

FAST ALGORITHMS FOR THE GENERALIZED FOLEY-SAMMON DISCRIMINANT ANALYSIS

LEI-HONG ZHANG*, LI-ZHI LIAO†, AND MICHAEL K. NG‡

Abstract. Linear Discriminant Analysis (LDA) is one of the most popular approaches for feature extraction and dimension reduction to overcome the curse of the dimensionality of the high-dimensional data in many applications of data mining, machine learning, and bioinformatics. In this paper, we made two main contributions to an important LDA scheme, the generalized Foley-Sammon transform (GFST [7, 13], or a trace ratio model [28]) and its regularization (RGFST) which handles the undersampled problem that involves small samples size n but with high number of features N ($N > n$) and arises frequently in many modern applications. Our first main result is to establish an equivalent reduced model for the RGFST which effectively improves the computational overhead. The iteration method proposed in [28] is applied to solve the GFST or the reduced RGFST. It has been proven [28] that this iteration converges globally and fast convergence was observed numerically, but there is no theoretical analysis on the convergence rate thus far. Our second main contribution completes this important and missing piece by proving the quadratic convergence even under two kinds of inexact computations. Practical implementations including computational complexity and storage requirement are also discussed. Our experimental results on several real world data sets indicate the efficiency of the algorithm and the advantages of the GFST model in classification.

Key words. Dimension reduction, linear discriminant analysis, regularization, Foley-Sammon transform, global convergence, quadratic convergence

AMS subject classifications. 65F15, 65F30, 62H30, 15A18

1. Introduction.

1.1. Background of LDA. A lot of practical applications of data mining, machine learning, bioinformatics require to deal with the high-dimensional data efficiently. Modern data sets are always very huge and hence dimension reduction seems imperative for efficiently manipulating and analyzing the massive quantity of data. Feature reduction commonly aims at reducing the dimension of the original features, while preserving the useful and necessary information as much as possible. It is largely applied in many applications and acts frequently as a preprocessing step to overcome the tremendous dimensionality. A lot of methods in this area, for example, the Principal Analysis (PCA) ([18]), Linear Discriminant Analysis (LDA) ([9]), have been proposed from different points of view and applied successfully in practice. LDA is one of the most popular approaches in pattern recognition (e.g., [9, 27]) whose goal is to find a proper linear transformation so that each sample vector with high dimension is projected into a low dimension vector, while preserving the original cluster structure as much as possible.

More precisely, suppose we are given a data matrix $A \in \mathbb{R}^{N \times n}$ in which each column $\mathbf{a}_i \in \mathbb{R}^N$ ($i = 1, 2, \dots, n$) corresponds to a training sample, while each row corresponds to a particular feature. In general, the number of the features N is very large and hence makes the analysis based

*Department of Applied Mathematics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, P. R. China (longzlh@gmail.com).

†Corresponding author. Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, P. R. China (liliao@hkbu.edu.hk). The work of this author has been partially supported by FRG grants from Hong Kong Baptist University and the Research Grant Council of Hong Kong.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. The work of this author has been partially supported by RGC grants 201508 and HKBU FRGs.

on this data rather difficult and inefficient. What we expect then is a linear transformation, say $G \in \mathbb{R}^{N \times l}$ (generally $l \ll N$), so that it maps each sample $\mathbf{a}_i \in \mathbb{R}^N$ in the data matrix A to a new reduced ‘sample’

$$\mathbf{y}_i = G^T \mathbf{a}_i \in \mathbb{R}^l.$$

The technique to employ a linear transformation to reduce the features also appears in maximal correlation analysis (MCA) where coefficients are to be determined so that the resulting linear combinations of sets of random variables are maximally correlated (see e.g., [3]). MCA differs greatly from LDA in that MCA deals with only features and pays no attention to the classes (or classification); LDA, however, focuses on how to find the optimal linear transformation G to preserve the cluster structure in A .

Suppose $A = [A_1, \dots, A_c] \in \mathbb{R}^{N \times n}$, where each $A_j \in \mathbb{R}^{N \times n_j}$ for $j = 1, \dots, c$, $c \leq n$, represents an independent class data set and n_j denotes the number of the samples of the j th class in A and $\sum_{j=1}^c n_j = n$. Define

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in A_j} \mathbf{x} = \frac{1}{n_j} A_j \mathbf{e}^{(j)}, \quad \text{where } \mathbf{e}^{(j)} = (1, \dots, 1)^T \in \mathbb{R}^{n_j}$$

to be the *centroid* of cluster A_j , and

$$\mathbf{m} = \frac{1}{n} \sum_{j=1}^c \sum_{\mathbf{x} \in A_j} \mathbf{x} = \frac{1}{n} A \mathbf{e}, \quad \text{where } \mathbf{e} = (1, \dots, 1)^T \in \mathbb{R}^n$$

be the *global centroid* of all objects. Then the *within-class scatter matrix* S_w , the *between-class scatter matrix* S_b , and the *total scatter matrix* S_t [9] are defined as

$$S_w = \frac{1}{n} \sum_{j=1}^c \sum_{\mathbf{x} \in A_j} (\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T = H_w H_w^T \in \mathbb{R}^{N \times N}, \quad (1.1)$$

$$S_b = \frac{1}{n} \sum_{j=1}^c n_j (\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T = H_b H_b^T \in \mathbb{R}^{N \times N}, \quad (1.2)$$

$$S_t = \frac{1}{n} \sum_{j=1}^c (\mathbf{a}_j - \mathbf{m})(\mathbf{a}_j - \mathbf{m})^T = H_t H_t^T \in \mathbb{R}^{N \times N}, \quad (1.3)$$

respectively, where

$$\begin{aligned} H_w &= \frac{1}{\sqrt{n}} [A_1 - \mathbf{m}_1 (\mathbf{e}^{(1)})^T, \dots, A_c - \mathbf{m}_c (\mathbf{e}^{(c)})^T] \in \mathbb{R}^{N \times n}, \\ H_b &= \frac{1}{\sqrt{n}} [\sqrt{n_1} (\mathbf{m}_1 - \mathbf{m}), \dots, \sqrt{n_c} (\mathbf{m}_c - \mathbf{m})] \in \mathbb{R}^{N \times c}, \\ H_t &= \frac{1}{\sqrt{n}} (A - \mathbf{m} \mathbf{e}^T) \in \mathbb{R}^{N \times n}. \end{aligned}$$

It is easy to verify ([9]) that

$$S_t = S_b + S_w. \quad (1.4)$$

To measure the within-class cohesion as well as the between-class separation, the trace operator is introduced. Therefore, for a given linear transformation G , $\text{tr}(G^T S_w G)$ and $\text{tr}(G^T S_b G)$ then measure the within-class cohesion and the between-class separation in the projected lower-dimensional space

respectively. The goal of the LDA is then to find a proper G , minimizing the within-class cohesion and maximizing the between-class separation simultaneously in the projected lower-dimensional space. To this end, different criteria have been proposed and studied in the literature, including

$$F_1(G) = \text{tr}((G^T S_w G)^{-1} (G^T S_b G)) \quad (\text{e.g., [4, 9, 15, 16, 24, 32]}), \quad (1.5)$$

$$F_2(G) = \frac{\text{tr}(G^T S_b G)}{\text{tr}(G^T S_w G)} \quad (\text{e.g., [7, 9, 13, 28]}), \quad (1.6)$$

$$F_3(G) = \text{tr}(G^T S_b G) - \beta \text{tr}(G^T S_w G), \quad \beta > 0 \quad (\text{e.g., [21, 22, 29]}). \quad (1.7)$$

It should be noted that criteria $F_1(G)$ and $F_2(G)$ can be regarded as the generalizations of the Fisher linear discriminant [6] for two-class problems. By maximizing one criterion in some proper subset of $\mathbb{R}^{N \times l}$, the corresponding optimal linear transformation G can be obtained. The orthogonal LDA (OLDA) (e.g., [31, 32]) is to find an orthonormal linear transformation G^* ; therefore, for criteria (1.5), (1.6), and (1.7), we have the following corresponding optimization problems

$$G^* = \arg \max_{G^T G = I_l} F_i(G), \quad i = 1, 2, 3, \quad (1.8)$$

where $I_l \in \mathbb{R}^{l \times l}$ denotes the l -by- l identity matrix. It is well-known that if S_w is nonsingular, the columns of the optimal G^* of (1.8) for the popular criterion $F_1(G)$ are the orthonormal eigenvectors of $S_w^{-1} S_b$ corresponding to its l -largest eigenvalues (see e.g., [4, 9, 16]), and the columns of the solution to (1.8) with $F_3(G)$ are the orthonormal eigenvectors of $(S_b - \beta S_w)$ corresponding to its l -largest eigenvalues (see [22, 30]). The criterion (1.8) with $F_2(G)$ which has been studied, e.g., in [13, 28], as the generalized Foley-Sammon transform [7] (GFST), is claimed to possess the preferred discriminant ability in global sense, and [28] further discusses the advantages of trace ratio $F_2(G)$ over ratio trace $F_1(G)$. As shown in [28], the GFST is the original model that is derived from the unified framework of most dimensionality reduction algorithms, namely graph embedding [29]; the criterion $F_1(G)$, however, is only its alternative, and may deviate from the original objectives and suffers from the fact that it is invariant under any non-singular transformation, which may lead to uncertainty in subsequent processing such as classification and clustering. When S_w is nonsingular, the columns of any global solution of (1.8) with $F_2(G)$ are the orthonormal eigenvectors of $(S_b - F_2^* S_w)$ corresponding to its l -largest eigenvalues, where F_2^* is the optimal objective function value of (1.8), and therefore, this criterion in this case is actually a special case of $F_3(G)$ with $\beta = F_2^*$; see Section 3 and [13].

However, both $F_1(G)$ and $F_2(G)$ suffer from the singularity of S_w , which is always the case for the *undersampled problem* where $N > n$ (see (1.1)). In many applications, collecting data is expensive, and it then involves high-dimensional data with small samples, i.e., $N \gg n$. Such is the case for the image databases of facial recognition, gene expression data, as well as the text documents, in which N could be up to several thousands and S_w becomes singular. Various approaches (e.g., [8, 9, 15, 16, 24, 26, 31, 34]) have been proposed to overcome this difficulty, and among them, a simple remedy is just to apply the regularization technique [8] by adding a regularized term μI_N , ($\mu > 0$), to S_w , hence arriving at

$$\max_{G^T G = I_l} \text{tr}((G^T (S_w + \mu I_N) G)^{-1} G^T S_b G), \quad (1.9)$$

for $F_1(G)$, and

$$\text{RGFST} : \max_{G^T G = I_l} \frac{\text{tr}(G^T S_b G)}{\text{tr}(G^T S_w G) + \mu l}, \quad (1.10)$$

for $F_2(G)$, where $\mu > 0$ is known as the *regularization parameter*. According to the recent work in [24], a reduced model of (1.9) for the undersampled problem can be established as the first stage

$$\max_{U^T U = I_l, U \in \mathbb{R}^{n \times l}} \text{tr}((U^T (\hat{S}_w + \mu I_n) U)^{-1} U^T \hat{S}_b U), \quad (1.11)$$

where $\hat{S}_w = Q_1^T S_w Q_1 \in \mathbb{R}^{n \times n}$, $\hat{S}_b = Q_1^T S_b Q_1 \in \mathbb{R}^{n \times n}$, and $Q_1 \in \mathbb{R}^{N \times n}$ is from the reduced QR decomposition of A , i.e., $A = Q_1 R$. Based on (1.11) and the technique used in building the LDA/QR-regGSVD [24], a fast and efficient algorithm can be easily established for (1.9) as the second stage. Apart from the iteration used in GSVD, the approach is direct and roughly requires the orders of storage $\mathcal{O}(Nn)$ and of computational complexity $\mathcal{O}(Nn^2)$.

