\mathbf{V}_{k+1} \in \underset{\mathbf{V}\in\mathbb{R}^{d\times p}}{\mathrm{argmax}}\left\{\mathrm{Tr}\left(\mathbf{V}^T\left[\mathbf{B} - \frac{\phi_B(\mathbf{V}_k)}{\phi_W(\mathbf{V}_k)}\mathbf{W}\right]\mathbf{V}\right) : \mathbf{V}^T\mathbf{V} = \mathbf{I}_p\right\}$$

The method was first introduced in [10], where convergence to the global optimal solution was established. An interpretation of the scheme as a Newton method was shown in [8]. Local quadratic and superlinear convergence rates of the method were established in the works [12] and [13] respectively.

- **Contribution 2.** In this paper we quantify the superlinear rate of convergence of the fixed point scheme. Specifically, we prove the following upper bound on the rate of convergence (f denotes the objective function and f_{opt} is the optimal value):

  $$f_{\text{opt}} - f(\mathbf{V}_n) \le (f_{\text{opt}} - f(\mathbf{V}_0))\alpha^n \beta^{n^2},$$

  where $\alpha > 0$ and $\beta \in (0,1)$ are constants that will be explicitly given.

Notice that since there exists proof of a local quadratic convergence under the same assumption we use in this paper, and hence the upper bound above is not tight.

1.4. Preliminaries on eigenvalues

The set of $n \times n$ symmetric matrices is denoted by $\mathbb{S}^n$, the set of all $n \times n$ positive semidefinite (definite) matrices by $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$), and the set of $n \times n$ orthogonal matrices by $\mathbb{O}^n$. For two integers $d \ge p$, the *Stiefel manifold* is the set $\mathbb{S}_{d,p} = \{\mathbf{X} \in \mathbb{R}^{d\times p} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_p\}$. The eigenvalues of a symmetric matrix $\mathbf{A} \in \mathbb{S}^n$ are denoted by $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots \ge \lambda_n(\mathbf{A})$. Given an integer $k \in \{1, 2, ..., n\}$, a set of "k leading eigenvectors" of $\mathbf{A}$ is a set of eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ that correspond to the eigenvalues $\lambda_1(\mathbf{A}), \lambda_2(\mathbf{A}), \ldots, \lambda_k(\mathbf{A})$ respectively. Suppose that we are given two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{S}^n$. Then a generalized eigenvalue of the matrix pair $(\mathbf{A}, \mathbf{B})$ is a nonzero vector $\mathbf{v}$ satisfying $\mathbf{Av} = \lambda\mathbf{Bv}$ for some $\lambda \in \mathbb{R}$. The scalar λ is referred to as a "generalized eigenvalue of the matrix pair $(\mathbf{A}, \mathbf{B})$". The concept of "leading generalized eigenvectors" is defined similarly. Later on, we will use the fact described in Remark 1.1 below that connects generalized eigenvectors with eigenvectors.

Remark 1.1. Let $\mathbf{A} \in \mathbb{S}^n$ and $\mathbf{B} \in \mathbb{S}^n_{++}$. If $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p\}$ is an orthonormal set of p leading eigenvectors of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, then $\{\mathbf{B}^{-1/2}\mathbf{v}_1, \mathbf{B}^{-1/2}\mathbf{v}_2, \ldots, \mathbf{B}^{-1/2}\mathbf{v}_p\}$ is a $\mathbf{B}$-orthonormal set of p leading generalized eigenvalues of the matrix pair $(\mathbf{A}, \mathbf{B})$. We recall that a set $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$ is called $\mathbf{B}$-orthonormal if $\mathbf{w}_i^T\mathbf{B}\mathbf{w}_i = 1$ for all i and $\mathbf{w}_i^T\mathbf{B}\mathbf{w}_j = 0$ for any $i \ne j$.

2. Solution of (LDA-T) and (LDA-D)

We now consider the formulations (LDA-T) and (LDA-D) ((1.4) and (1.5) respectively) in the multidimensional case.

We first note that both (LDA-T) and (LDA-D) can be written as

$$\begin{array}{ll} \max_{\mathbf{V}} & \mathcal{F}((\mathbf{V}^T\mathbf{W}\mathbf{V})^{-\frac{1}{2}}(\mathbf{V}^T\mathbf{B}\mathbf{V})(\mathbf{V}^T\mathbf{W}\mathbf{V})^{-\frac{1}{2}}) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}, \mathbf{V} \in \mathbb{R}^{d\times p}, \end{array} \tag{2.1}$$

where $\mathcal{F}(\cdot) = \text{Tr}(\cdot)$ in (LDA-T) and $\mathcal{F}(\cdot) = \det(\cdot)$ in (LDA-D). In the above presentation we also used the cyclic property of the determinant and trace functions. We will show how to solve problem (2.1) for the class of problems in which $\mathcal{F}$ is an isotonic spectral function, whose definition is given below. This class of functions includes the trace and determinant functions as special cases.

Definition 2.1 *(isotonic spectral function).* A function $\mathcal{F} : \mathbb{S}^p_+ \to \mathbb{R}$ is called an **isotonic spectral function** if there exists a continuous function $\mathcal{G} : \mathbb{R}^p_+ \to \mathbb{R}$ such that

$$\mathcal{F}(\mathbf{H}) = \mathcal{G}(\lambda_1(\mathbf{H}), \ldots, \lambda_p(\mathbf{H})),$$

and $\mathcal{G}$ is isotonic over $\mathbb{R}^p_+$, meaning that $\mathcal{G}(\mathbf{x}) \ge \mathcal{G}(\mathbf{y})$ whenever $\mathbf{x} \ge \mathbf{y}$ and $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^p_+$.

