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. The undersampled problem, which arises frequently in many modern applications, involves small samples size n with high number of features N ($N > n$) and limits the application of the linear discriminant analysis. In this paper, we investigate the generalized Foley-Sammon transform (GFST, [7, 12]) and its regularization (RGFST) for undersampled problems. The optimal linear transformations of RGFST are characterized completely and an equivalent reduced RGFST is established, based on which a global and superlinear convergence algorithm is proposed. Practical implementations including computational complexity and storage of our method are discussed and experimental results on several real world data sets indicate the efficiency of the algorithm and the advantages of RGFST in classification.

Key words. Dimension reduction, linear discriminant analysis, Foley-Sammon transform, superlinear convergence

1. Introduction. A lot of practical applications of data mining, machine learning, bioinformatics require to deal with the high-dimensional data efficiently. Feature reduction commonly aims at reducing the dimension of the original features, preserving the useful and necessary information as much as possible. Linear Discriminant Analysis (LDA) ([9]) is one of the most popular approaches in this area and has been applied successfully in practice. Its goal is to find a proper linear transformation to project each sample vector with high dimension 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 $a_i \in \mathbb{R}^N$ ($i = 1, 2, \cdots, n$) 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 on this data rather difficult and inefficient. What we expect is a linear transformation, say $G \in \mathbb{R}^{N\times l}$ (generally $l \ll N$), so that it maps each sample vector $a_i \in \mathbb{R}^N$ in A to a new reduced 'sample': $y_i = G^T a_i \in \mathbb{R}^l$. A natural question arises here is how to find the optimal linear transformation G to preserve the cluster structure in A.

Suppose $A = [A_1, \cdots, A_c] \in \mathbb{R}^{N\times n}$, where $A_j \in \mathbb{R}^{N\times n_j}$ for $j = 1, \cdots, c,\ c \le n$, represents an independent class data set and n_j denotes the number of the samples of the jth class and $\sum_{j=1}^{c} n_j = n$. Define $m_j = \frac{1}{n_j} A_j e^{(j)}$ and $m = \frac{1}{n} Ae$, to be the *centroid* of cluster A_j and the *global centroid* of all objects respectively, where $e^{(j)} = (1, \cdots, 1)^T \in \mathbb{R}^{n_j}$ and $e = (1, \cdots, 1)^T \in \mathbb{R}^n$. 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

$$S_w = H_w H_w^T \in \mathbb{R}^{N\times N}, \quad H_w = \frac{1}{\sqrt{n}}[A_1 - m_1(e^{(1)})^T, \cdots, A_k - m_k(e^{(k)})^T] \in \mathbb{R}^{N\times n}, \tag{1.1}$$

$$S_b = H_b H_b^T \in \mathbb{R}^{N\times N}, \quad H_b = \frac{1}{\sqrt{n}}[\sqrt{n_1}(m_1 - m), \cdots, \sqrt{n_k}(m_k - m)] \in \mathbb{R}^{N\times c}, \tag{1.2}$$

$$S_t = H_t H_t^T \in \mathbb{R}^{N\times N}, \quad H_t = \frac{1}{\sqrt{n}}(A - me^T) \in \mathbb{R}^{N\times n}, \tag{1.3}$$

respectively, and it is easy to verify ([9]) that $S_t = S_b + S_w$. 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, $\mathrm{tr}(G^T S_w G)$ and $\mathrm{tr}(G^T S_b G)$ then measure the within-class cohesion and the between-class separation in the projected lower-dimensional space respectively. To find a proper G, minimizing the within-class cohesion and maximizing the between-class separation simultaneously in the projected space, different criteria have been

*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, P. R. China (lhzhang@math.hkbu.edu.hk).

†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, P. R. China (liliao@hkbu.edu.hk). Research was supported in part 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. Research was supported in part by RGC grants 7035/04P, 7035/05P and HKBU FRGs.

proposed and studied in the literature (e.g., [4, 7, 9, 12, 13, 15, 16, 20, 28]), including

$$F_1(G) = \mathrm{tr}((G^T S_w G)^{-1}(G^T S_b G)),\ F_2(G) = \frac{\mathrm{tr}(G^T S_b G)}{\mathrm{tr}(G^T S_w G)},\ F_3(G) = \mathrm{tr}(G^T(S_b - \beta S_w)G),\ \beta > 0.$$