1.2. Main contributions. The present paper is concerned with the GFST for the general case (S_w is positive definite) as well as the RGFST for the undersampled problem. For the latter case, following the same idea in building the reduced model (1.11) for (1.9), our first main result is to establish an equivalent reduced model as the first stage

$$\max_{U^T U = I_l, U \in \mathbb{R}^{n \times l}} \frac{\text{tr}(U^T \hat{S}_b U)}{\text{tr}(U^T \hat{S}_w U) + \mu l} \quad (1.12)$$

for the RGFST, where \hat{S}_b and \hat{S}_w have the same definitions as in (1.11). This equivalent version (1.12) reduces the working dimension from N to n and plays a critical role in effectively improving the computational overhead in the second stage. The key Theorem 4.2 asserts that identically the same between-class and within-class information can be retrieved from (1.12), and therefore justifies the reduced RGFST is another two-stage LDA/QR approach [17] for handling the undersampled problem.

From the computational point of view, we only need to consider the following optimization problem

$$\max_{V^T V = I_l, V \in \mathbb{R}^{m \times l}} \psi(V) := \frac{\text{tr}(V^T B V)}{\text{tr}(V^T W V)} \quad (m \geq l), \quad (1.13)$$

where $B = B^T \in \mathbb{R}^{m \times m}$ is positive semi-definite, and $W = W^T \in \mathbb{R}^{m \times m}$ is positive definite. It is easy to see that both the GFST and the reduced RGFST (1.12) fall into this unified form by only interpreting $B = S_b$, $W = S_w$, $V = G$, and $m = N$ for the GFST (S_w is positive definite), and $B = \hat{S}_b$, $W = (\hat{S}_w + \mu I_m)$, $V = U$, and $m = n$ for (1.12), respectively.

Two iterative algorithms are available to solve the problem (1.13) thus far. Both algorithms have been proven to converge globally to global solutions. The first iteration [13] is based on the bisection technique, which is only of linear convergence and requires moreover, an initial interval containing the global optimal objective function value ψ^* of (1.13). The second iteration method summarized as Algorithm 1 is proposed in [28]. It has been proven that this iteration monotonically converges

Algorithm 1 A fast iterative scheme

Given a symmetric and positive semi-definite $B \in \mathbb{R}^{m \times m}$, and a symmetric and positive definite $W \in \mathbb{R}^{m \times m}$, this algorithm computes a global solution to (1.13).

1. Select any V_0 satisfying $V_0^T V_0 = I_l$, and the tolerance $\varepsilon > 0$. Set $k = 0$.
2. Compute an orthonormal eigenbasis V_{k+1} corresponding to the l -largest eigenvalues of

$$E_{\psi_k} = B - \psi_k W, \quad \psi_k := \psi(V_k). \quad (1.14)$$

3. If $\psi_{k+1} - \psi_k < \varepsilon$, then stop; (if $\psi_{k+1} - \psi_k = 0$, then V_{k+1} solves (1.13) globally.) otherwise, set $k = k + 1$ and go to 2.
-

to a global solution and numerical testings show that the convergence is very fast. However, no theoretical analysis on the convergence rate has been established and furthermore, no discussion on the convergence in inexact computation has been carried out either.

Our second main result offers a formal proof of the global and quadratic convergence of Algorithm 1, even under the presence of round-off errors. Two kinds of inexact computations are provided, both of which are of locally quadratic convergence. This result then completes an important and missing piece for the GFST model. Our numerical experiments on several real world data sets confirm our theoretical analysis and indicate the efficiency and effectiveness of the GFST model in classification.

1.3. Paper organization. The rest of the paper is organized as follows. In Section 2, we provide some preliminary results. In Section 3, we shall present an equivalent characterization for the global solution of (1.13). The equivalent reduced model (1.12) of the RGFST for the undersampled problem is established in Section 4, and an algorithm (Algorithm 2) based on Algorithm 1 and (1.12) is thus proposed. In Section 5, we prove the local quadratic convergence both in exact and two kinds of inexact computations of Algorithm 1. Practical implementations including the storage requirement and computational complexity are discussed in Section 6. Experimental results on several real world data sets are reported in Section 7. Finally, some concluding remarks are drawn in Section 8.

2. Preliminaries. We first define the constraint of (1.13) as

$$St(l, m) = \{V \in \mathbb{R}^{m \times l} | V^T V = I_l\}. \quad (2.1)$$

It should be noted that $St(l, m)$ is a compact smooth manifold called the *compact Stiefel manifold*, and its *tangent space* $\mathcal{T}_V St(l, m)$ at any $V \in St(l, m)$ can be expressed by (see e.g., [5, 14])

$$\mathcal{T}_V St(l, m) = \{X \in \mathbb{R}^{m \times l} | X^T V + V^T X = 0\}. \quad (2.2)$$

Viewing the manifold $St(l, m)$ as an embedded submanifold of the Euclidean space, the standard inner product (or the Frobenius inner product) for m -by- l matrices

$$\langle X, Y \rangle = \text{tr}(X^T Y), \quad \forall X, Y \in \mathcal{T}_V St(l, m) \quad (2.3)$$

is induced and referred as the induced Riemannian metric on $St(l, m)$.

Suppose a smooth function $\phi : St(l, m) \rightarrow \mathbb{R}$, is defined on $St(l, m)$, then the gradient $\text{grad}(\phi(V))$ of ϕ at $V \in St(l, m)$ is given by

$$\text{grad}(\phi(V)) = \Pi_{\mathcal{T}} \left(\frac{\partial \phi(V)}{\partial V} \right), \quad (2.4)$$

where

$$\Pi_{\mathcal{T}}(Z) = V \left(\frac{V^T Z - Z^T V}{2} \right) + (I_m - VV^T)Z \in \mathcal{T}_V St(l, m), \quad \forall Z \in \mathbb{R}^{m \times l} \quad (2.5)$$

is the orthogonal projection of $Z \in \mathbb{R}^{m \times l}$ onto the tangent space $\mathcal{T}_V St(l, m)$ at V ; furthermore, any local minimizer (or local maximizer) $V \in St(l, m)$ of ϕ on $St(l, m)$ must be a critical point (see [2, 5, 14]); in other words, the gradient at V vanishes at this point $V \in St(l, m)$, i.e., $\text{grad}(\phi(V)) = 0$.

We next define the distance between two subspaces [12].

DEFINITION 2.1. Let \mathcal{M}_1 and \mathcal{M}_2 be two subspaces of \mathbb{R}^m with the same dimension, the distance between \mathcal{M}_1 and \mathcal{M}_2 is defined by

$$\text{dist}(\mathcal{M}_1, \mathcal{M}_2) = \|\pi_{\mathcal{M}_1} - \pi_{\mathcal{M}_2}\|_2, \quad (2.6)$$

where $\pi_{\mathcal{M}_1}$ and $\pi_{\mathcal{M}_2}$ are the orthogonal projections onto \mathcal{M}_1 and \mathcal{M}_2 respectively.

Also, the *separation* between two symmetric matrices C_1 and C_2 is given by

$$\text{sep}(C_1, C_2) = \min_{\lambda \in \lambda(C_1), \nu \in \lambda(C_2)} |\lambda - \nu|, \quad (2.7)$$

where $\lambda(C_1)$ and $\lambda(C_2)$ are the *spectrums* of the matrices C_1 and C_2 respectively.

3. Characterization of the global solution. The goal of this section is to characterize the global solution of the problem (1.13). For simplicity, we denote

$$B_V := V^T B V, \quad W_V := V^T W V, \quad \text{and hence} \quad \psi(V) = \frac{\text{tr}(B_V)}{\text{tr}(W_V)}. \quad (3.1)$$

LEMMA 3.1. *The set of the critical points of the function $\psi(V) : St(l, m) \rightarrow \mathbb{R}$ (or the stationary points of (1.13)) is given by*

$$\mathcal{S} = \{V \in St(l, m) \mid (I_m - VV^T)[B - \psi(V) \cdot W]V = 0\}. \quad (3.2)$$

Proof. It is straightforward to have

$$\frac{\partial \psi(V)}{\partial V} = 2 \left[\frac{BV}{\text{tr}(W_V)} - \frac{\text{tr}(B_V)}{(\text{tr}(W_V))^2} W_V \right],$$

and by taking the advantage of $V^T \frac{\partial \psi(V)}{\partial V} - (\frac{\partial \psi(V)}{\partial V})^T V = 0$, the gradient of $\psi : St(l, m) \rightarrow \mathbb{R}$ is

$$\text{grad}(\psi(V)) = \Pi_{\mathcal{T}} \left(\frac{\partial \psi(V)}{\partial V} \right) = 2(I_m - VV^T) \left[\frac{BV}{\text{tr}(W_V)} - \frac{\text{tr}(B_V)}{(\text{tr}(W_V))^2} W_V \right],$$

where the orthogonal projection $\Pi_{\mathcal{T}}(Z)$ is defined in (2.5). Since $\text{tr}(W_V) \neq 0$, for any $V \in St(l, m)$, (3.2) follows. \square

Denote

$$E_{\psi(V)} := B - \psi(V) \cdot W, \quad (3.3)$$

then for any $V \in \mathcal{S}$, it follows that

$$E_{\psi(V)} V = V(V^T E_{\psi(V)} V) = V M_V, \quad \text{where} \quad M_V := B_V - \psi(V) \cdot W_V. \quad (3.4)$$

This implies that $E_{\psi(V)} V \in \text{span}(V)$. Therefore, for any $V \in \mathcal{S}$, V is an *orthonormal eigenbasis* for the matrix $E_{\psi(V)}$ with the corresponding *eigenblock* M_V , and thus (M_V, V) is an *orthonormal eigenpair* (see [25]) of $E_{\psi(V)}$. Moreover, it follows that

$$\text{tr}(M_V) = \sum_{i=1}^l \lambda_{J_i}(E_{\psi(V)}) = \text{tr}(B_V) - \text{tr}(W_V) = 0,$$

where $\lambda_{J_1}(E_{\psi(V)}), \dots, \lambda_{J_l}(E_{\psi(V)})$ are the eigenvalues corresponding to the orthonormal eigenbasis V , with $\lambda_i(E_{\psi(V)})$, by counting the *algebraic multiplicity* for each eigenvalue, standing for the i -th largest eigenvalue of $E_{\psi(V)}$. The discussion then leads to the following Lemma 3.2.