Example 2.2. The determinant and trace functions are isotonic spectral functions since for any symmetric matrix $\mathbf{H} \in \mathbb{S}^p_+$ it holds that

$$\det(\mathbf{H}) = \prod_{i=1}^p \lambda_i(\mathbf{H}),$$

$$\text{Tr}(\mathbf{H}) = \sum_{i=1}^p \lambda_i(\mathbf{H}),$$

and both the product and sum functions are isotonic over the nonnegative orthant. Another class of isotonic spectral functions is the class of Schatten norms [6]. For any $m \ge 1$, the Schatten m-norm is the function $\|\mathbf{H}\|_{S_m} \equiv \sqrt[m]{\sum_{i=1}^p \lambda_i(\mathbf{H})^m}$ ($\mathbf{H} \in \mathbb{S}^p_+$), and the isotonicity of $\|\cdot\|_{S_m}$ follows by the isotonicity of the ℓ_m-norm over $\mathbb{R}^p_+$.

We begin by showing that problem (2.1) can be transformed into a problem over the Stiefel manifold.

Lemma 2.3. *Let $\mathbf{W} \in \mathbb{S}^d_{++}$, $\mathbf{B} \in \mathbb{S}^d_+$, and let $\mathcal{F} : \mathbb{S}^p_+ \to \mathbb{R}$ ($p \le d$) be an arbitrary function. Consider the following two optimization problems:*

$$\text{(P1)} \quad \begin{array}{ll} \max_{\mathbf{V}} & \mathcal{F}\left(\left(\mathbf{V}^T\mathbf{W}\mathbf{V}\right)^{-\frac{1}{2}}\mathbf{V}^T\mathbf{B}\mathbf{V}\left(\mathbf{V}^T\mathbf{W}\mathbf{V}\right)^{-\frac{1}{2}}\right) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}, \mathbf{V} \in \mathbb{R}^{d\times p}, \end{array}$$

$$\text{(P2)} \quad \begin{array}{ll} \max_{\mathbf{X}} & \mathcal{F}(\mathbf{X}^T\tilde{\mathbf{B}}\mathbf{X}) \\ \text{s.t.} & \mathbf{X}^T\mathbf{X} = \mathbf{I}_p, \mathbf{X} \in \mathbb{R}^{d\times p}, \end{array}$$

where $\tilde{\mathbf{B}} = \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$. Then if $\mathbf{X}^*$ is an optimal solution of (P2), then $\mathbf{V} = \mathbf{W}^{-1/2}\mathbf{X}^*$ is an optimal solution of (P1).

Proof. We begin by showing that (P1) is equivalent to the problem

$$\text{(Q1)} \quad \begin{array}{ll} \max_{\mathbf{V}} & \mathcal{F}(\mathbf{V}^T\mathbf{B}\mathbf{V}) \\ \text{s.t.} & \mathbf{V}^T\mathbf{W}\mathbf{V} = \mathbf{I}_p, \mathbf{V} \in \mathbb{R}^{d\times p} \end{array}$$

in the sense that any optimal solution of (Q1) is an optimal solution of (P1). Denote the objective functions of problems (P1) and (Q1) by

$$f_{P1}(\mathbf{V}) \equiv \mathcal{F}\left(\left(\mathbf{V}^T\mathbf{W}\mathbf{V}\right)^{-\frac{1}{2}}\mathbf{V}^T\mathbf{B}\mathbf{V}\left(\mathbf{V}^T\mathbf{W}\mathbf{V}\right)^{-\frac{1}{2}}\right),$$

$$f_{Q1}(\mathbf{V}) \equiv \mathcal{F}\left(\mathbf{V}^T\mathbf{B}\mathbf{V}\right)$$

respectively. Suppose that $\mathbf{V}^*$ is an optimal solution of problem (Q1). We will show that $\mathbf{V}^*$ is an optimal solution of (P1) as well. We first note that

$$\text{val}(Q1) = f_{Q1}(\mathbf{V}^*) = \mathcal{F}((\mathbf{V}^*)^T\mathbf{B}\mathbf{V}^*).$$

Also, since $(\mathbf{V}^*)^T\mathbf{W}\mathbf{V}^* = \mathbf{I}$, then

$$\begin{aligned} f_{P1}(\mathbf{V}^*) &= \mathcal{F}\left(\left((\mathbf{V}^*)^T\mathbf{W}\mathbf{V}^*\right)^{-\frac{1}{2}}(\mathbf{V}^*)^T\mathbf{B}\mathbf{V}^*\left((\mathbf{V}^*)^T\mathbf{W}\mathbf{V}^*\right)^{-\frac{1}{2}}\right) \\ &= \mathcal{F}((\mathbf{V}^*)^T\mathbf{B}\mathbf{V}^*) = \text{val}(Q1). \end{aligned} \tag{2.2}$$

Now, let $\mathbf{V}$ be any feasible solution of problem (P1), meaning that $\mathbf{V}^T\mathbf{W}\mathbf{V} \succ \mathbf{0}$. Then $\tilde{\mathbf{V}} \equiv \mathbf{V}(\mathbf{V}^T\mathbf{W}\mathbf{V})^{-\frac{1}{2}}$ is a feasible solution of (Q1), which by the optimality of $\mathbf{V}^*$ implies that

$$f_{Q1}(\tilde{\mathbf{V}}) \le \text{val}(Q1). \tag{2.3}$$

Finally, since $f_{Q1}(\tilde{\mathbf{V}}) = \mathcal{F}(\tilde{\mathbf{V}}^T\mathbf{B}\tilde{\mathbf{V}}) = \mathcal{F}((\mathbf{V}^T\mathbf{W}\mathbf{V})^{-\frac{1}{2}}\mathbf{V}^T\mathbf{B}\mathbf{V}(\mathbf{V}^T\mathbf{W}\mathbf{V})^{-\frac{1}{2}}) = f_{P1}(\mathbf{V})$ and $\text{val}(Q1) = f_{P1}(\mathbf{V}^*)$ (see (2.2)), then (2.3) translates to

$f_{P1}(\mathbf{V}) \le f_{P1}(\mathbf{V}^*)$.