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., [26, 27]) is to find an orthonormal linear transformation G^*; therefore, for criteria $F_1(G)$, $F_2(G)$ and $F_3(G)$, we have the following corresponding optimization problems

$$G^* = \arg\max_{G^TG=I_l} F_i(G), \quad i = 1, 2, 3, \tag{1.4}$$

where $I_l \in R^{l\times l}$ is the identity matrix. If S_w is nonsingular, the columns of G^* of (1.4) for the popular criterion $F_1(G)$ are the orthonormal eigenvectors of $S_w^{-1}S_b$ corresponding to its l-largest eigenvalues (see [4, 9, 16]), and the columns of G^* to (1.4) with $F_3(G)$ are the orthonormal eigenvectors of $(S_b - \beta S_w)$ corresponding to its l-largest eigenvalues ([20]). The criterion (1.4) with $F_2(G)$ which has been studied in [12], as the generalized Foley-Sammon transform [7] (GFST), is claimed to possess preferred discriminant ability in global sense, and if S_w is nonsingular it is actually a particular case of $F_3(G)$ with special β; see Section 3 and [12].

However, both $F_1(G)$ and $F_2(G)$ suffer from the singularity of S_w, which is always the case when $N > n$ (see (1.1)). In the *undersampled problem*, collecting data is expensive, and it then involves high-dimensional data with small samples, i.e., $N > n$. Such is the case for the image databases of facial recognition, gene expression datas, as well as the text documents. Various approaches (e.g., [8, 15, 16, 22, 24, 26, 28]) have been proposed to overcome this difficulty. Among them, a simple remedy is just to apply the regularization technique (see [8]) by adding a regularized term $\mu I_N,\ (\mu > 0)$, to S_w, hence arriving at

$$\max_{G^TG=I_l} \mathrm{tr}((G^T(S_w + \mu I_N)G)^{-1}G^T S_b G), \tag{1.5}$$

for $F_1(G)$, and

$$\text{RGFST}: \quad \max_{G^TG=I_l} \frac{\mathrm{tr}(G^T S_b G)}{\mathrm{tr}(G^T S_w G) + \mu l}, \tag{1.6}$$

for $F_2(G)$, where $\mu > 0$ is known as the *regularization parameter*. Most recent work in [22] establishes an equivalent reduced model of (1.5) for undersampled problems

$$\max_{\hat{G}^T\hat{G}=I_l} \mathrm{tr}((\hat{G}^T(\hat{S}_w + \mu I_n)\hat{G})^{-1}\hat{G}^T\hat{S}_b\hat{G}), \tag{1.7}$$

where $\hat{G} \in \mathbb{R}^{n\times l}$, $\hat{S}_w = Q_1^T S_w Q_1,\ \hat{S}_b = Q_1^T S_b Q_1 \in \mathbb{R}^{n\times n}$ 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 (1.7) and the generalized SVD, an efficient and direct method (LDA/QR-regGSVD) is proposed for (1.5) with orders of storage $\mathcal{O}(Nn)$ and computational complexity $\mathcal{O}(Nn^2)$.

For RGFST (1.6), an iterative algorithm [12] (stated as Algorithm 1 in Section 4) based on the bisection technique is available. However, the bisection technique is only of linear convergence; moreover, one must know an initial interval containing the optimal objective value of (1.6) in advance to start the iteration. But for the RGFST (1.6), we will see in Section 7 that either small change of the regularization parameter μ or l may tremendously change the optimal objective value, and hence it becomes difficult to determine a proper initial interval to perform the iteration successfully. These limitations imply that such iteration is not a fast and efficient algorithm for RGFST. Furthermore, the performance of RGFST (1.6) in terms of classification accuracy still remains open.

Interestingly, we will see in this paper that an equivalent reduced model of (1.6) of order n instead of N can also be established for undersampled problems; moreover, an iterative algorithm with global convergence and superlinear convergence under mild condition can be developed to solve it, which possesses the same orders of storage and computational complexity as the LDA/QR-regGSVD for (1.5). Experimental results on several real world data sets indicate that this iterative algorithm is efficient and that the equivalent reduced model of RGFST (1.6) can generate better classification results than (1.7).

The rest of the paper is organized as followed. In Section 2, we present the preliminaries that are required in characterizing the global solutions of (1.6) in Section 3. In Section 4, we propose the global and superlinear convergence iteration. The equivalent reduced model of (1.6) for undersampled problems is developed in Section 5, and an algorithm based on the superlinear convergence iteration in Section 4 is also proposed to solve the reduced RGFST. Practical implementations including the storage and computational complexity are discussed in Section 6. We present experimental results in Section 7, and draw a conclusion in Section 8.