LEMMA 3.2. *Any stationary point V of (1.13) is an orthonormal eigenbasis for the matrix $E_{\psi(V)}$ defined by (3.3), and the sum of the eigenvalues of $E_{\psi(V)}$ corresponding to the orthonormal eigenbasis V is zero.*

As a direct application, we have the following result for the special case $l = 1$.

COROLLARY 3.3. *If $l = 1$, then V^* is a global solution of (1.13) if and only if it is a normalized eigenvector of $W^{-1}B \in \mathbb{R}^{m \times m}$ corresponding to the largest eigenvalue.*

Proof. For any $V \in \mathcal{S}$, from Lemma 3.2 and $l = 1$, it follows that $M_V = 0$ and hence $BV = \psi(V)WV$, or $W^{-1}BV = \psi(V)V$, which implies that V is an eigenvector of $W^{-1}B$ with

the corresponding eigenvalue $\psi(V)$. On the other hand, it can be verified from Lemma 3.1 that any normalized eigenvector of $W^{-1}B$ is a stationary point. (Note that all the eigenvalues of $W^{-1}B$ are real; see [12].) Therefore, any normalized eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue is a global solution of (1.13); and vice versa. \square

In general, we offer a necessary and sufficient condition for V^* to be a global solution of (1.13).

THEOREM 3.4. *Let ψ^* be the global optimal objective function value of (1.13). Then any $V^* \in \mathbb{R}^{m \times l}$ solves (1.13) globally if and only if V^* is an orthonormal eigenbasis corresponding to the l -largest eigenvalues of the matrix*

$$E^* := (B - \psi^* \cdot W) \in \mathbb{R}^{m \times m}. \quad (3.5)$$

Moreover, the sum of the l -largest eigenvalues of the matrix E^* is zero.

Proof. Let V^* be any orthonormal eigenbasis of E^* corresponding to the l -largest eigenvalues, with the associated eigenblock $M_{V^*} = (V^*)^T E^* V^*$, i.e.,

$$E^* V^* = (B - \psi^* \cdot W) V^* = V^* M_{V^*}. \quad (3.6)$$

Suppose \tilde{V} is an arbitrary global solution that solves (1.13), i.e., $\psi(\tilde{V}) = \psi^*$. Since $\tilde{V} \in \mathcal{S}$, it follows from Lemma 3.2 that

$$E_{\psi(\tilde{V})} \tilde{V} = E^* \tilde{V} = [B - \psi(\tilde{V}) \cdot W] \tilde{V} = \tilde{V} M_{\tilde{V}}, \quad \text{and} \quad \text{tr}(M_{\tilde{V}}) = 0.$$

Obviously $\text{tr}(M_{V^*}) = \text{tr}((V^*)^T E^* V^*) \geq \text{tr}((\tilde{V})^T E^* \tilde{V}) = \text{tr}(M_{\tilde{V}}) = 0$.

Premultiplying $(V^*)^T$ and taking trace operator on both sides of (3.6) yield

$$\text{tr}(B_{V^*}) - \psi^* \cdot \text{tr}(W_{V^*}) = \text{tr}(M_{V^*}) \geq 0,$$

and hence

$$\psi(V^*) = \frac{\text{tr}(B_{V^*})}{\text{tr}(W_{V^*})} \geq \psi^* + \frac{\text{tr}(M_{V^*})}{\text{tr}(W_{V^*})} \geq \psi^*,$$

where the equality holds if $\text{tr}(M_{\tilde{V}}) = \text{tr}(M_{V^*}) = 0$. Since ψ^* is the global optimal value, it consequently leads to the fact that $\psi(V^*) = \psi^*$ and $\text{tr}(M_{\tilde{V}}) = \text{tr}(M_{V^*}) = 0$, which prove the sufficient part of the theorem.

Moreover, $\tilde{V} \in \mathcal{S}$ and $\text{tr}(M_{\tilde{V}}) = \text{tr}(M_{V^*}) = 0$ imply that \tilde{V} is also an orthonormal eigenbasis corresponding to the l -largest eigenvalues of the matrix E^* , and hence we complete the proof. \square

Let V^* be any orthonormal eigenbasis corresponding to the l -largest eigenvalues of E^* . The eigenspace $\text{span}(V^*)$ is said to be a *simple eigenspace* if the corresponding eigenvalues $\lambda_1(E^*) \geq \dots \geq \lambda_l(E^*)$ are disjoint from the other eigenvalues $\lambda_{l+1}(E^*) \geq \dots \geq \lambda_m(E^*)$ of E^* (see [25]), i.e., if the following condition

$$\lambda_l(E^*) - \lambda_{l+1}(E^*) = \delta > 0 \quad (3.7)$$

holds. When $\text{span}(V^*)$ is a simple eigenspace, then it is known ([25], p.244) that the eigenspace $\text{span}(V^*)$ is uniquely determined by its eigenvalues $\lambda_1(E^*) \geq \dots \geq \lambda_l(E^*)$. In this case, the set of the global solutions to (1.13) can be completely expressed by

$$\{V^* X | X^T X = I_l, X \in \mathbb{R}^{l \times l}\} = \text{span}(V^*) \cap St(l, m).$$

However, if the eigenspace corresponding to the l -largest eigenvalues of E^* is not a simple eigenspace, the global solutions set becomes relatively complicated and does not possess the simple form as in the first case.

Theorem 3.4 states that if the optimal objective function value ψ^* of (1.13) is available, a global solution can be computed; however, in practice, we may only have an approximation, say α , of ψ^* . The following theorem provides useful information when an orthonormal eigenbasis corresponding to the l -largest eigenvalues of the matrix $E_\alpha = B - \alpha W$ is obtained, which is the theoretical foundation of the bisection-based algorithm [13].

THEOREM 3.5. *Let ψ^* be the global optimal objective function value of (1.13). For any $\alpha \in \mathbb{R}$, suppose $V \in \mathbb{R}^{m \times l}$ is any orthonormal eigenbasis corresponding to the l -largest eigenvalues of the matrix E_α ; we have (i) if $\text{tr}(V^T E_\alpha V) > 0$, then $\psi^* > \alpha$; (ii) if $\text{tr}(V^T E_\alpha V) < 0$, then $\psi^* < \alpha$; (iii) if $\text{tr}(V^T E_\alpha V) = 0$, then $\psi^* = \alpha$, and V solves (1.13) globally.*

Proof. Let E^* be defined by (3.5), and V^* be an orthonormal eigenbasis corresponding to the l -largest eigenvalues of E^* . Noting from

$$0 = \text{tr}((V^*)^T (B - \psi^* \cdot W) V^*) = \text{tr}((V^*)^T E_\alpha V^*) + (\alpha - \psi^*) \cdot \text{tr}((V^*)^T W V^*),$$

one then has

$$(\psi^* - \alpha) \cdot \text{tr}((V^*)^T W V^*) = \text{tr}((V^*)^T E_\alpha V^*) \leq \text{tr}(V^T E_\alpha V). \quad (3.8)$$

Similarly, it follows that

$$\begin{aligned} \text{tr}(V^T E_\alpha V) &= \text{tr}(V^T E^* V) + (\psi^* - \alpha) \cdot \text{tr}(V^T W V) \\ &\leq \text{tr}(M_{V^*}) + (\psi^* - \alpha) \cdot \text{tr}(V^T W V) \\ &= (\psi^* - \alpha) \cdot \text{tr}(V^T W V). \end{aligned} \quad (3.9)$$

Consequently, from (3.8) and (3.9), it yields

$$(\psi^* - \alpha) \cdot \text{tr}((V^*)^T W V^*) \leq \text{tr}(V^T E_\alpha V) \leq (\psi^* - \alpha) \cdot \text{tr}(V^T W V).$$

This together with the positive definiteness of W leads to the result. \square

It should be pointed out that the results of Theorem 3.4 and Theorem 3.5 are also explored in [13]; the arguments developed here are based on the analysis of the smooth function on the Stiefel manifold $St(l, m)$. More importantly, these arguments lead us to establish the global and quadratic convergence of Algorithm 1, which we shall discuss in Section 5.

4. An equivalent reduced RGFST for the undersampled problem. This section is dedicated to establish the equivalent reduced RGFST model (1.12) and an efficient algorithm for the undersample problem under the assumption

$$\text{rank}(A) = n, \quad (4.1)$$

which frequently holds in practice (see [22, 32]), and guarantees furthermore, the following conclusion (Proposition 3, [22]).

LEMMA 4.1. *When $\text{rank}(A) = n$ and β is positive, the matrix $S_b - \beta S_w$ exactly has $c-1$ positive, $n-c$ negative and $N-n+1$ zero eigenvalues.*

Let $Q_1 \in \mathbb{R}^{N \times n}$ be any orthonormal basis for $\text{span}(A)$; i.e.,

$$\text{span}(Q_1) = \text{span}(A), \quad Q_1^T Q_1 = I_n. \quad (4.2)$$

It is obvious that Q_1 can be computed from either a reduced QR decomposition or a reduced SVD of A , and it is true that for such Q_1 there is a nonsingular matrix $R \in \mathbb{R}^{n \times n}$ such that $A = Q_1 R$. Suppose $[Q_1, Q_2] \in \mathbb{R}^{N \times N}$ is orthogonal. According to the definition of S_t in (1.3), we have

$$S_t = H_t H_t^T = \frac{1}{n} A (I_n - \frac{1}{n} \mathbf{e} \mathbf{e}^T)^2 A^T = \frac{1}{n} Q_1 R (I_n - \frac{1}{n} \mathbf{e} \mathbf{e}^T)^2 R^T Q_1^T = Q_1 \hat{R} Q_1^T,$$

where $\hat{R} = \frac{1}{n} R (I_n - \frac{1}{n} \mathbf{e} \mathbf{e}^T)^2 R^T$. From

$$0 = Q_2^T Q_1 \hat{R} Q_1^T Q_2 = Q_2^T S_t Q_2 = Q_2^T S_w Q_2 + Q_2^T S_b Q_2,$$

and the positive semi-definiteness of S_w and S_b , it then follows that

$$S_w Q_2 = 0 \quad \text{and} \quad S_b Q_2 = 0. \quad (4.3)$$

The key result of this section, Theorem 4.2, then can be stated using the notation

$$F_{2,\mu}(G) = \frac{\text{tr}(G^T S_b G)}{\text{tr}(G^T S_w G) + \mu l}, \quad \hat{F}_{2,\mu}(U) = \frac{\text{tr}(U^T \hat{S}_b U)}{\text{tr}(U^T \hat{S}_w U) + \mu l}. \quad (4.4)$$

THEOREM 4.2. *Suppose $\text{rank}(A) = n$ and $l \leq c - 1$, then for any $\mu > 0$ and any orthonormal matrix $Q_1 \in \mathbb{R}^{N \times n}$ satisfying (4.2), it follows that*

$$\max_{U \in St(l, n)} \hat{F}_{2,\mu}(U) = \max_{G \in St(l, N)} F_{2,\mu}(G). \quad (4.5)$$

Moreover for any global solution U^* of (1.12), $Q_1 U^*$ is a global solution of (1.10); while for any global solution G^* of (1.10), $Q_1^T G^*$ is a global solution of (1.12).