As the above was proven for any feasible solution $\mathbf{V}$ of (P1), it follows that $\mathbf{V}^*$ is an optimal solution of (P1). The result now follows by noting that making the change of variable $\mathbf{X} = \mathbf{W}^{1/2}\mathbf{V}$ transforms problem (Q1) into problem (P2). □

Note that the equivalency result of Lemma 2.3 holds for an arbitrary function $\mathcal{F}$. In Theorem 2.5 below we show that when $\mathcal{F}$ is an isotonic spectral function, an optimal solution of problem (P2) is a matrix whose columns constitute an orthonormal set of p leading eigenvectors of $\tilde{\mathbf{B}}$. The proof of the theorem requires the eigenvalue interlacing theorem, which is now recalled.

Theorem 2.4 *(Eigenvalue Interlacing Theorem [11, p. 269]). Let $\mathbf{A} \in \mathbb{S}^d$ be symmetric and partitioned as*

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{D} \end{pmatrix},$$

where $\mathbf{B} \in \mathbb{S}^p$, $\mathbf{D} \in \mathbb{S}^{d-p}$, $\mathbf{C} \in \mathbb{R}^{p\times(d-p)}$. Then $\lambda_i(\mathbf{B}) \le \lambda_i(\mathbf{A})$ for any $i = 1, 2, \dots, p$.

Theorem 2.5. *Let $\mathbf{C} \in \mathbb{S}^d_+$. Consider the optimization problem*

$$\begin{array}{ll} \max_{\mathbf{X}} & \mathcal{F}(\mathbf{X}^T\mathbf{C}\mathbf{X}) \\ \text{s.t.} & \mathbf{X}^T\mathbf{X} = \mathbf{I}_p, \mathbf{X} \in \mathbb{R}^{d\times p} \end{array} \tag{2.4}$$

where $\mathcal{F} : \mathbb{S}^p_+ \to \mathbb{R}$ is an isotonic spectral function. Then

(a) any matrix whose columns constitute an orthonormal set of p leading eigenvector of $\mathbf{C}$ is an optimal solution of (2.4);
(b) the maximal value of (2.4) is $\mathcal{F}(\Omega)$ where $\Omega = diag(\lambda_1(\mathbf{C}), \dots, \lambda_p(\mathbf{C}))$.

Proof. We will use the simplified notation $\lambda_i = \lambda_i(\mathbf{C})$ for the ith eigenvalue of $\mathbf{C}$. Let $\mathbf{V} \equiv (\mathbf{v}_1, \dots, \mathbf{v}_p)$ with $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ being p orthogonal leading eigenvectors of $\mathbf{C}$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ respectively. Then,

$$\mathbf{V}^T\mathbf{C}\mathbf{V} = \text{diag}(\lambda_1, \dots, \lambda_p) \equiv \Omega. \tag{2.5}$$

For any $\mathbf{X} \in \mathbb{S}_{d,p}$, we can expand $\mathbf{X}$ to complete an orthonormal basis for $\mathbb{R}^d$:

$$\hat{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & \mathbf{X}' \end{pmatrix}, \quad \hat{\mathbf{X}}^T\hat{\mathbf{X}} = \mathbf{I}_d.$$

Then $\hat{\mathbf{X}}^T\mathbf{C}\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^T\mathbf{C}\mathbf{X} & \mathbf{X}^T\mathbf{C}\mathbf{X}' \\ \mathbf{X}'^T\mathbf{C}\mathbf{X}^T & \mathbf{X}'^T\mathbf{C}\mathbf{X}' \end{bmatrix}$.

We can now apply the interlacing eigenvalue theorem (Theorem 2.4) and obtain that

$$\lambda_i(\mathbf{X}^T\mathbf{C}\mathbf{X}) \le \lambda_i(\hat{\mathbf{X}}^T\mathbf{C}\hat{\mathbf{X}}), \quad i = 1, \dots, p. \tag{2.6}$$

Since $\hat{\mathbf{X}}$ is orthogonal, it holds that $\lambda_i(\hat{\mathbf{X}}^T\mathbf{C}\hat{\mathbf{X}}) = \lambda_i(\mathbf{C}) = \lambda_i$. Plugging this in (2.6), we conclude that

$$\lambda_i(\mathbf{X}^T\mathbf{C}\mathbf{X}) \le \lambda_i, \quad i = 1, \dots, p. \tag{2.7}$$

Since $\mathcal{F}$ is an isotonic spectral function, it follows that there exists a continuous and isotonic function $\mathcal{G} : \mathbb{R}^p_+ \to \mathbb{R}$ such that $\mathcal{F}(\mathbf{H}) \equiv \mathcal{G}(\lambda_1(\mathbf{H}), \dots, \lambda_p(\mathbf{H}))$. Thus, by (2.7) and the isotonicity of $\mathcal{G}$ over $\mathbb{R}^p_+$.

$$\begin{aligned} \mathcal{F}(\mathbf{X}^T\mathbf{C}\mathbf{X}) &= \mathcal{G}(\lambda_1(\mathbf{X}^T\mathbf{C}\mathbf{X}), \dots, \lambda_p(\mathbf{X}^T\mathbf{C}\mathbf{X})) \\ &\le \mathcal{G}(\lambda_1, \dots, \lambda_p) = \mathcal{F}(\Omega) \end{aligned}$$

The above holds for all $\mathbf{X} \in \mathbb{S}_{d,p}$, and thus, using (2.5),

$$\max_{\mathbf{X}\in\mathbb{S}_{d,p}} \mathcal{F}(\mathbf{X}^T\mathbf{C}\mathbf{X}) \le \mathcal{F}(\Omega) = \mathcal{F}(\mathbf{V}^T\mathbf{C}\mathbf{V}), \tag{2.8}$$

implying that the optimal value of problem (2.4) is $\mathcal{F}(\Omega)$, and that it is attained at $\mathbf{V}$. □

Combining Lemma 2.3 and Theorem 2.5, we finally obtain the main result showing that the optimal solution of (P1) can be expressed in terms of generalized eigenvalues of the matrix pair $(\mathbf{B}, \mathbf{W})$ as long as $\mathcal{F}$ is an isotonic spectral function.