2. Preliminaries. We first define the set of all orthonormal N-by-l matrices as

$$St(l,N) = \{G \in \mathbb{R}^{N\times l} | G^T G = I_l\}. \tag{2.1}$$

It should be noted that $St(l,N)$ is a compact smooth manifold called the *compact Stiefel manifold*, and its *tangent space* $\mathcal{T}_G St(l,N)$ at any $G \in St(l,N)$ can be expressed by $\mathcal{T}_G St(l,N) = \{X \in \mathbb{R}^{N\times l} | X^T G + G^T X = 0\}$ (see e.g., [5, 14]). Viewing the manifold $St(l,N)$ as an embedded submanifold of the Euclidean space, the standard inner product (or the Frobenius inner product) for N-by-l matrices: $< X, Y >= \mathrm{tr}(X^T Y)$, $\forall\, X, Y \in \mathcal{T}_G St(l,N)$, is induced and referred as the induced Riemannian metric on $St(l,N)$.

Suppose a smooth function $\phi : St(l,N) \to \mathbb{R}$, is defined on $St(l,N)$, then the gradient $\mathrm{grad}(\phi(G))$ of ϕ at $G \in St(l,N)$ is given by $\mathrm{grad}(\phi(G)) = \Pi_{\mathcal{T}}\left(\frac{\partial \phi(G)}{\partial G}\right)$, where

$$\Pi_{\mathcal{T}}(Z) = G\left(\frac{G^T Z - Z^T G}{2}\right) + (I_N - GG^T)Z \in \mathcal{T}_G St(l,N), \quad \forall Z \in \mathbb{R}^{N\times l} \tag{2.2}$$

is the orthogonal projection of $Z \in \mathbb{R}^{N\times l}$ onto $\mathcal{T}_G St(l,N)$ at G; furthermore, any local maximizer (or local minimizer) $G \in St(l,N)$ of ϕ on $St(l,N)$ must be a critical point (see [3, 5, 14]); i.e., $\mathrm{grad}(\phi(G)) = 0$.

3. Characterization of global solutions. In this section, we will characterize completely the global solutions of the following form of optimization problem,

$$\max_{G\in St(l,N)} \frac{\mathrm{tr}(G^T B G)}{\mathrm{tr}(G^T W G)}, \tag{3.1}$$

where $B = B^T \in \mathbb{R}^{N\times N}$ is positive semi-definite, and $W = W^T \in \mathbb{R}^{N\times N}$ is positive definite. It is clear that both RGFST (1.6) and the reduced RGFST (5.2) which will be developed in Section 5 fall into this general form (3.1). Using the notations in (3.2) and recalling the preliminaries in Section 2, we then have Lemma 3.1.

$$\psi(G) := \frac{\mathrm{tr}(B_G)}{\mathrm{tr}(W_G)}, \quad B_G := G^T B G,\ W_G := G^T W G, \quad \forall G \in St(l,N). \tag{3.2}$$

LEMMA 3.1. *The set of the critical points of the function $\psi(G) : St(l,N) \to \mathbb{R}$ (or the stationary points of (3.1)) is given by $\mathcal{S} = \{G \in St(l,N) | (I_N - GG^T)[B - \psi(G)\cdot W]G = 0\}$.*

Proof. By calculations, one has $\frac{\partial \psi(G)}{\partial G} = 2[\frac{BG}{\mathrm{tr}(W_G)} - \frac{\mathrm{tr}(B_G)}{(\mathrm{tr}(W_G))^2} WG]$ and the gradient of $\psi : St(l,N) \to \mathbb{R}$ is

$$\mathrm{grad}(\psi(G)) = \Pi_{\mathcal{T}}(\frac{\partial \psi(G)}{\partial G}) = 2(I - GG^T)[\frac{BG}{\mathrm{tr}(W_G)} - \frac{\mathrm{tr}(B_G)}{(\mathrm{tr}(W_G))^2} WG],$$

where $\Pi_{\mathcal{T}}(Z)$ is defined in (2.2). By noting that $\mathrm{tr}(W_G) \neq 0, \forall G \in St(l,N)$, the proof is complete. ∎

Denote $E_G = B - \psi(G)\cdot W$, then for any $G \in \mathcal{S}$, it follows that

$$E_G G = G(G^T E_G G) = G M_G, \quad \text{where } M_G := B_G - \psi(G)\cdot W_G. \tag{3.3}$$