Proof. The conclusion is trivial for the case $S_b = 0$. Suppose $S_b \neq 0$, and for any $\mu > 0$, denote

$$F_{2,\mu}^* = \max_{G \in St(l, N)} F_{2,\mu}(G) \quad \text{and} \quad \hat{F}_{2,\mu}^* = \max_{U \in St(l, n)} \hat{F}_{2,\mu}(U).$$

Clearly, $F_{2,\mu}^* \geq \hat{F}_{2,\mu}^*$ and $F_{2,\mu}^* > 0$. Let G^* be an arbitrary global solution to (1.10), which, by Theorem 3.4, is an orthonormal eigenbasis corresponding to the l -largest eigenvalues of the matrix

$$D_\mu^* := S_b - F_{2,\mu}^*(S_w + \mu I_N) \in \mathbb{R}^{N \times N}. \quad (4.6)$$

Moreover, there must be an orthogonal matrix, say $Q \in \mathbb{R}^{l \times l}$, such that the new orthogonal matrix $\hat{G}^* := G^* Q \in St(l, N)$ satisfies $F_{2,\mu}(\hat{G}^*) = F_{2,\mu}^*$ and

$$D_\mu^* \hat{G}^* = \hat{G}^* \cdot \text{diag}\{\lambda_1(D_\mu^*), \dots, \lambda_l(D_\mu^*)\}. \quad (4.7)$$

Now, we can decompose \hat{G}^* as

$$\hat{G}^* = Q_1 \hat{G}_1^* + Q_2 \hat{G}_2^*, \quad \text{where} \quad \hat{G}_1^* = Q_1^T \hat{G}^* \quad \text{and} \quad \hat{G}_2^* = Q_2^T \hat{G}^* := [\mathbf{g}_1, \dots, \mathbf{g}_l]. \quad (4.8)$$

From (4.6), (4.7), (4.8) and (4.3), we have

$$\begin{aligned} Q_2^T D_\mu^* \hat{G}^* &= Q_2^T \hat{G}^* \cdot \text{diag}\{\lambda_1(D_\mu^*), \dots, \lambda_l(D_\mu^*)\} \\ &= \hat{G}_2^* \cdot \text{diag}\{\lambda_1(D_\mu^*), \dots, \lambda_l(D_\mu^*)\} = -F_{2,\mu}^* \mu \hat{G}_2^*. \end{aligned}$$

Thus, if there is a $\mathbf{g}_j \neq 0$ for some $j = 1, \dots, l$, the above relation implies that $-F_{2,\mu}^* \mu = \lambda_j(D_\mu^*)$.

On the other hand, it is easy to see that

$$\lambda_j(D_\mu^*) = \lambda_j(S_b - F_{2,\mu}^* S_w) - F_{2,\mu}^* \mu, \quad j = 1, 2, \dots, l,$$

which, together with $-F_{2,\mu}^* \mu = \lambda_j(D_\mu^*)$, implies that $\lambda_j(S_b - F_{2,\mu}^* S_w) = 0$. This is a contradiction of Lemma 4.1, since $1 \leq j \leq l \leq c - 1$. Therefore, we conclude that $\hat{G}_2^* = 0$ and $\hat{G}^* = Q_1 \hat{G}_1^*$ and $Q_1^T \hat{G}^* = \hat{G}_1^* \in St(l, n)$. Note then that

$$F_{2,\mu}^* = \frac{\text{tr}((\hat{G}^*)^T S_b \hat{G}^*)}{\text{tr}((\hat{G}^*)^T S_w \hat{G}^*) + \mu l} = \frac{\text{tr}((\hat{G}_1^*)^T \hat{S}_b \hat{G}_1^*)}{\text{tr}((\hat{G}_1^*)^T \hat{S}_w \hat{G}_1^*) + \mu l} \leq \hat{F}_{2,\mu}^* \leq F_{2,\mu}^*,$$

from which we claim that $\hat{F}_{2,\mu}^* = F_{2,\mu}^*$, and both $Q_1^T \hat{G}^*$ and $Q_1^T G^*$ are global solutions of (1.12).

Furthermore, if $U^* \in \mathbb{R}^{n \times l}$ is a global solution to (1.12), by noting (4.5) and $(Q_1 U^*)^T (Q_1 U^*) = I_l$, it follows that $Q_1 U^* \in \mathbb{R}^{N \times l}$ is a global solution to (1.10). This completes the proof. \square

It deserves to note that Theorem 4.2 offers a representation of the global solution of (1.10), that is, any global solution G^* lies in the linear combination of the data points in the training set. Theorem 4.2 further leads to an efficient algorithm, Algorithm 2, for the RGFST based on (1.12).

Algorithm 2 Based on the reduced QR decomposition of A

Given a regularization parameter $\mu > 0$, and an undersampled data matrix $A \in \mathbb{R}^{N \times n}$, $N > n$, where the columns are partitioned into c classes and are linearly independent, this algorithm computes a global solution $G^* \in \mathbb{R}^{N \times l}$ for $l \leq c - 1$, of (1.10) based on its equivalent problem (1.12).

1. Compute the reduced QR decomposition of A , i.e., $A = Q_1 R$.
 2. Form $\hat{S}_b = Q_1^T H_b H_b^T Q_1 \in \mathbb{R}^{n \times n}$, $\hat{S}_w = Q_1^T H_w H_w^T Q_1 \in \mathbb{R}^{n \times n}$.
 3. Compute a global solution $U^* \in \mathbb{R}^{n \times l}$ to (1.12) by using Algorithm 1.
 4. $G^* = Q_1 U^*$.
-

5. Convergence analysis. We now turn back to our unified optimization problem (1.13) and investigate the convergence of Algorithm 1 in this section.

5.1. Global and linear convergence. The following theorem first states that Algorithm 1 is globally and at least linearly convergent.

THEOREM 5.1. *The sequence $\{\psi_k\}$ generated by Algorithm 1 is monotonically increasing to the global optimal objective function value of (1.13) ψ^* , and satisfies*

$$\psi^* - \psi_{k+1} \leq (1 - \gamma)(\psi^* - \psi_k), \quad k = 0, 1, \dots, \quad (5.1)$$

where

$$\gamma = \frac{\sum_{i=1}^l \lambda_{m-i+1}(W)}{\sum_{i=1}^l \lambda_i(W)} \in (0, 1].$$

Proof. For each $k \geq 0$ during the iteration, we have

$$\text{tr}(V_{k+1}^T E_{\psi_k} V_{k+1}) = \text{tr}(B_{V_{k+1}}) - \psi_k \cdot \text{tr}(W_{V_{k+1}}),$$

or equivalently

$$\psi_{k+1} = \frac{\text{tr}(B_{V_{k+1}})}{\text{tr}(W_{V_{k+1}})} = \psi_k + \frac{\text{tr}(V_{k+1}^T E_{\psi_k} V_{k+1})}{\text{tr}(W_{V_{k+1}})}. \quad (5.2)$$

Again, let V^* be a global solution, then by Theorem 3.4, we have

$$E^* V^* = (B - \psi^* W) V^* = V^* M_{V^*}, \quad \text{and} \quad \text{tr}(M_{V^*}) = 0.$$

Substituting $\alpha = \psi_k$ in (3.8), and noticing $\psi_k \leq \psi^*$, we have

$$\text{tr}(V_{k+1}^T E_{\psi_k} V_{k+1}) \geq \text{tr}((V^*)^T E_{\psi_k} V^*) = (\psi^* - \psi_k) \cdot \text{tr}(W_{V^*}) \geq 0. \quad (5.3)$$

Consequently, from (5.2) and (5.3), it yields that

$$\psi_{k+1} = \psi_k + \frac{\text{tr}(V_{k+1}^T E_{\psi_k} V_{k+1})}{\text{tr}(W_{V_{k+1}})} \geq \psi_k + (\psi^* - \psi_k) \frac{\text{tr}(W_{V^*})}{\text{tr}(W_{V_{k+1}})}. \quad (5.4)$$

Moreover, from (5.4) and

$$\frac{\text{tr}(W_{V^*})}{\text{tr}(W_{V_{k+1}})} \geq \gamma = \frac{\sum_{i=1}^l \lambda_{m-i+1}(W)}{\sum_{i=1}^l \lambda_i(W)} \in (0, 1],$$

we have

$$\psi_{k+1} \geq \psi_k + (\psi^* - \psi_k) \frac{\text{tr}(W_{V^*})}{\text{tr}(W_{V_{k+1}})} \geq \psi_k + (\psi^* - \psi_k) \gamma, \quad (5.5)$$

which leads to (5.1).

In (5.5), if $\psi^* > \psi_k$, then $\psi_{k+1} > \psi_k$; if $\gamma = 1$, then $\psi_{k+1} = \psi^*$; and if $\psi_{k+1} = \psi_k$, then $\psi^* = \psi_k$, which implies by Theorem 3.4 that V_k or V_{k+1} solves the problem (1.13) globally. This completes the proof. \square

5.2. Local quadratic convergence. Note that the main step (Step 2) of Algorithm 1 involves computing the l -largest eigenvalues and the corresponding orthonormal eigenvectors (this issue will be addressed in Section 6 in more details), where the rounding error cannot be avoided in practice; moreover, even though the relationship (5.1) is shown to hold in the entire iteration, it seems that the convergence would still be slow as γ could be close to zero. Fortunately, we will show in this subsection that Algorithm 1 stays out of these troubles by proving the local quadratic convergence even under the presence of round-off errors.

First, Theorem 5.1 indicates that Algorithm 1 generates a matrix sequence $\{E_{\psi_k}\}$ converging to E^* . It is clear then by Corollary 8.1.6 [12] that $\lim_{k \rightarrow +\infty} \lambda_j(E_{\psi_k}) = \lambda_j(E^*)$, for $j = 1, \dots, m$, and hence

$$\lim_{k \rightarrow +\infty} \text{tr}(V_{k+1}^T E_{\psi_k} V_{k+1}) = \lim_{k \rightarrow +\infty} \sum_{j=1}^l \lambda_j(E_{\psi_k}) = \lim_{k \rightarrow +\infty} \sum_{j=1}^l \lambda_j(E^*) = 0. \quad (5.6)$$

We next provide a key theorem for proving the quadratic convergence in both exact and inexact computations of Step 2 in Algorithm 1.