Theorem 2.6. *Let $\mathbf{B} \in \mathbb{S}^d_+$, $\mathbf{W} \in \mathbb{S}^d_{++}$, and denote $\tilde{\mathbf{B}} = \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$. Let $\mathcal{F} : \mathbb{S}^p_+ \to \mathbb{R}$ ($p \le d$) be an isotonic spectral function. Then an optimal solution of problem (2.1) is any matrix $\mathbf{V}$ whose columns constitute a set of $\mathbf{W}$-orthonormal vectors that are p leading generalized eigenvalues of the matrix pair $(\mathbf{B}, \mathbf{W})$.*

Proof. By Lemma 2.3 and Theorem 2.5, for any matrix $\mathbf{X} \in \mathbb{S}_{d,p}$ whose columns constitute an orthonormal set of p leading eigenvectors of $\tilde{\mathbf{B}} \equiv \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, the matrix $\mathbf{V} = \mathbf{W}^{-1/2}\mathbf{X}$ is an optimal solution of (2.1). The result now follows by Remark 1.1. □

3. The trace ratio problem (FLDA)

In this section, we will present an algorithm for finding an optimal solution of the trace ratio problem (FLDA) (see (1.3)). We begin by introducing the method.

3.1. The FPR method

We recall the formulation of the trace ratio problem:

$$f_{\text{opt}} = \max_{\mathbf{X}\in\mathbb{S}_{d,p}} \frac{\text{Tr}(\mathbf{X}^T\mathbf{B}\mathbf{X})}{\text{Tr}(\mathbf{X}^T\mathbf{W}\mathbf{X})}, \tag{3.1}$$

where $\mathbf{B} \in \mathbb{S}^p_+, \mathbf{W} \in \mathbb{S}^p_{++}$. For the sake of simplicity, denote by $f_1(\mathbf{X})$ the denominator of the ratio, by $f_2(\mathbf{X})$ the nominator, and by $f(\mathbf{X})$ the ratio between $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$, meaning

$$\begin{aligned} &f_1(\mathbf{X}) \equiv \text{Tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}), \quad f_2(\mathbf{X}) \equiv \text{Tr}(\mathbf{X}^T\mathbf{W}\mathbf{X}), \\ &f(\mathbf{X}) \equiv \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}. \end{aligned} \tag{3.2}$$

Notice that (3.1) consists of maximizing a continuous function over a nonempty compact set, and therefore it has a maximal value, which we denote by f_{opt}. To define a method for solving (3.2), we begin by presenting a simple optimality condition for general problems consisting of maximizing the ratio of two functions:

$$\text{(G)} \quad \max\left\{ g(\mathbf{x}) \equiv \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} : \mathbf{x} \in S \right\},$$

where the only assumptions are that (1) g_1, g_2 are continuous; (2) S is compact and (3) $g_2(\mathbf{x}) > 0$ for any $\mathbf{x} \in S$. Then it is easy to show the following optimality condition.

Lemma 3.1. *The vector $\mathbf{x}^*$ is an optimal solution of (G) if and only if the following relation holds:*

$$\mathbf{x}^* \in \text{argmax}\{g_1(\mathbf{x}) - g(\mathbf{x}^*)g_2(\mathbf{x}) : \mathbf{x} \in S\}. \tag{3.3}$$

Moreover, $\max_{\mathbf{x}\in S}\{g_1(\mathbf{x}) - g(\mathbf{x}^*)g_2(\mathbf{x})\} = 0$.

Relation (3.3) naturally suggests the following fixed point method for solving (G):

$$\text{(FPR)}\quad \mathbf{x}_{k+1} \in \operatorname{argmax}\{g_1(\mathbf{x}) - g(\mathbf{x}_k)g_2(\mathbf{x}) : \mathbf{x} \in S\}.$$

We will refer to algorithm above as the FPR method as it is a Fixed Point method for Ratio optimization problems. The FPR method was suggested in [9] for solving regularized total least squares problems and it was generalized and further analyzed in [1] to problems consisting of optimizing a ratio of indefinite quadratic functions over a convex homogeneous quadratic constraint.

For the specific case of the trace ratio problem (3.1), the FPR method takes the form

$$\mathbf{X}_{k+1} \in \operatorname{argmax}\left\{f_1(\mathbf{X}) - f(\mathbf{X}_k)f_2(\mathbf{X}) : \mathbf{X} \in \mathbb{S}_{d,p}\right\}. \tag{3.4}$$

We call the above procedure as FPR-TR, as it is the FPR method employed on the trace ratio problem. The procedure can also be written slightly more explicitly as

$$\text{(FPR-TR)}\quad \mathbf{X}_{k+1} \in \operatorname*{argmax}_{\mathbf{X}\in\mathbb{S}_{d,p}} \left\{\operatorname{Tr}\left(\mathbf{X}^T(\mathbf{B} - f(\mathbf{X}_k)\mathbf{W})\mathbf{X}\right)\right\}. \tag{3.5}$$

The type of problem in FPR-TR is known to be solved via the eigenvectors of the associated matrix. This is recalled in the next lemma.

Lemma 3.2 *([8, p. 549]). Let $\mathbf{C} \in \mathbb{S}^d$. Then*

$$\max\{\operatorname{Tr}(\mathbf{X}^T\mathbf{C}\mathbf{X}) : \mathbf{X} \in \mathbb{S}_{d,p}\} = \textstyle\sum_{i=1}^p \lambda_i(\mathbf{C}),$$

and the optimal value is attained at a matrix whose columns form an orthonormal set of p leading eigenvectors of $\mathbf{C}$. Consequently, for any $\mathbf{X} \in \mathbb{S}_{d,p}$,

$$\textstyle\sum_{i=1}^p \lambda_{d-i+1}(\mathbf{C}) \le \operatorname{Tr}(\mathbf{X}^T\mathbf{C}\mathbf{X}) \le \sum_{i=1}^p \lambda_i(\mathbf{C}).$$

Using Lemma 3.2, we can conclude that an optimal solution of (3.5) is a matrix whose columns form an orthogonal set of p leading eigenvectors of $\mathbf{B} - f(\mathbf{X}_k)\mathbf{W}$. Thus, the FPR-TR method for solving the trace ratio problem (1.3) reiterates the following step:

FPR-TR. $\mathbf{X}_{k+1}$ is a $d \times p$ matrix whose columns form an orthonormal set of p leading eigenvectors of $\mathbf{B} - f(\mathbf{X}_k)\mathbf{W}$.