This implies that $E_G G \in \mathrm{span}(G)$. Therefore, for any $G \in \mathcal{S}$, G is an *orthonormal eigenbasis* for the matrix E_G with the corresponding *eigenblock* M_G, and thus (M_G, G) is an *orthonormal eigenpair* (see [23]) of E_G. Moreover, from (3.3), it yields that $\mathrm{tr}(M_G) = \sum_{i=1}^{l} \lambda_{I_i}(E_G) = \mathrm{tr}(B_G) - \mathrm{tr}(B_G) = 0$, where $\lambda_1(E_G) \geq \lambda_2(E_G) \geq \cdots \geq \lambda_N(E_G)$ are the ordered eigenvalues of E_G, and $\lambda_{I_1}(E_G), \cdots, \lambda_{I_l}(E_G)$ are the eigenvalues associated with the orthonormal eigenbasis G. Here, we emphasize that for a symmetric N-by-N matrix, we count the *algebraic multiplicity* for each eigenvalue, and hence totally have N ordered eigenvalues. The previous discussion interestingly leads to the following Lemma 3.2.

LEMMA 3.2. *Any stationary point G of (3.1) is an orthonormal eigenbasis for the matrix $E_G = B - \psi(G)\cdot W$, and the sum of the eigenvalues of E_G corresponding to the orthonormal eigenbasis G is zero.*

In general, we provide a necessary and sufficient condition for $G^* \in \mathbb{R}^{N\times l}$ to be a global solution of (3.1).

THEOREM 3.3. *Suppose ψ^* is the global optimal value of (3.1). Then any $G^* \in \mathbb{R}^{N\times l}$ solves (3.1) globally if and only if G^* is an orthonormal eigenbasis corresponding to the l-largest eigenvalues of the matrix*

$$E^* := (B - \psi^*\cdot W) \in \mathbb{R}^{N\times N}. \tag{3.4}$$

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

Proof. Let G^* be any orthonormal eigenbasis of E^* corresponding to the l-largest eigenvalues, with the associated eigenblock M^*, i.e.,

$$E^*G^* = (B - \psi^* \cdot W)G^* = G^*M^*. \tag{3.5}$$

Suppose $\tilde{G}$ is an arbitrary global maximizer of (3.1), i.e., $\psi(\tilde{G}) = \psi^*$. Since $\tilde{G} \in \mathcal{S}$, it follows from Lemma 3.2 that there is a matrix $M_{\tilde{G}} \in \mathbb{R}^{l\times l}$ such that $E_{\tilde{G}}\tilde{G} = E^*\tilde{G} = [B - \psi(\tilde{G}) \cdot W]\tilde{G} = \tilde{G}M_{\tilde{G}}$, and $\mathrm{tr}(M_{\tilde{G}}) = 0$. Obviously $\mathrm{tr}(M^*) \geq \mathrm{tr}(M_{\tilde{G}}) = 0$.

Premultiplying $(G^*)^T$ and taking trace operator to both sides on (3.5) yield $\mathrm{tr}(B_{G^*}) - \psi^* \cdot \mathrm{tr}(W_{G^*}) = \mathrm{tr}(M^*) \geq 0$, and the equality holds if $\mathrm{tr}(M_{\tilde{G}}) = \mathrm{tr}(M^*) = 0$. Thus one has $\psi(G^*) = \frac{\mathrm{tr}(B_{G^*})}{\mathrm{tr}(W_{G^*})} \geq \psi^*$, and the equality holds if $\mathrm{tr}(M_{\tilde{G}}) = \mathrm{tr}(M^*) = 0$. Since ψ^* is the global optimal value, it consequently leads to $\psi(G^*) = \psi^*$, and $\mathrm{tr}(M_{\tilde{G}}) = \mathrm{tr}(M^*) = 0$, which prove the sufficient part of the theorem.

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

Let G^* be any orthonormal eigenbasis associated with the l-largest eigenvalues of E^*. The eigenspace $\mathrm{span}(G^*)$ is said to be a *simple eigenspace* if the following condition

$$\lambda_l(E^*) - \lambda_{l+1}(E^*) = \delta > 0 \tag{3.6}$$

holds. When $\mathrm{span}(G^*)$ is a simple eigenspace, then it is uniquely determined by its eigenvalues $\lambda_1(E^*)$, $\cdots$, $\lambda_l(E^*)$ (see [23], p.244). In this case, the global solutions set of (3.1) can be completely expressed by $\{G^*V | V^TV = I_l,\ V \in \mathbb{R}^{l\times l}\}$. However, if the eigenspace associated with the l-largest eigenvalues of E^* is not a simple eigenspace, the complete global solutions set does not possess the simple form as in the first case.