THEOREM 5.2. *Let $V \in \mathcal{S}$ be a stationary point of (1.13). For any $\epsilon_0 \in [0, 1)$, there is $K_1 = K_1(\epsilon_0) > 0$ which is nonincreasing as ϵ_0 decreases such that for any $\bar{V} \in St(l, m)$ satisfying*

$$\text{dist}(\text{span}(V), \text{span}(\bar{V})) = \epsilon \leq \epsilon_0, \quad (5.7)$$

it follows that

$$|\psi(V) - \psi(\bar{V})| \leq K_1 \epsilon^2. \quad (5.8)$$

Proof. For any $V \in \mathcal{S}$, it follows from (3.4) that $E_{\psi(V)}V = VV^T E_{\psi(V)}V$. Let the real Schur decomposition of $V^T E_{\psi(V)}V$ be $V^T E_{\psi(V)}V = Q^T \Lambda_1 Q$, and hence $E_{\psi(V)}VQ^T = VQ^T \Lambda_1$. Since $\psi(V) = \psi(VQ^T)$, $E_{\psi(V)} = E_{\psi(VQ^T)}$ and $\text{span}(V) = \text{span}(VQ^T)$, we can assume without loss of generality that the stationary point V satisfies $E_{\psi(V)}V = V\Lambda_1$.

Let $[V, V_\perp] \in \mathbb{R}^{m \times m}$ be orthogonal such that $E_{\psi(V)}V_\perp = V_\perp \Lambda_2$, where Λ_2 is diagonal. Suppose $\bar{V} = VY_1 + V_\perp Y_2$. From (5.7) and Theorem 2.6.1 [12], we have

$$\|Y_2\|_2 = \epsilon. \quad (5.9)$$

Moreover, since $\bar{V} \in St(l, m)$, it yields

$$Y_1^T Y_1 + Y_2^T Y_2 = I_l, \quad \text{or} \quad Y_1^T Y_1 = I_l - Y_2^T Y_2. \quad (5.10)$$

Since $\|Y_2^T Y_2\|_2 = \|Y_2\|_2^2 = \epsilon^2 < 1$, it is clear that $Y_1^T Y_1$ is invertible and so is Y_1 , and

$$\|Y_1^{-1}\|_2^2 = \|(Y_1^T Y_1)^{-1}\|_2 \leq \frac{1}{1 - \|Y_2^T Y_2\|_2} \leq \frac{1}{1 - \epsilon_0^2}. \quad (5.11)$$

Furthermore,

$$Y_1 Y_1^T = Y_1 (Y_1^T Y_1) Y_1^{-1} = I_l - Y_1 Y_2^T Y_2 Y_1^{-1}. \quad (5.12)$$

Notice from (5.9), (5.11), (5.12), and $\text{tr}(\Lambda_1) = 0$ (by Lemma 3.2) that

$$\begin{aligned} |\text{tr}(\bar{V}^T E_{\psi(V)} \bar{V})| &= |\text{tr}(B_{\bar{V}}) - \psi(V) \text{tr}(W_{\bar{V}})| \\ &= |\text{tr}(Y_1^T \Lambda_1 Y_1) + \text{tr}(Y_2 Y_2^T \Lambda_2)| \\ &= |\text{tr}(Y_1 Y_1^T \Lambda_1) + \text{tr}(Y_2 Y_2^T \Lambda_2)| \\ &= |\text{tr}(\Lambda_1) - \text{tr}(Y_1 Y_2^T Y_2 Y_1^{-1} \Lambda_1) + \text{tr}(Y_2 Y_2^T \Lambda_2)| \end{aligned} \quad (5.13)$$

$$\begin{aligned} &= \left| - \sum_{j=1}^l (Y_1 Y_2^T Y_2 Y_1^{-1})_{jj} (\Lambda_1)_{jj} + \sum_{j=1}^l (Y_2 Y_2^T)_{jj} (\Lambda_2)_{jj} \right| \\ &\leq \left(\frac{l \|\Lambda_1\|_2}{\sqrt{1 - \epsilon_0^2}} + l \|\Lambda_2\|_2 \right) \epsilon^2. \end{aligned} \quad (5.14)$$

Therefore,

$$|\psi(\bar{V}) - \psi(V)| \leq \frac{\frac{l \|\Lambda_1\|_2}{\sqrt{1 - \epsilon_0^2}} + l \|\Lambda_2\|_2}{\text{tr}(W_{\bar{V}})} \epsilon^2 \leq \frac{\frac{l \|\Lambda_1\|_2}{\sqrt{1 - \epsilon_0^2}} + l \|\Lambda_2\|_2}{\sum_{i=1}^l \lambda_{m+i-1}(W)} \epsilon^2 = K_1(\epsilon_0) \epsilon^2, \quad (5.15)$$

where $K_1 = K_1(\epsilon_0) = \frac{\frac{l \|\Lambda_1\|_2}{\sqrt{1 - \epsilon_0^2}} + l \|\Lambda_2\|_2}{\sum_{i=1}^l \lambda_{m+i-1}(W)}$. The proof is then completed. \square

Based on Theorem 5.2, we have another important theorem.

THEOREM 5.3. *Let V^* be any global solution of (1.13) and suppose (3.7) holds. Then there exist an $\epsilon_1 \in (0, 1)$ and a constant $K_2 > 0$ such that for any $V \in St(l, m)$ satisfying*

$$\text{dist}(\text{span}(V), \text{span}(V^*)) = \epsilon < \epsilon_1,$$

any orthonormal eigenbasis \tilde{V} of $E_{\psi(V)}$ corresponding to the l -largest eigenvalues satisfies

$$\text{dist}(\text{span}(\tilde{V}), \text{span}(V^*)) < K_2 \epsilon^2.$$

Proof. By the continuity of eigenvalues of a matrix, $\sum_{i=1}^l \lambda_i(E^*) = 0$ and (3.7), it follows that there is a $\sigma_0 > 0$ such that for any θ with $|\theta - \psi^*| < \sigma_0$, it holds that

$$\left| \sum_{i=1}^l \lambda_i(B - \theta W) \right| < \frac{\delta}{4} \quad \text{and} \quad \lambda_l(B - \theta W) - \lambda_{l+1}(B - \theta W) > \frac{\delta}{2}. \quad (5.16)$$

Let $\Delta E_{\psi(V)} = E_{\psi(V)} - E^* = (\psi(V^*) - \psi(V))W = \Delta\psi_V W$, and $[V^*, V_\perp^*] \in \mathbb{R}^{m \times m}$ be the orthogonal matrix. Denote

$$[V^*, V_\perp^*]^T E^* [V^*, V_\perp^*] = \text{diag}\{(V^*)^T E^* V^*, (V_\perp^*)^T E^* V_\perp^*\} = \text{diag}\{M_{V^*}, M_\perp^*\}$$

and

$$\begin{bmatrix} (V^*)^T \\ (V_\perp^*)^T \end{bmatrix} \Delta E_{\psi(V)} [V^*, V_\perp^*] := \begin{bmatrix} \Delta E_{\psi(V)}^{11}, & (\Delta E_{\psi(V)}^{21})^T \\ \Delta E_{\psi(V)}^{21}, & \Delta E_{\psi(V)}^{22} \end{bmatrix}.$$

Obviously, $\text{sep}(M_{V^*}, M_\perp^*) = \delta > 0$. Let $K_0 := K_1(\frac{1}{2})$ in Theorem 5.2, and let

$$\bar{\epsilon}_1 = \min\left\{\frac{1}{2}, \sqrt{\frac{\sigma_0}{2K_0}}, \sqrt{\frac{\delta}{5K_0\|W\|_2}}\right\}.$$

It then follows from Theorem 5.2 that for any $V \in St(l, m)$ satisfying

$$\text{dist}(\text{span}(V), \text{span}(V^*)) = \epsilon < \bar{\epsilon}_1,$$

we have

$$\Delta\psi_V = \psi(V^*) - \psi(V) \leq K_1(\bar{\epsilon}_1)\epsilon^2 \leq K_0\epsilon^2 \leq K_0\bar{\epsilon}_1^2 < \sigma_0, \quad (5.17)$$

which implies that (5.16) holds for the matrix $E_{\psi(V)}$; moreover,

$$\|\Delta E_{\psi(V)}\|_2 = \|\Delta\psi_V W\|_2 \leq \|W\|_2 K_1(\bar{\epsilon}_1)\epsilon^2 \leq \frac{\delta}{5} = \frac{\text{sep}(M_{V^*}, M_\perp^*)}{5},$$

which by Theorem 8.1.10 and Corollary 8.1.11 in [12] implies that there exists a matrix $P \in \mathbb{R}^{(m-l) \times l}$ with

$$\|P\|_2 \leq \frac{4\|\Delta E_{\psi(V)}^{21}\|_2}{\delta} \leq \frac{4K_0\|W\|_2}{\delta}\epsilon^2 \quad (5.18)$$

such that the matrix

$$\hat{V} := (V^* + V_\perp^* P)(I_l + P^T P)^{-\frac{1}{2}} \in \mathbb{R}^{m \times l} \quad (5.19)$$

defines an orthonormal basis of an eigenspace of $E_{\psi(V)}$, and

$$\text{dist}(\text{span}(V^*), \text{span}(\hat{V})) \leq \frac{4\|\Delta E_{\psi(V)}^{21}\|_2}{\delta} \leq \frac{4K_0\|W\|_2}{\delta}\epsilon^2 = K_2\epsilon^2, \quad (5.20)$$

with $K_2 = \frac{4K_0\|W\|_2}{\delta}$. Noting from (5.18), (5.19), $\text{tr}(M_{V^*}) = 0$, and $\Delta\psi_V \leq K_0\epsilon^2$ that

$$\begin{aligned}\text{tr}(\hat{V}^T E_{\psi(V)} \hat{V}) &= \text{tr}((V^* + V_\perp^* P)^T E_{\psi(V)} (V^* + V_\perp^* P) (I_l + P^T P)^{-1}) \\ &= \text{tr}((M_{V^*} + P^T M_\perp^* P) (I_l + P^T P)^{-1}) \\ &\quad + \text{tr}((V^* + V_\perp^* P)^T \Delta E_{\psi(V)} (V^* + V_\perp^* P) (I_l + P^T P)^{-1}) \rightarrow 0, \text{ as } \epsilon \rightarrow 0,\end{aligned}$$

there must exist an $\epsilon_1 \leq \bar{\epsilon}_1$ such that for any $V \in St(l, m)$ with

$$\text{dist}(\text{span}(V), \text{span}(V^*)) = \epsilon < \epsilon_1,$$

the matrix \hat{V} defined in (5.19) satisfies

$$|\text{tr}(\hat{V}^T E_{\psi(V)} \hat{V})| < \frac{\delta}{4}. \quad (5.21)$$

Suppose now \tilde{V} is any orthonormal eigenbasis of $E_{\psi(V)}$ corresponding to the l -largest eigenvalues, and suppose also that $\text{span}(\tilde{V}) \neq \text{span}(\hat{V})$, then by (5.17), (5.16), and (5.21), we have

$$\frac{\delta}{2} = \frac{\delta}{4} + \frac{\delta}{4} > \text{tr}(\tilde{V}^T E_{\psi(V)} \tilde{V}) - \text{tr}(\hat{V}^T E_{\psi(V)} \hat{V}) > \frac{\delta}{2},$$

which is a contradiction and leads to the conclusion of $\text{span}(\tilde{V}) = \text{span}(\hat{V})$. Therefore, by (5.20), it follows that there are $\epsilon_1 > 0$ and constant $K_2 = \frac{4K_0\|W\|_2}{\delta} > 0$ such that for any $V \in St(l, m)$ satisfying $\text{dist}(\text{span}(V), \text{span}(V^*)) = \epsilon < \epsilon_1$, the subspace, $\text{span}(\tilde{V})$, satisfies $\text{dist}(\text{span}(\tilde{V}), \text{span}(V^*)) < K_2\epsilon^2$. This completes the proof. \square