The FPR-TR method was first introduced in [10], where convergence to the global optimal solution was established. The work [12] was able to show a linear rate of convergence of the sequence of function values of the sequence generated by the method as well as a local quadratic convergence with [12], and a superlinear rate of convergence was discussed in [13]. An interpretation of the scheme as a Newton method was discussed in [8] In our analysis we will exploit the linear convergence rate result of the FPR-TR method that was established in [12].

Theorem 3.3 *([12, Theorem 5.1]). Let $\{\mathbf{X}_k\}_{k\ge 0}$ be the sequence generated by (3.4) and $\mathbf{X}^*$ be an optimal solution of problem (3.1). Then*

$$f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_{k+1}\right) \le \left(1 - \frac{1}{\kappa(\mathbf{W})}\right)\left(f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_k\right)\right), \tag{3.6}$$

where $\kappa(\mathbf{W}) \equiv \frac{\sum_{i=1}^p \lambda_i(\mathbf{W})}{\sum_{i=1}^p \lambda_{d-i+1}(\mathbf{W})}$.

3.2. A quadratic growth inequality

We will first prove a quadratic growth inequality, which we will use in the next section in order to prove an $O((1-\kappa(\mathbf{W})^{-1})^{k^2})$ convergence rate of the algorithm given in (3.4). Suppose that $\mathbf{C} \in \mathbb{S}^d$. Consider the problem

$$g_{\text{opt}} = \max_{\mathbf{X}\in\mathbb{S}_{d,p}} \{g(\mathbf{X}) \equiv \operatorname{Tr}(\mathbf{X}^T\mathbf{C}\mathbf{X})\}. \tag{3.7}$$

The optimal value of the problem (see Lemma 3.2) is $g_{\text{opt}} \equiv \sum_{i=1}^p \lambda_i(\mathbf{C})$. Let Ω be the set of all optimal solutions of problem (3.7). Our objective is to show that there exists a positive constant α such that for all $\mathbf{X} \in \mathbb{S}_{d,p}$,

$$g_{\text{opt}} - g(\mathbf{X}) \ge \alpha \cdot \min_{\mathbf{X}^*\in\Omega} \|\mathbf{X} - \mathbf{X}^*\|_F^2. \tag{3.8}$$

The above inequality is often referred to as “a quadratic growth condition”, and it is well known to hold for convex problems with a strongly convex objective. A study of this property, as well as other related properties can be found in [7]. Although our problem is not convex and the objective function is not strongly convex, we will show that we can still prove it under the mild condition $\lambda_p(\mathbf{C}) > \lambda_{p+1}(\mathbf{C})$. This is done in Theorem 3.5. The inequality will be key in establishing the $O(q^{k^2})$ of the FPR-TR method for the choice of $q = 1 - \kappa(\mathbf{W})^{-1}$. We first require to establish the following technical lemma.

Lemma 3.4. *For any $\mathbf{H} \in \mathbb{R}^{p\times p}$, it holds that*

$$\max\{\operatorname{Tr}(\mathbf{R}^T\mathbf{H}) : \mathbf{R} \in \mathbb{O}_p\} = \textstyle\sum_{i=1}^p \sigma_i(\mathbf{H}). \tag{3.9}$$

Proof. Let the singular value decomposition of $\mathbf{H}$ be given by $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{U}, \mathbf{S}, \mathbf{V} \in \mathbb{R}^{p\times p}$ are such that $\mathbf{U}$, $\mathbf{V} \in \mathbb{O}_p$ and $\mathbf{S} = \operatorname{diag}(\sigma_1(\mathbf{H}), \dots, \sigma_p(\mathbf{H}))$. Then for any $\mathbf{R} \in \mathbb{O}^p$,

$$\begin{aligned}\operatorname{Tr}(\mathbf{R}^T\mathbf{H}) &= \operatorname{Tr}\left(\mathbf{R}^T\mathbf{U}\mathbf{S}\mathbf{V}^T\right) = \operatorname{Tr}\left((\mathbf{R}^T\mathbf{U}\mathbf{S}^{\frac{1}{2}})(\mathbf{V}\mathbf{S}^{\frac{1}{2}})^T\right)\\ &\le \|\mathbf{R}^T\mathbf{U}\mathbf{S}^{\frac{1}{2}}\|_F\|\mathbf{V}\mathbf{S}^{\frac{1}{2}}\|_F\\ &= \|\mathbf{S}^{\frac{1}{2}}\|_F\|\mathbf{S}^{\frac{1}{2}}\|_F = \textstyle\sum_{i=1}^p \sigma_i(\mathbf{H}),\end{aligned} \tag{3.10}$$

where we used Cauchy-Schwarz inequality and the fact that the Frobenius norm is invariant under orthogonal transformations. Finally, $\mathbf{R} = \tilde{\mathbf{R}} \equiv \mathbf{U}\mathbf{V}^T \in \mathbb{O}^p$ attains the upper bound (3.10) as it satisfies

$$\begin{aligned}\operatorname{Tr}(\tilde{\mathbf{R}}^T\mathbf{H}) &= \operatorname{Tr}\left(\tilde{\mathbf{R}}^T\mathbf{U}\mathbf{S}\mathbf{V}^T\right) = \operatorname{Tr}\left(\mathbf{V}^T\tilde{\mathbf{R}}^T\mathbf{U}\mathbf{S}\right)\\ &= \operatorname{Tr}\left(\mathbf{V}^T\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{S}\right)\\ &= \operatorname{Tr}(\mathbf{S}) = \textstyle\sum_{i=1}^p \sigma_i(\mathbf{H}). \quad \square\end{aligned}$$