A more practical conclusion is stated in Theorem 3.4.

THEOREM 3.4. *For any $\alpha \in \mathbb{R}$, suppose (M_α, G_α) with $M_\alpha \in \mathbb{R}^{l\times l}$ and $G_\alpha \in \mathbb{R}^{N\times l}$, is an orthonormal eigenpair associated with the l-largest eigenvalues of the matrix*

$$E_\alpha := B - \alpha W. \tag{3.7}$$

Then (i) if $\mathrm{tr}(M_\alpha) > 0$, $\psi^* > \alpha$; (ii) if $\mathrm{tr}(M_\alpha) < 0$, $\psi^* < \alpha$; (iii) if $\mathrm{tr}(M_\alpha) = 0$, $\psi^* = \alpha$.

Proof. Let (M^*, G^*) be an orthonormal eigenpair associated with the l-largest eigenvalues of E^*. Since

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

it follows

$$(\psi^* - \alpha) \cdot \mathrm{tr}((G^*)^T W G^*) = \mathrm{tr}((G^*)^T E_\alpha G^*) \leq \mathrm{tr}(G_\alpha^T E_\alpha G_\alpha) = \mathrm{tr}(M_\alpha). \tag{3.8}$$

Similarly, from $\mathrm{tr}(M_\alpha) = \mathrm{tr}((G_\alpha)^T E_\alpha G_\alpha)$ and $\mathrm{tr}((G_\alpha)^T E^* G_\alpha) \leq \mathrm{tr}((G^*)^T E^* G^*) = \mathrm{tr}(\mathrm{M}^*) = 0$, one has

$$\mathrm{tr}(M_\alpha) = \mathrm{tr}((G_\alpha)^T E^* G_\alpha) + (\psi^* - \alpha) \cdot \mathrm{tr}((G_\alpha)^T W G_\alpha) \leq (\psi^* - \alpha) \cdot \mathrm{tr}((G_\alpha)^T W G_\alpha). \tag{3.9}$$

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

$$(\psi^* - \alpha) \cdot \mathrm{tr}((G^*)^T W G^*) \leq \mathrm{tr}(M_\alpha) \leq (\psi^* - \alpha) \cdot \mathrm{tr}((G_\alpha)^T W G_\alpha)$$

which, together with the positive definiteness of W, completes the proof. □

It should be pointed out that the results of Theorem 3.3 and Theorem 3.4 are also explored in [12]; the arguments developed here are based on the analysis of the smooth function on the Stiefel manifold $St(l, N)$, and the importance is that these proofs inspire us to develop efficient algorithms in next sections.

4. A fast iterative scheme. According to Theorem 3.4, an iterative algorithm, Algorithm 1, based on the bisection technique can be designed for solving (3.1), which is essentially proposed and utilized in [12].

However, as discussed in Section 1, Algorithm 1 is inefficient and impracticable for RGFST (1.6). We next propose an algorithm, Algorithm 2, which is shown to be of global convergence and superlinear convergence under mild condition (3.6). The global convergence is stated in Theorem 4.1.

THEOREM 4.1. *The sequence $\{\psi_k\}$ generated by Algorithm 2 is monotonically increasing to ψ^*, and satisfies*

$$\frac{\psi^* - \psi_{k+1}}{\psi^* - \psi_k} \leq 1 - \gamma \quad \text{with} \quad \gamma := \frac{\sum_{i=1}^{l} \lambda_{N-i+1}(W)}{\sum_{i=1}^{l} \lambda_i(W)} \in (0, 1], \qquad k = 0, 1, \cdots, \tag{4.2}$$

Algorithm 1 Based on the bisection technique

Given a symmetric and positive semi-definite $B \in \mathbb{R}^{N\times N}$, and a symmetric and positive definite $W \in \mathbb{R}^{N\times N}$, this algorithm computes a global solution to (3.1).