With the aid of these results, we are able to prove the local quadratic convergence of Algorithm 1 both in exact and inexact computations of Step 2 of Algorithm 1. In the following theorem, $\{\bar{V}_k\}$ represents the computed sequence in which each matrix \bar{V}_k ($k = 0, 1, \dots$) is allowed to have an error expressed by (5.23).

THEOREM 5.4. *Let V^* be any global solution of (1.13) and suppose (3.7) holds. Then there is an $\eta_0 > 0$ such that for any $\eta \in [0, \eta_0)$, there exists an $\epsilon_2 = \epsilon_2(\eta) > 0$ such that if $\bar{V}_0 \in St(l, m)$ satisfies*

$$\text{dist}(\text{span}(\bar{V}_0), \text{span}(V^*)) = \epsilon < \epsilon_2, \quad (5.22)$$

and if $\bar{V}_{k+1} \in St(l, m)$ satisfies

$$\text{dist}(\text{span}(\bar{V}_{k+1}), \text{span}(V_{k+1})) \leq \eta |\text{tr}(\bar{V}_{k+1}^T E_{\psi(\bar{V}_k)} \bar{V}_{k+1})|, \quad k = 0, 1, \dots, \quad (5.23)$$

where V_{k+1} is an orthonormal eigenbasis of $E_{\psi(\bar{V}_k)}$ corresponding to the l -largest eigenvalues, it then follows that

$$\text{dist}(\text{span}(\bar{V}_k), \text{span}(V^*)) \rightarrow 0 \quad (5.24)$$

quadratically, and moreover, $\psi(\bar{V}_k) \rightarrow \psi^*$ quadratically as $k \rightarrow +\infty$.

Proof. Let $\epsilon_1 > 0$ and $K_2 > 0$ be given in Theorem 5.3. It is obvious from Theorem 5.3 that for any $\epsilon_2 \in (0, \epsilon_1]$ and any $\bar{V}_0 \in St(l, m)$ satisfying (5.22), any orthonormal eigenbasis V_1 of $E_{\psi(\bar{V}_0)}$ corresponding to the l -largest eigenvalues satisfies

$$\text{dist}(\text{span}(V_1), \text{span}(V^*)) < K_2\epsilon^2. \quad (5.25)$$

Reduce ϵ_2 if necessary so that $K_2\epsilon_2^2 < \epsilon_1$. Thus by Theorem 5.2, we have

$$|\psi(V^*) - \psi(\bar{V}_0)| \leq K_1(\epsilon_1)\epsilon^2, \quad \text{and} \quad |\psi(V^*) - \psi(V_1)| \leq K_1(\epsilon_1)K_2^2\epsilon^4. \quad (5.26)$$

Moreover, (5.23) and (5.25) imply that

$$\begin{aligned} \text{dist}(\text{span}(\bar{V}_1), \text{span}(V^*)) &\leq \text{dist}(\text{span}(\bar{V}_1), \text{span}(V_1)) + \text{dist}(\text{span}(V_1), \text{span}(V^*)) \\ &\leq \text{dist}(\text{span}(\bar{V}_1), \text{span}(V_1)) + K_2\epsilon^2. \end{aligned} \quad (5.27)$$

We next prove that there is an $\eta_0 > 0$ such that for any $\eta \in [0, \eta_0)$, if the condition (5.23) holds, there is a constant $K_3 > 0$ which is independent of the iteration number k , such that $\text{dist}(\text{span}(\bar{V}_1), \text{span}(V_1)) < K_3\epsilon^2$, by which we can conclude together with (5.27) that

$$\text{dist}(\text{span}(\bar{V}_1), \text{span}(V^*)) < (K_2 + K_3) \cdot \text{dist}(\text{span}(\bar{V}_0), \text{span}(V^*))^2. \quad (5.28)$$

To this end, we note first of all that for any $\bar{V}_0, \bar{V}_1 \in St(l, m)$, $|\text{tr}(\bar{V}_1^T E_{\psi(\bar{V}_0)} \bar{V}_1)|$ is bounded which implies that there exists an $\eta_1 > 0$ such that for all $\eta \in [0, \eta_1)$, we have

$$\eta |\text{tr}(\bar{V}_1^T E_{\psi(\bar{V}_0)} \bar{V}_1)| < \frac{1}{2}, \quad \forall \bar{V}_0, \bar{V}_1 \in St(l, m). \quad (5.29)$$

Let $[V_1, V_{1\perp}] \in \mathbb{R}^{m \times m}$ be orthogonal, and we can assume without loss of generality that it satisfies $E_{\psi(\bar{V}_0)} V_1 = V_1 \Lambda_1$ and $E_{\psi(\bar{V}_0)} V_{1\perp} = V_{1\perp} \Lambda_2$ with Λ_1 and Λ_2 being diagonal. From (5.26) and $\text{tr}(W_{V_1}) \leq \sum_{i=1}^l \lambda_i(W) = \tau$, it yields that

$$\begin{aligned} |\text{tr}(\Lambda_1)| &= |\text{tr}(V_1^T E_{\psi(\bar{V}_0)} V_1)| = |\text{tr}(B_{V_1}) - \psi(\bar{V}_0)\text{tr}(W_{V_1})| \\ &= \text{tr}(W_{V_1}) |\psi(V_1) - \psi(\bar{V}_0)| \\ &\leq \text{tr}(W_{V_1}) (|\psi(V_1) - \psi(V^*)| + |\psi(V^*) - \psi(\bar{V}_0)|) \\ &\leq \tau K_1(\epsilon_1) (K_2^2\epsilon^2 + 1)\epsilon^2 \\ &\leq \tau K_1(\epsilon_1) (K_2^2 + 1)\epsilon^2 = K_4\epsilon^2, \end{aligned} \quad (5.30)$$

with $K_4 = \tau K_1(\epsilon_1) (K_2^2 + 1)$. Suppose $\bar{V}_1 = V_1 \bar{Y}_{11} + V_{1\perp} \bar{Y}_{21}$ and $\eta \in [0, \eta_1)$. Then (5.23), (5.29), (5.11), and (5.12) imply that $\|\bar{Y}_{21}\|_2 < \frac{1}{2}$ and $\|\bar{Y}_{11}^{-1}\|_2 < \frac{2}{\sqrt{3}}$. Moreover, following the same arguments as (5.13) and (5.14), we know that if (5.23) holds for $\eta \in [0, \eta_1)$, we have from (5.30)

$$\begin{aligned} |\text{tr}(\bar{V}_1^T E_{\psi(\bar{V}_0)} \bar{V}_1)| &= |\text{tr}(\bar{Y}_{11}^T \Lambda_1 \bar{Y}_{11}) + \text{tr}(\bar{Y}_{21}^T \Lambda_2 \bar{Y}_{21})| \\ &= |\text{tr}(\Lambda_1) - \text{tr}(\bar{Y}_{11} \bar{Y}_{21}^T \bar{Y}_{21} \bar{Y}_{11}^{-1} \Lambda_1) + \text{tr}(\bar{Y}_{21} \bar{Y}_{21}^T \Lambda_2)| \\ &\leq K_4\epsilon^2 + K_5 \|\bar{Y}_{21}\|_2^2, \end{aligned}$$

where $K_5 > 0$ can be chosen as a constant independent of the iteration number k . Thus (5.23) implies that

$$\|\bar{Y}_{21}\|_2 \leq \eta K_4\epsilon^2 + \eta K_5 \|\bar{Y}_{21}\|_2^2, \quad \text{or} \quad \|\bar{Y}_{21}\|_2 (1 - \eta K_5 \|\bar{Y}_{21}\|_2) \leq \eta K_4\epsilon^2.$$

Let $\eta_0 = \min\{\eta_1, \frac{1}{2K_5}\}$. Then for any $\eta \in [0, \eta_0)$, we have

$$\|\bar{Y}_{21}\|_2 \leq \frac{\eta K_4\epsilon^2}{(1 - \eta K_5 \|\bar{Y}_{21}\|_2)} < \frac{\eta K_4\epsilon^2}{(1 - \eta K_5)} < 2\eta K_4\epsilon^2 = K_3\epsilon^2$$

with $K_3 = 2\eta K_4$. Therefore, (5.28) holds. Reduce ϵ_2 if necessary (which does not change K_3 and K_2) to ensure that $(K_3 + K_2)\epsilon_2^2 < \epsilon_2$, by which we can use the mathematical induction to conclude for $k = 0, 1, \dots$,

$$\text{dist}(\text{span}(\bar{V}_{k+1}), \text{span}(V^*)) < (K_3 + K_2) \cdot \text{dist}(\text{span}(\bar{V}_k), \text{span}(V^*))^2. \quad (5.31)$$

This proves (5.24).

For the last part of the proof, we express $\bar{V}_k = \tilde{V}^* Y_{1k} + \tilde{V}_\perp^* Y_{2k}$, where $[\tilde{V}^*, \tilde{V}_\perp^*] \in \mathbb{R}^{m \times m}$ is orthogonal and

$$[\tilde{V}^*, \tilde{V}_\perp^*]^T E^* [\tilde{V}^*, \tilde{V}_\perp^*] = \text{diag}\{\Lambda_1^*, \Lambda_2^*\}$$

with $\Lambda_1^* = \text{diag}\{\lambda_1(E^*), \dots, \lambda_l(E^*)\}$ and $\Lambda_2^* = \text{diag}\{\lambda_{l+1}(E^*), \dots, \lambda_m(E^*)\}$. From $\text{span}(V^*) = \text{span}(\tilde{V}^*)$, (5.24) and Theorem 2.6.1 [12], we have $\|Y_{2k}\|_2 \rightarrow 0$ quadratically as $k \rightarrow +\infty$. Applying (5.15), we obtain

$$|\psi^* - \psi(\bar{V}_k)| \leq K_1(\epsilon_1) \|Y_{2k}\|_2^2, \quad (5.32)$$

which implies that $\psi(\bar{V}_k)$ converges to ψ^* at least quadratically when $k \rightarrow +\infty$. This completes the proof. \square

Directly based on Theorem 5.4, we can show that another type of inexact computation (5.34) for Step 2 of Algorithm 1 also leads to the local quadratic convergence.