Theorem 3.5 *(quadratic growth). Let $\mathbf{C} \in \mathbb{S}^p$ and denote its eigenvectors by $\lambda_i = \lambda_i(\mathbf{C}), i = 1, 2, ..., p$. Let Ω be the optimal set of problem (3.7). Then,*

$$\sum_{i=1}^p \lambda_i - g(\mathbf{X}) \ge \frac{\lambda_p - \lambda_{p+1}}{2} \cdot \min_{\mathbf{X}^*\in\Omega} \|\mathbf{X} - \mathbf{X}^*\|_F^2. \tag{3.11}$$

Proof. Let $\{\mathbf{v}_i\}_{i=1}^d \subseteq \mathbb{R}^d$ be an orthonormal basis of $\mathbb{R}^d$ such that $\mathbf{v}_i$ is an eigenvector of $\mathbf{C}$ corresponding to λ_i, and define the matrices

$$\mathbf{V}_1 = (\mathbf{v}_1, \dots, \mathbf{v}_p), \mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_d).$$

Note that $\mathbf{V}$ is an orthogonal matrix, and in particular,

$$\textstyle\sum_{i=1}^d \mathbf{v}_i\mathbf{v}_i^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}. \tag{3.12}$$

Now, for any $\mathbf{X} \in \mathbb{S}_{d,p}$,

$$\begin{aligned}\sum_{i=1}^{p}\lambda_i-\mathrm{Tr}(\mathbf{X}^T\mathbf{C}\mathbf{X}) &= \sum_{i=1}^{p}\lambda_i-\mathrm{Tr}\left(\mathbf{X}^T\left[\sum_{i=1}^{d}\lambda_i\mathbf{v}_i\mathbf{v}_i^T\right]\mathbf{X}\right)\\ &= \sum_{i=1}^{p}\lambda_i-\sum_{i=1}^{d}\lambda_i\|\mathbf{X}^T\mathbf{v}_i\|_2^2\\ &= \sum_{i=1}^{p}\lambda_i(1-\|\mathbf{X}^T\mathbf{v}_i\|_2^2)-\sum_{i=p+1}^{d}\lambda_i\|\mathbf{X}^T\mathbf{v}_i\|_2^2\\ &\overset{\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d}{\geq} \lambda_p\left(p-\sum_{i=1}^{p}\|\mathbf{X}^T\mathbf{v}_i\|_2^2\right)-\lambda_{p+1}\left(\sum_{i=p+1}^{d}\|\mathbf{X}^T\mathbf{v}_i\|_2^2\right)\\ &= (\lambda_p-\lambda_{p+1})\left(p-\sum_{i=1}^{p}\|\mathbf{X}^T\mathbf{v}_i\|_2^2\right), \qquad (3.13)\end{aligned}$$

where the last equality follows from the following argument:

$$\begin{aligned}\sum_{i=1}^{d}\|\mathbf{X}^T\mathbf{v}_i\|_2^2 &= \sum_{i=1}^{d}\mathrm{Tr}(\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{X})\\ &= \mathrm{Tr}\left(\mathbf{X}^T\left[\sum_{i=1}^{d}\mathbf{v}_i\mathbf{v}_i^T\right]\mathbf{X}\right)\overset{(3.12)}{=}\mathrm{Tr}(\mathbf{X}^T\mathbf{X})\\ &= \mathrm{Tr}(\mathbf{I}_p)=p.\end{aligned}$$

On the other hand, note that $\{\mathbf{V}_1\mathbf{R}:\mathbf{R}\in\mathbb{O}_p\}\subseteq\Omega$, and thus,

$$\begin{aligned}&\min_{\mathbf{X}^*}\{\|\mathbf{X}-\mathbf{X}^*\|_2^2:\mathbf{X}^*\in\Omega\}\leq\min_{\mathbf{R}\in\mathbb{O}_p}\{\|\mathbf{X}-\mathbf{V}_1\mathbf{R}\|_F^2\}\\ &=\min_{\mathbf{R}\in\mathbb{O}_p}\{\|\mathbf{X}\|_F^2+\|\mathbf{V}_1\mathbf{R}\|_F^2-2\,\mathrm{Tr}(\mathbf{R}^T\mathbf{V}_1^T\mathbf{X})\}\\ &=\min_{\mathbf{R}\in\mathbb{O}_p}\{2p-2\,\mathrm{Tr}(\mathbf{R}^T\mathbf{V}_1^T\mathbf{X})\}\\ &=2p-2\max_{\mathbf{R}\in\mathbb{O}_p}\{\mathrm{Tr}(\mathbf{R}^T\mathbf{V}_1^T\mathbf{X})\} \qquad [\mathbf{X},\mathbf{V}_1\mathbf{R}\in\mathbb{S}_{d,p}]\\ &=2p-2\sum_{i=1}^{p}\sigma_i(\mathbf{V}_1^T\mathbf{X}) \qquad [\text{Lemma } 3.4]\\ &\leq 2p-2\sum_{i=1}^{p}\sigma_i^2(\mathbf{V}_1^T\mathbf{X}), \qquad (3.14)\end{aligned}$$

where the last inequality follows by the fact that $\|\mathbf{V}_1^T\mathbf{X}\|_2\leq\|\mathbf{V}_1\|_2\|\mathbf{X}\|_2=1$, and therefore $\sigma_i(\mathbf{V}_1^T\mathbf{X})\leq 1$ for any i. Note that $\sum_{i=1}^{p}\sigma_i^2(\mathbf{V}_1^T\mathbf{X})=\|\mathbf{V}_1^T\mathbf{X}\|_F^2=\sum_{i=1}^{p}\|\mathbf{X}^T\mathbf{v}_i\|_2^2$, which together with (3.14), implies the inequality

$$\min_{\mathbf{X}^*}\{\|\mathbf{X}-\mathbf{X}^*\|_2^2:\mathbf{X}^*\in\Omega\}\leq 2p-2\textstyle\sum_{i=1}^{p}\|\mathbf{X}^T\mathbf{v}_i\|_2^2. \qquad (3.15)$$

Finally, combining (3.13) and (3.15), the desired result (3.11) readily follows. □

3.3. $O(q^{k^2})$ convergence rate

We will use the quadratic growth condition obtained in Theorem 3.5 to show an $O(q^{k^2})$ convergence rate of the sequence generated by (3.5).