1. *Initial step*: Select the tolerance $\epsilon > 0$ and an initial interval with $\psi^* \in [\alpha_l, \alpha_r]$. Set $k = 0$ and $\Delta\alpha_k = \alpha_r - \alpha_l$.
2. Compute an orthonormal eigenbasis G_{k+1} corresponding to the l-largest eigenvalues of $E_\alpha = B - \frac{\alpha_l+\alpha_r}{2}W$.
3. If $\sum_{i=1}^{l}\lambda_i(E_\alpha) = 0$, then stop ($G_{k+1}$ solves (3.1) globally); if $\sum_{i=1}^{l}\lambda_i(E_\alpha) > 0$ then set $\alpha_l = \alpha$; if $\sum_{i=1}^{l}\lambda_i(E_\alpha) < 0$ then set $\alpha_r = \alpha$.
4. If $\Delta\alpha_k = \alpha_r - \alpha_l < \epsilon$, then stop; otherwise, set $k = k + 1$ and goto step 2.

Algorithm 2 A fast iterative scheme

Given a symmetric and positive semi-definite $B \in \mathbb{R}^{N\times N}$, and a symmetric and positive definite $W \in \mathbb{R}^{N\times N}$, this algorithm computes a global solution to (3.1).

1. *Initial step*: Select any $G_0 \in St(l, N)$, and the tolerance $\epsilon > 0$. Set $k = 0$.
2. Compute an orthonormal eigenbasis G_{k+1} corresponding to the l-largest eigenvalues of the matrix
$$E_k := B - \psi_k W, \quad \psi_k := \psi(G_k). \tag{4.1}$$
3. Stop if $\psi_{k+1} - \psi_k < \epsilon$; ($G_k$ solves (3.1) globally if $\psi_{k+1} - \psi_k = 0$.) otherwise, set $k = k + 1$ and goto step 2.

where ψ^* is the global optimal value of (3.1), and $\lambda_1(W) \geq \cdots \geq \lambda_N(W) > 0$ are the eigenvalues of W.

Proof. For each $k \geq 0$ during the iteration, denote $M_{k+1} \in \mathbb{R}^{l\times l}$ the eigenblock corresponding to G_{k+1} which is generated by Algorithm 2, i.e.,
$$E_k G_{k+1} = (B - \psi_k W)G_{k+1} = G_{k+1}M_{k+1}. \tag{4.3}$$
By premultiplying G_{k+1}^T and taking trace operator on both sides on (4.3), one has
$$\mathrm{tr}(B_{G_{k+1}}) - \psi_k \cdot \mathrm{tr}(W_{G_{k+1}}) = \mathrm{tr}(M_{k+1}), \quad \text{or} \quad \psi_{k+1} = \frac{\mathrm{tr}(B_{G_{k+1}})}{\mathrm{tr}(W_{G_{k+1}})} = \psi_k + \frac{\mathrm{tr}(M_{k+1})}{\mathrm{tr}(W_{G_{k+1}})}. \tag{4.4}$$
Again, let (M^*, G^*) be an orthonormal eigenpair associated with the l-largest eigenvalues of E^* (3.4). Let $\alpha = \psi_k$ in (3.7), (3.8) and $\psi_k \leq \psi^*$ then give
$$\mathrm{tr}(M_{k+1}) \geq \mathrm{tr}((G^*)^T E_k G^*) = (\psi^* - \psi_k)\cdot \mathrm{tr}(W_{G^*}) \geq 0. \tag{4.5}$$
Consequently, from (4.4) and (4.5), it yields that
$$\psi_{k+1} = \psi_k + \frac{\mathrm{tr}(M_{k+1})}{\mathrm{tr}(W_{G_{k+1}})} \geq \psi_k + (\psi^* - \psi_k)\frac{\mathrm{tr}(W_{G^*})}{\mathrm{tr}(W_{G_{k+1}})}. \tag{4.6}$$
Moreover, from $\frac{\mathrm{tr}(W_{G^*})}{\mathrm{tr}(W_{G_{k+1}})} \geq \gamma = \frac{\sum_{i=1}^{l}\lambda_{N-i+1}(W)}{\sum_{i=1}^{l}\lambda_i(W)} \in (0, 1]$ and (4.6), it then results to
$$\psi_{k+1} \geq \psi_k + (\psi^* - \psi_k)\frac{\mathrm{tr}(W_{G^*})}{\mathrm{tr}(W_{G_{k+1}})} \geq \psi_k + (\psi^* - \psi_k)\gamma, \tag{4.7}$$
which then proves (4.2). In (4.7), 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 3.3 that G_k or G_{k+1} solves (3.1) globally. We complete the proof. □

Theorem 4.1 and Corollary 8.1.6 [11] imply that $\lim_{k\to+\infty}\lambda_j(E_k) = \lambda_j(E^*),\ j = 1, \cdots, N$, and hence
$$\lim_{k\to+\infty}\mathrm{tr}(G_{k+1}^T E_k G_{k+1}) = \lim_{k\to+\infty}\sum_{j=1}^{l}\lambda_j(E_k) = \lim_{k\to+\infty}\sum_{j=1}^{l}\lambda_j(E^*) = 0, \tag{4.8}$$
where E_k and G_{k+1} are generated by Algorithm 2. We next prove that the algorithm enjoys the superlinear convergence under (3.6). Before stating the theorem, we first present some preliminaries (see e.g., [2, 11]).