COROLLARY 5.5. *Let V^* be any global solution of (1.13) and suppose (3.7) holds. Then there is an $\bar{\eta}_0 > 0$ such that for any $\eta \in [0, \bar{\eta}_0)$, there exists an $\epsilon_3 = \epsilon_3(\eta) > 0$ such that if $\bar{V}_0 \in \text{St}(l, m)$ satisfies*

$$\text{dist}(\text{span}(\bar{V}_0), \text{span}(V^*)) = \epsilon < \epsilon_3, \quad (5.33)$$

and if $\bar{V}_{k+1} \in \text{St}(l, m)$ satisfies

$$\text{dist}(\text{span}(\bar{V}_{k+1}), \text{span}(V_{k+1})) \leq \eta |\psi(\bar{V}_{k+1}) - \psi(\bar{V}_k)|, \quad k = 0, 1, \dots, \quad (5.34)$$

where V_{k+1} is an orthonormal eigenbasis of $E_{\psi(\bar{V}_k)}$ corresponding to the l -largest eigenvalues, it then follows that $\text{dist}(\text{span}(\bar{V}_k), \text{span}(V^*)) \rightarrow 0$ quadratically, and moreover, $\psi(\bar{V}_k) \rightarrow \psi^*$ quadratically as $k \rightarrow +\infty$.

Proof. The proof is simple if we note

$$\begin{aligned} |\text{tr}(\bar{V}_{k+1}^T E_{\psi(\bar{V}_k)} \bar{V}_{k+1})| &= \text{tr}(W_{\bar{V}_{k+1}}) |\psi(\bar{V}_{k+1}) - \psi(\bar{V}_k)| \\ &\geq \sum_{i=1}^l \lambda_{m+i-1}(W) \cdot |\psi(\bar{V}_{k+1}) - \psi(\bar{V}_k)|. \end{aligned}$$

Thus, from Theorem 5.4, the proof is complete after setting $\bar{\eta}_0 = \eta_0 \sum_{i=1}^l \lambda_{m+i-1}(W)$, where η_0 is defined in Theorem 5.4. \square

A few remarks may be helpful in understanding the mechanism behind the quadratic convergence results, Theorem 5.4 and Corollary 5.5. The principal tool used to bring forth the arguments is the generalization (Theorem 5.2) of the notion that Rayleigh quotients should improve twice-power as fast as the eigenvectors. Theorem 5.3 then serves as a relationship of residual between two successive iterations in exact arithmetic. To extend this relationship of residual to the entire iteration and to allow, furthermore, the round-off errors in Step 2 of Algorithm 1, further efforts were made and led to our final convergence results, Theorem 5.4 and Corollary 5.5.

6. Practical implementations. A practical implementation of Algorithm 2 should concern how to compute efficiently the l -largest eigenvalues and the corresponding orthonormal eigenvectors of a sequence of matrices, say

$$\hat{E}_k(\mu) = \hat{S}_b - \hat{F}_{2,\mu}(U_k)(\hat{S}_w + \mu I_n), \quad k = 1, 2, \dots, \quad (6.1)$$

where U_k is generated successively by Algorithm 1. A naive way is to compute the full eigensystem of $\hat{E}_k(\mu)$ which requires $\mathcal{O}(n^3)$ flops and $\mathcal{O}(n^2)$ storage. An alternative and efficient way is the *Implicitly Restarted Lanczos Method (IRLM)*, which is particularly appropriate for large scale problems with special structure, and has been successfully incorporated into the MATLAB platform (**eigs.m**); see [12, 19, 20, 25] and the references therein for a detailed discussion.

After specifying the matrix-vector product for $\hat{E}_k(\mu)\hat{\mathbf{x}}$ via

$$\hat{E}_k(\mu)\hat{\mathbf{x}} = \hat{H}_b\hat{H}_b^T\hat{\mathbf{x}} - \hat{F}_{2,\mu}(U_k)\hat{H}_w\hat{H}_w^T\hat{\mathbf{x}} - \mu\hat{F}_{2,\mu}(U_k)\hat{\mathbf{x}}, \quad \hat{\mathbf{x}} \in \mathbb{R}^n,$$

for (1.12) with $\hat{H}_b = Q_1^T H_b$, $\hat{H}_w = Q_1^T H_w$, IRLM can produce the l -largest eigenvalues and corresponding eigenvectors numerically orthogonal to working precision with $n\mathcal{O}(l) + \mathcal{O}(l^2)$ storage. Though IRLM is iterative, it requires roughly $\mathcal{O}(n^2\hat{m})$ operations in each iteration, where $l \leq \hat{m} \leq n$ is a parameter and recommended to be the same order as l .

To sum up, Algorithm 2 requires $\mathcal{O}(Nn)$ storage and has computational complexity as summarized in Table 6.1, where \hat{I} denotes the iteration number and \hat{T} denotes the maximal iteration number of IRLM for solving an orthonormal l -largest eigenbasis of $\hat{E}_k(\mu)$, $1 \leq k \leq \hat{I}$.

TABLE 6.1

Summary of orders of flops for Algorithm 2.

Step No.	Step 1	Step 2	Step 3	Step 4
Order of flops	$\mathcal{O}(Nn^2)$	$\mathcal{O}(Nnc) + \mathcal{O}(Nn^2)$	$\hat{I}\hat{T} \cdot \mathcal{O}(n^2\hat{m}), \quad l \leq \hat{m} \leq n$	$\mathcal{O}(Nn^2)$

7. Experiments. We evaluate the performance of the proposed algorithm and the reduced RGFST (1.12) on nine public data sets from face image, microarray, and text document databases. Table 7.1 describes in details the information in our testing.

TABLE 7.1

Summary of real world data sets.

	Data set	Dimension (N)	Training (n)	Test	Number of classes (c)
Image	ORL	10304	280	120	40
	Yale	16384	105	60	15
	Yale B	10304	300	150	10
Gene	Leukaemia	7129	32	40	2
	Colon	2000	22	40	2
Text	Text A2	1145	100	100	2
	Text B2	1067	150	50	2
	Text A4	1795	240	160	4
	Text A4-U	708	208	92	4

7.1. Data sets. There are three image data sets in Table 7.1. The ORL database of faces [35] contains 400 face images taken from 40 distinct subjects. Each person has 10 images taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). For Yale data set, there are totally 165 images from 15 individuals, 11 images with different facial expression or configuration (center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink) per subject. The third data set,

the Yale Face Database B [10], contains 405 viewing conditions (9 poses \times 45 illumination conditions) for 10 individuals. We use the frontal pose with 45 illumination conditions for each individual, and totally have 450 images.

Two microarray data sets are colon cancer data [1] and Leukaemia MIT AML/ ALL data [11]. The colon cancer data set contains 62 subject samples and 2000 gene expression values of each sample. Among the data, there are 40 cancer samples while the rest of them are normal samples. MIT AML/ALL data set contains 7129 genes expression of 72 samples, including 47 AML samples and 25 ALL samples.

The text data was the publicly available *20-News* groups data. The original text data was first preprocessed to strip the news messages from the e-mail headers and special tags, and eliminate the stopwords and stem words to their root forms. Then, the words were sorted on the inverse document frequency (IDF) and some words were removed if the IDF values were too small or too large. The BOW toolkit [23] was used in preprocessing. Each data set contains 2 or 4 categories. Different data sets have different class structures. Data sets A2 and A4 contain categories which are semantically different, while data set B2 contains semantically close categories. Each category was described by a subset of words. (In this case, the documents in each category are one class in the data set and the words are the dimensions.) The data set A4-U contains unbalanced documents in each category.

7.2. Experimental results. Since all our testing cases are undersampled problems, we apply Algorithm 2 to implement (1.10). We set the tolerance $\varepsilon = 10^{-6}$ and the initial point $V_0 = [I_l, 0]^T \in \mathbb{R}^{n \times l}$ for Algorithm 1 that is incorporated in Step 3 of Algorithm 2. For each data set described in Table 7.1, we randomly partition the data into the training and testing parts, and evaluate their average classification accuracies with various l , over 10 random partitions, where the classification accuracy is evaluated by employing the K -Nearest-Neighbor (KNN) procedure (see [15, 27]) with $K = 3$ in all cases.

The 5-fold cross-validation (see e.g., [8, 33]) is used to choose the optimal regularization parameter μ for both compared models in the following way: We choose $\Gamma = \{10^{-4}, 10^{-3}, \dots, 10^3, 10^4\}$ as the candidate set for μ . For each random partition of the data set in Table 7.1, the part for training is randomly divided into 5 subsets of (approximately) equal size. All subsets are mutually exclusive, and in the i -th fold, the i -th subset is held out for testing while the rest subsets are used for training. For each $\mu \in \Gamma$, we compute the cross validation accuracy, namely $Accu(\mu)$, which is defined as the mean of the accuracies for all folds. The best regularization value μ^* is then defined by

$$\mu^* = \arg \max_{\mu \in \Gamma} Accu(\mu).$$

In Table 7.2, we recorded the average classification accuracy over the 10 random partitions. Recorded are also the average optimal regularization parameter and the average iteration number that is required in Algorithm 1 for each case. It is clear from Table 7.2 that the average iteration numbers from Algorithm 1 are less than 10 for almost all the cases, which is consistent with the numerical result in [28]. Moreover, the classification accuracies indicate that for most tested cases, (1.12) can generate better classification results than (1.11).

To show more clearly the quadratic convergence of Algorithm 1 numerically, we provide the iteration history of a particular partition of the Colon data set in Fig. 7.1. The y -axis of the left figure is $\frac{|\psi_{k+1} - \psi_k|}{|\psi_k - \psi_{k-1}|^2}$; while the y -axis of the right figure is

$$\text{dist}_k = \frac{\text{dist}(\text{span}(V_{k+1}), \text{span}(V_k))}{(\text{dist}(\text{span}(V_k), \text{span}(V_{k-1})))^2}.$$

TABLE 7.2
Experimental results.