Theorem 3.6. *Suppose that* $\mathbf{B}\in\mathbb{S}_+^d$, $\mathbf{W}\in\mathbb{S}_{++}^d$. *Let* $\{\mathbf{X}_k\}_{k\geq 0}$ *be the sequence generated by The FPR-TR method* (3.5) *for solving problem (3.1). Let* $\tilde{\lambda}_1\geq\cdots\geq\tilde{\lambda}_d$ *be the eigenvalues of the matrix* $\mathbf{B}-f_{\mathrm{opt}}\mathbf{W}$. *Then under the assumption that* $\tilde{\lambda}_p>\tilde{\lambda}_{p+1}$, *it holds that*

$$f_{\mathrm{opt}}-f(\mathbf{X}_n)\leq\left(f_{\mathrm{opt}}-f(\mathbf{X}_0)\right)D^n\left(1-\frac{1}{\kappa(\mathbf{W})}\right)^{\frac{n^2}{4}}, \qquad (3.16)$$

where

$$D=\sqrt{p}\,\frac{\lambda_{\max}(\mathbf{W})}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})}\sqrt{\frac{8\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\tilde{\lambda}_p-\tilde{\lambda}_{p+1}}}\cdot\sqrt[4]{1-\frac{1}{\kappa(\mathbf{W})}}\sqrt{f_{\mathrm{opt}}-f(\mathbf{X}_0)}$$

and $\kappa(\mathbf{W})\equiv\frac{\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})}$.

Proof. Denote $g(\mathbf{X})\equiv f_2(\mathbf{X})(f(\mathbf{X})-f_{\mathrm{opt}})=\mathrm{Tr}(\mathbf{X}^T(\mathbf{B}-f_{\mathrm{opt}}\mathbf{W})\mathbf{X})$, and let

$$\Omega=\mathrm{argmax}\{g(\mathbf{X}):\mathbf{X}\in\mathbb{S}_{d,p}\}.$$

Then using Lemma 3.1, $\sum_{i=1}^{p}\tilde{\lambda}_i=0$. By Theorem 3.5, for all $\mathbf{X}\in\mathbb{S}_{d,p}$,

$$\begin{aligned}&\frac{\tilde{\lambda}_p-\tilde{\lambda}_{p+1}}{2}\min_{\mathbf{X}^*\in\Omega}\|\mathbf{X}-\mathbf{X}^*\|_F^2\leq\textstyle\sum_{i=1}^{p}\tilde{\lambda}_i-g(\mathbf{X}) \qquad (3.17)\\ &=-g(\mathbf{X})=f_2(\mathbf{X})(f_{\mathrm{opt}}-f(\mathbf{X})).\end{aligned}$$

Using Lemma 3.2, for all $\mathbf{X}\in\mathbb{S}_{d,p}$ it holds that $f_2(\mathbf{X})=\mathrm{Tr}(\mathbf{X}^T\mathbf{W}\mathbf{X})\leq\sum_{i=1}^{p}\lambda_i(\mathbf{W})$, and under the assumption that $\tilde{\lambda}_p>\tilde{\lambda}_{p+1}$, (3.17) becomes

$$\min_{\mathbf{X}^*\in\Omega}\|\mathbf{X}-\mathbf{X}^*\|_F^2\leq\frac{2\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\tilde{\lambda}_p-\tilde{\lambda}_{p+1}}(f_{\mathrm{opt}}-f(\mathbf{X})). \qquad (3.18)$$

Plugging $\mathbf{X}=\mathbf{X}_k$ in the above and invoking (3.6), we obtain

$$\begin{aligned}&\min_{\mathbf{X}^*\in\Omega}\|\mathbf{X}_k-\mathbf{X}^*\|_F^2\leq\left(\frac{2\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\tilde{\lambda}_p-\tilde{\lambda}_{p+1}}\right)(f_{\mathrm{opt}}-f(\mathbf{X}_k))\\ &\leq\left(\frac{2\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\tilde{\lambda}_p-\tilde{\lambda}_{p+1}}\right)(f_{\mathrm{opt}}-f(\mathbf{X}_0))\left(1-\frac{1}{\kappa(\mathbf{W})}\right)^k. \qquad (3.19)\end{aligned}$$

Let $\mathbf{X}^*\in\Omega$. Using the mean value theorem, there exists $\omega\in[0,1]$ such that,

$$\begin{aligned}&\left|f_2\left(\mathbf{X}^*\right)-f_2\left(\mathbf{X}_k\right)\right|=\left|\langle\nabla f_2\left(\omega\mathbf{X}_k+(1-\omega)\mathbf{X}^*\right),\mathbf{X}_k-\mathbf{X}^*\rangle\right|\\ &\leq 2\left\|\mathbf{W}\left(\omega\mathbf{X}_k+(1-\omega)\mathbf{X}^*\right)\right\|_F\left\|\mathbf{X}_k-\mathbf{X}^*\right\|_F\\ &\leq 2\|\mathbf{W}\|_2\left(\|\omega\mathbf{X}_k\|_F+\|(1-\omega)\mathbf{X}^*\|_F\right)\|\mathbf{X}_k-\mathbf{X}^*\|_F\\ &=2\lambda_{\max}(\mathbf{W})\left(\|\omega\mathbf{X}_k\|_F+\|(1-\omega)\mathbf{X}^*\|_F\right)\|\mathbf{X}_k-\mathbf{X}^*\|_F\\ &\leq 2\sqrt{p}\lambda_{\max}(\mathbf{W})\|\mathbf{X}_k-\mathbf{X}^*\|_F, \qquad (3.20)\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality, the second is a result of the triangle inequality and the inequality $\|\mathbf{A}\mathbf{B}\|_F\leq\|\mathbf{A}\|_2\|\mathbf{B}\|_F$, and the last inequality follows by the fact that $\mathbf{X}_k,\mathbf{X}^*\in\mathbb{S}_{d,p}$.