DEFINITION 4.2. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two subspaces of $\mathbb{R}^N$ with the same dimension, the distance between $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined by* $\mathrm{dist}(\mathcal{M}_1, \mathcal{M}_2) = \|\pi_{\mathcal{M}_1} - \pi_{\mathcal{M}_2}\|_2$, *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 (4.9), where $\lambda(C_1)$ and $\lambda(C_2)$ are the *spectrums* of the matrices C_1 and C_2 respectively.

$$\mathrm{sep}(C_1, C_2) = \min_{\lambda\in\lambda(C_1),\ \nu\in\lambda(C_2)} |\lambda - \nu|. \tag{4.9}$$

LEMMA 4.3. *Under the assumptions of Theorem 4.1, if (3.6) holds additionally, then*

$$\lim_{k\to+\infty} \mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(G_k)) = 0, \tag{4.10}$$

where G_k is generated by Algorithm 2, and G^* is an orthonormal eigenbasis that corresponds to the l-largest eigenvalues of the matrix E^* in (3.4).

Proof. Let $\Delta E_k = E_k - E^* = (\psi^* - \psi_k)W$, and $[G^*, G_\perp^*] \in \mathbb{R}^{N\times N}$ be an orthogonal matrix. Denote

$$\begin{bmatrix}(G^*)^T\\ (G_\perp^*)^T\end{bmatrix} E^*[G^*, G_\perp^*] := \begin{bmatrix}M^*, & 0\\ 0, & M_\perp^*\end{bmatrix}; \quad \begin{bmatrix}(G^*)^T\\ (G_\perp^*)^T\end{bmatrix} \Delta E_k[G^*, G_\perp^*] := \begin{bmatrix}\Delta E_k^{11}, & (\Delta E_k^{21})^T\\ \Delta E_k^{21}, & \Delta E_k^{22}\end{bmatrix}.$$

Obviously, $\mathrm{sep}(M^*, M_\perp^*) = \delta > 0$. Suppose now $k_0 > 0$ is sufficiently large so that for all $k > k_0$ it follows that

$$\lambda_l(E_k) > \lambda_{l+1}(E_k) + \frac{\delta}{2}, \quad \text{and} \quad \|\Delta E_k\|_2 < \frac{\mathrm{sep}(M^*, M_\perp^*)}{5} = \frac{\delta}{5},$$

and by Theorem 8.1.10 and Corollary 8.1.11 in [11], it implies that there exists a matrix $P_k \in \mathbb{R}^{(N-l)\times l}$ with $\|P_k\|_2 \le \frac{4\|\Delta E_k^{21}\|_2}{\delta}$ such that the matrix $\hat{G}_k := (G^* + G_\perp^* P_k)(I_l + P_k^T P_k)^{-\frac{1}{2}} \in \mathbb{R}^{N\times l}$ defines an orthonormal basis of an eigenspace of E_k, and $\mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(\hat{G}_k)) \le \frac{4\|\Delta E_k^{21}\|_2}{\delta}$. Since $\Delta E_k \to 0$ as $k \to +\infty$, it follows

$$\lim_{k\to+\infty} P_k = 0, \quad \text{and} \quad \lim_{k\to+\infty} \mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(\hat{G}_k)) = 0. \tag{4.11}$$

Moreover, $\mathrm{tr}(\hat{G}_k^T E_k \hat{G}_k)$ is the sum of the eigenvalues of E_k associated with the eigenbasis $\hat{G}_k$, and

$$\begin{aligned}\mathrm{tr}(\hat{G}_k^T E_k \hat{G}_k) &= \mathrm{tr}((G^* + G_\perp^* P_k)^T E_k (G^* + G_\perp^* P_k)(I_l + P_k^T P_k)^{-1})\\ &= \mathrm{tr}((M^* + P_k^T M_\perp^* P_k)(I_l + P_k^T P_k)^{-1}) + \mathrm{tr}((G^* + G_\perp^* P_k)^T \Delta E_k (G^* + G_\perp^* P_k)(I_l + P_k^T P_k)^{-1}),\end{aligned}$$

which results to

$$\lim_{k\to+\infty} \mathrm{tr}(\hat{G}_k^T E_k \hat{G}_k) = \mathrm{tr}(M^*) = \lim_{k\to+\infty} \sum_{j=1}^{l} \lambda_j(E^*) = 0. \tag{4.12}$$