Data set	Reduced RGFST (1.12)				Model (1.11)		
	Accuracy (%)	l	μ	\hat{I}	Accuracy (%)	l	μ
ORL	85.667	5	$2.310e+2$	8.0	90.750	5	$4.100e+2$
	97.167	10	$1.510e+3$	8.1	96.000	10	$7.100e+3$
	96.333	15	$3.410e+3$	6.9	95.417	15	$6.100e+3$
	96.333	20	$2.130e+3$	7.5	95.750	20	$9.000e+3$
	97.417	25	$3.310e+3$	7.2	94.917	25	$7.200e+3$
	95.750	30	$3.300e+3$	6.6	95.500	30	$9.400e+3$
	96.000	35	$3.400e+3$	6.3	95.000	35	$9.600e+3$
Yale	97.000	40	$1.411e+3$	7.6	95.250	40	$9.100e+3$
	84.333	4	$7.100e+3$	8.0	78.667	4	$6.100e+3$
	81.333	6	$7.110e+3$	7.5	77.667	6	$6.200e+3$
	85.667	8	$7.100e+3$	7.6	80.667	8	$4.100e+3$
	83.667	10	$2.100e+3$	8.8	82.667	10	$7.100e+3$
	87.000	12	$3.310e+3$	8.1	81.333	12	$6.000e+3$
YaleB	88.333	14	$4.200e+3$	8.6	84.667	14	$5.000e+3$
	88.400	2	$3.700e+3$	9.4	88.133	2	$4.600e+3$
	96.200	8	$6.300e+3$	8.7	96.200	4	$5.410e+3$
	97.933	6	$3.501e+3$	8.8	97.467	6	$3.320e+3$
Leukaemia	98.667	8	$6.300e+3$	8.0	98.667	8	$5.400e+3$
	97.000	1	$1.000e-4$	8.6	97.000	1	$1.000e-4$
Colon	97.500	3	$1.000e-4$	5.8	95.750	3	$1.022e+1$
	79.500	1	$1.000e-4$	10.1	79.500	1	$1.000e-4$
Text A2	83.000	3	$1.000e-4$	7.4	74.750	3	$1.122e+0$
	92.400	1	$8.002e+3$	3.0	92.400	1	$8.002e+3$
Text B2	94.200	3	$4.011e+3$	3.1	93.300	3	$1.441e+3$
	88.600	1	$2.170e+1$	5.2	88.600	1	$2.170e+1$
Text A4	91.000	3	$1.720e+0$	5.0	89.800	3	$2.314e+3$
	84.813	2	$1.126e+2$	4.3	85.000	2	$1.102e+3$
	90.938	4	$2.161e+1$	4.7	90.500	4	$1.027e+2$
Text A4-U	88.625	6	$1.033e+2$	4.5	88.625	6	$3.124e+2$
	88.804	2	$2.530e+1$	3.1	88.043	2	$1.242e+2$
	90.435	4	$1.032e+3$	3.0	90.000	4	$1.031e+3$
	91.413	6	$2.103e+3$	3.2	90.000	6	$1.234e+2$

8. Conclusions. In this paper, we investigated another generalization of Fisher linear discriminant [6] GFST and its regularization form RGFST (1.10). We discussed the global solutions set of the unified optimization problem (1.13) and proved the global and quadratic convergence of Algorithm 1 even under the presence of round-off errors. For the undersampled problem, we established an equivalent reduced RGFST model (1.10) with working dimension n instead of N in the first stage; based on the reduced QR decomposition of the data matrix A , Algorithm 2 was proposed in the second stage for classification. Finally, we observed the quadratic convergence of Algorithm 1 in numerical experimental results, and the advantages of the GFST model in classification.

Acknowledgements. The authors are very grateful to Professor C. Tim Kelley and Professor Moody T. Chu for their thoughtful and constructive comments on proving the quadratic convergence of Algorithm 1. They also thank the Associate Editor and two anonymous referees for many constructive comments and suggestions.

REFERENCES

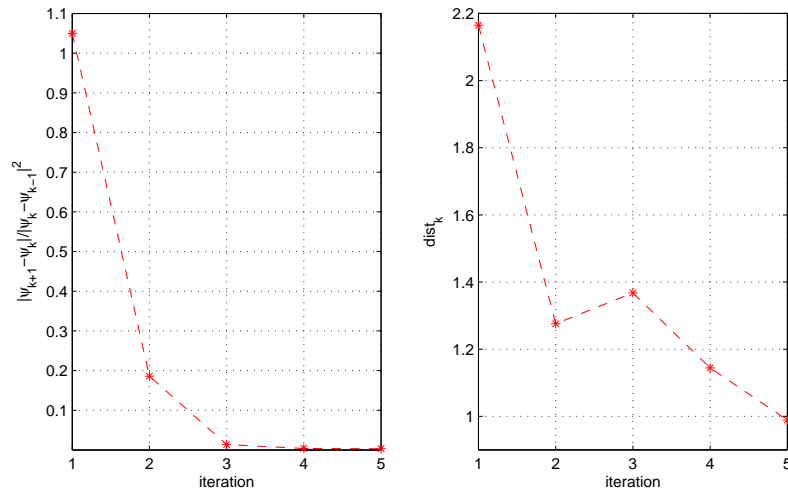


FIG. 7.1. Convergence demonstration of Algorithm 1 for the Colon data set with $l = 3$, $\mu = 10^{-4}$.

- [1] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, *Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays*, Proc. Nat. Acad. Sci. USA, vol. 96(1999), pp. 6745–6750.
- [2] M. T. Chu and N. T. Trendafilov, *The orthogonally constrained regression revisited*, J. Comput. Graph. Stat., vol. 10(2001), pp. 746–771.
- [3] M. T. Chu and J. L. Watterson, *On a multivariate eigenvalue problem: I. Algebraic Theory and Power method*, SIAM J. Sci. Comput., vol. 14(1993), pp. 1089–1106.
- [4] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Wiley-interscience, New York, 2001.
- [5] A. Edelman, T. A. Arias and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., vol. 20(1998), pp. 303–353.
- [6] R. A. Fisher, *The use of multiple measurements in taxonomic problems*, Annual of Eugenics, vol. 7(1936), pp. 179–188.
- [7] D. Foley and J. Sammon, *An optimal set of discriminant vectors*, IEEE Trans Computers, vol. 24(1975), pp. 281–289.
- [8] J. Friedman, *Regularized Discriminant Analysis*, J. Am. Statistical Assoc., vol. 84(1989), no. 405, pp. 165–175.
- [9] K. Fukunaga, *Introduction to Statistical Pattern Classification*, Academic Press, 1990.
- [10] A. S. Georghiades, P. N. Belhumeur and D. J. Kriegman, *From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23(2001), pp. 643–660.
- [11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Gaasenbeek, J. Mesirov, H. oller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield and E. Lander, *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science, vol. 286(1999), pp. 531–537.
- [12] G. H. Golub and C. F. Van Loan, *Matrix Computations, 3rd ed.*, Johns Hopkins University Press, Baltimore, MD, 1996.
- [13] Y.-F. Guo, S.-J. Li, J.-Y. Yang, T.-T. Shu and L.-D. Wu, *A Generalized Foley-Sammon Transform Based on Generalized Fisher Discriminant Criterion and Its Application to Face Recognition*, Pattern Recognition Letter, vol. 24(2003), pp. 147–158.
- [14] U. Helmke and J. B. Moore, *Optimization and Dynamical systems*, Springer-Verlag, London, UK, 1994.

- [15] P. Howland, M. Jeon, and H. Park, *Structure Preserving Dimension Reduction for Clustered Text Data Based on the Generalized Singular Value Decomposition*, SIAM J. Matrix Analysis and Applications, vol. 25(2003), no. 1, pp. 165-179.
- [16] P. Howland and H. Park, *Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26(2004), no. 8, pp. 995-1006.
- [17] P. Howland and H. Park, *Equivalence of several two-stage methods for linear discriminant analysis*, in Proceedings for the Fourth SIAM International Conference on Data Mining, Kissimmee, FL, pp. 69-77, April, 2004.
- [18] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.
- [19] R. B. Lehoucq and D. C. Sorensen, *Deflation Techniques for an Implicitly Re-Started Arnoldi Iteration*, SIAM J. Matrix Analysis and Applications, vol. 17(1996), pp. 789-821.
- [20] R. B. Lehoucq, D. C. Sorensen and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM Publications, Philadelphia, 1998.
- [21] H.-F. Li, T. Jiang and K.-S. Zhang, *Efficient and robust feature extraction by maximum margin criterion*, IEEE Trans. Neural Networks, vol. 17(2006), pp. 157-165.
- [22] J. Liu, S.-C. Chen and X.-Y. Tan, *A study on three linear discriminant analysis based methods in small sample size problem*, Pattern Recognition, vol. 41(2008), pp. 102-116.
- [23] A. K. McCallum, *A toolkit for statistical language modeling, text retrieval, classification and clustering*, 1996.
- [24] H. Park, B. L. Drake, S. Lee and C. H. Park, *Fast Linear Discriminant Analysis using QR Decomposition and Regularization*, Technical Report GT-CSE-07-21, 2007.
- [25] G. W. Stewart, *Matrix Algorithms: Vol. II, Eigensystems*, SIAM, Philadelphia, PA, 2001.
- [26] D. L. Swets and J. Weng, *Using Discriminant Eigenfeatures for Image Retrieval*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18(1996), no. 8, pp. 831-836.
- [27] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, 1999.
- [28] H. Wang, S.-C. Yan, D. Xu, X. Tang and T. Huang, *Trace Ratio vs. Ratio Trace for Dimensionality Reduction*, In Proc. International Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1-8.
- [29] S. Yan, D. Xu, B. Zhang, and H. Zhang, *Graph embedding: A general framework for dimensionality reduction*, Proceedings of Conference on Computer Vision and Pattern Recognition, 2005, pp. 830-837.
- [30] J. Yan, B.-Y. Zhang, S.-C. Yan, Z. Chen, W.-G. Fan, W.-S. Xi, Q. Yang, W.-Y. Ma and Q.-S. Cheng, *IMMC: Incremental maximum margin criterion*, in the Proceedings of the Tenth ACM SIGKDD international conference on knowledge discovery and data mining, 2004.
- [31] J.-P. Ye, *Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems*, Journal of Machine Learning Research, vol. 6(2005), pp. 483-502.
- [32] J.-P. Ye and T. Xiong, *Computational and theoretical analysis of null space and orthogonal linear discriminant analysis*, Journal of Machine Learning Research, vol. 7(2006), pp. 1183-1204.
- [33] J.-P. Ye, T. Xiong, Q. Li, R. Janardan, J.-B. Bi, V. Cherkassky and C. Kambhamettu, *Efficient model selection for regularized linear discriminant analysis*, The Fifteenth ACM International Conference on Information and Knowledge Management (CIKM 2006), pp. 532-539.
- [34] J.-P. Ye, R. Janardan, C. Park and H. Park, *An Optimization Criterion for Generalized Discriminant Analysis on Undersampled Problems*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26(2004), no. 8, pp. 982-994.
- [35] <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>