By the definition of $\mathbf{X}_{k+1}$ as the maximizer of (3.4), $f_2\left(\mathbf{X}_{k+1}\right)\left(f\left(\mathbf{X}_{k+1}\right)-f\left(\mathbf{X}_k\right)\right)\geq f_2\left(\mathbf{X}^*\right)\left(f\left(\mathbf{X}^*\right)-f\left(\mathbf{X}_k\right)\right)$. Dividing this inequality by $-f_2(\mathbf{X}_{k+1})$ leads to

$$
\begin{aligned}
&f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_{k+1}\right) \leq \\
&\leq \frac{f_2\left(\mathbf{X}^*\right)}{f_2\left(\mathbf{X}_{k+1}\right)}\left(f\left(\mathbf{X}_k\right) - f\left(\mathbf{X}^*\right)\right) + f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_k\right) \\
&= \left(\frac{f_2\left(\mathbf{X}_{k+1}\right) - f_2\left(\mathbf{X}^*\right)}{f_2\left(\mathbf{X}_{k+1}\right)}\right)\left(f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_k\right)\right).
\end{aligned}
\tag{3.21}
$$

Consequently,

$$
\begin{aligned}
\frac{f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_{k+1}\right)}{f\left(\mathbf{X}^*\right) - f\left(\mathbf{X}_k\right)} &\overset{(3.21)}{\leq} \frac{f_2\left(\mathbf{X}_{k+1}\right) - f_2\left(\mathbf{X}^*\right)}{f_2\left(\mathbf{X}_{k+1}\right)} \\
&\overset{(3.20),\text{Lemma 3.2}}{\leq} \frac{2\sqrt{p}\lambda_{\max}(\mathbf{W})\|\mathbf{X}_{k+1} - \mathbf{X}^*\|_F}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})} \\
&= \quad 2\sqrt{p}\frac{\lambda_{\max}(\mathbf{W})}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})}\|\mathbf{X}_{k+1} - \mathbf{X}^*\|_F.
\end{aligned}
$$

Since the above holds for any $\mathbf{X}^* \in \Omega$ and for any such $\mathbf{X}^*$, $f\left(\mathbf{X}^*\right) = f_{\text{opt}}$, we conclude that

$$
\frac{f_{\text{opt}} - f\left(\mathbf{X}_{k+1}\right)}{f_{\text{opt}} - f\left(\mathbf{X}_k\right)} \leq 2\sqrt{p}\frac{\lambda_{\max}(\mathbf{W})}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})}\min_{\mathbf{X}^* \in \Omega}\|\mathbf{X}_{k+1} - \mathbf{X}^*\|_F.
$$

Combining the above with (3.19), and by denoting

$$
\gamma \equiv 1 - 1/\kappa(\mathbf{W})
$$

and

$$
D_1 \equiv 2\sqrt{p}\frac{\lambda_{\max}(\mathbf{W})}{\sum_{i=1}^{p}\lambda_{d-i+1}(\mathbf{W})}\sqrt{\frac{2\sum_{i=1}^{p}\lambda_i(\mathbf{W})}{\tilde{\lambda}_p - \tilde{\lambda}_{p+1}}}\sqrt{f_{\text{opt}} - f(\mathbf{X}_0)},
$$

we obtain that

$$
\frac{f_{\text{opt}} - f\left(\mathbf{X}_{k+1}\right)}{f_{\text{opt}} - f\left(\mathbf{X}_k\right)} \leq D_1\left(1 - \frac{1}{\kappa(\mathbf{W})}\right)^{\frac{k+1}{2}} = D_1\gamma^{\frac{k+1}{2}}.
$$

Then for any $n \geq 1$,

$$
\frac{f_{\text{opt}} - f\left(\mathbf{X}_n\right)}{f_{\text{opt}} - f\left(\mathbf{X}_0\right)} = \prod_{k=0}^{n-1}\frac{f_{\text{opt}} - f\left(\mathbf{X}_{k+1}\right)}{f_{\text{opt}} - f\left(\mathbf{X}_k\right)} \leq \prod_{k=0}^{n-1} D_1\gamma^{\frac{k+1}{2}} = D_1^n\gamma^{\frac{n^2+n}{4}},
$$

from which the desired result (3.16) follows. □

Acknowledgements

This research is supported by the Israel Science Foundation (Grant No. 926/21).

References

[1] A. Beck, M. Teboulle, A convex optimization approach for minimizing the ratio of indefinite quadratic functions over an ellipsoid, Math. Program. 118 (1) (2009) 13–35.

[2] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[3] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.

[4] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.

[5] K. Fukunaga, Introduction to Statistical Pattern Recognition, Elsevier, 2013.

[6] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 2012.

[7] I. Necoara, Y. Nesterov, F. Glineur, Linear convergence of first order methods for non-strongly convex optimization, Math. Program., Ser. A 175 (1–2) (2019) 69–107.

[8] T.T. Ngo, M. Bellalij, Y. Saad, The trace ratio optimization problem, SIAM Rev. 54 (3) (2012) 545–569.

[9] D.M. Sima, S. Van Huffel, G.H. Golub, Regularized total least squares based on quadratic eigenvalue problem solvers, BIT 44 (4) (2004) 793–812.

[10] H. Wang, S. Yan, D. Xu, X. Tang, T. Huang, Trace ratio vs. ratio trace for dimensionality reduction, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[11] F. Zhang, Matrix Theory: Basic Results and Techniques, Springer Science & Business Media, 2011.

[12] L.H. Zhang, L.Z. Liao, M.K. Ng, Fast algorithms for the generalized Foley-Sammon discriminant analysis, SIAM J. Matrix Anal. Appl. 31 (4) (2009/10) 1584–1605.

[13] L.H. Zhang, L.Z. Liao, M.K. Ng, Superlinear convergence of a general algorithm for the generalized Foley-Sammon discriminant analysis, J. Optim. Theory Appl. 157 (3) (2013) 853–865.