We next prove that there exists a $k_1 \ge k_0$ such that $\forall k > k_1$, it follows that $\mathrm{span}(G_k) = \mathrm{span}(\hat{G}_k)$. If the statement is not true, then there must exist a subsequence, say $\{\mathrm{span}(\hat{G}_{k_i})\}$, of $\{\mathrm{span}(\hat{G}_k)\}$ $(k > k_0)$, in which each subspace $\mathrm{span}(\hat{G}_{k_i})$ contains at least a one-dimensional eigenspace of E_k with the associated eigenvalue less than $\lambda_l(E_k) - \frac{\delta}{2}$. This directly results to the fact that for $i = 1, 2, \cdots$,

$$\mathrm{tr}(\hat{G}_{k_i}^T E_{k_i} \hat{G}_{k_i}) < \mathrm{tr}(G_{k_i}^T E_{k_i} G_{k_i}) - \frac{\delta}{2}, \quad \text{and} \quad \lim_{i\to+\infty} \mathrm{tr}(\hat{G}_{k_i}^T E_{k_i} \hat{G}_{k_i}) \le \sum_{j=1}^{l} \lambda_j(E^*) - \frac{\delta}{2} = -\frac{\delta}{2},$$

a contradiction with (4.12). Consequently, (4.11) and $\mathrm{span}(G_k) = \mathrm{span}(\hat{G}_k)$ $(\forall k > k_1)$, complete the proof. ∎

THEOREM 4.4. *Under the assumptions of Lemma 4.3, the sequence $\{\psi_k\}$ generated by Algorithm 2 is monotonically increasing and converges superlinearly to ψ^*, i.e.,* $\lim_{k\to+\infty} \frac{\psi^* - \psi_{k+1}}{\psi^* - \psi_k} = 0$. *Moreover, there exists a constant $\hat{\delta} \ge 0$ such that for any sufficiently large k, it follows*

$$\mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(G_k)) \le (\psi^* - \psi_k)\hat{\delta},$$

and hence $\mathrm{span}(G_k)$ *converges to* $\mathrm{span}(G^*)$ *at least superlinearly.*

Proof. From Lemma 4.3, (4.10) holds. By Theorem 1.5.2 [2], (4.10) implies that there exists a sequence of orthogonal matrices $V_k \in \mathbb{R}^{l\times l}$ such that $V_k^T V_k = I_l,\ G_k V_k \to G^*,$ as $k \to +\infty$. Therefore,

$$\mathrm{tr}(W_{G_{k+1}}) = \mathrm{tr}(G_{k+1}^T W G_{k+1}) = \mathrm{tr}((G_k V_k)^T W (G_k V_k)) \to \mathrm{tr}(W_{G^*}), \quad \text{as } k \to +\infty.$$

On the other hand, from (4.4), it follows that

$$0 \le \frac{\psi^* - \psi_{k+1}}{\psi^* - \psi_k} \le 1 - \frac{\mathrm{tr}(W_{G^*})}{\mathrm{tr}(W_{G_{k+1}})} \to 0, \quad \text{as } k \to +\infty,$$

which completes the first part of the proof. For the second part, we just need to note that in the argument of Theorem 4.3, $\|\Delta E_k^{21}\|_2 = (\psi^* - \psi_k)\|(G^*)^T W G_\perp^*\|_2$, and for any sufficiently large $k > 0$,

$$\mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(G_k)) \le \frac{4\|\Delta E_k^{21}\|_2}{\delta} = (\psi^* - \psi_k)\frac{4\|(G^*)^T W G_\perp^*\|_2}{\delta} := (\psi^* - \psi_k)\hat{\delta},$$

which implies that the sequence $\{\mathrm{dist}(\mathrm{span}(G^*), \mathrm{span}(G_k))\}$ converges to zero at least with the same order as $\{\psi^* - \psi_k\}$. Therefore, we complete the proof. □

5. An equivalent reduced RGFST for undersampled problems. According to Algorithm 2, when S_w is positive definite, we then can compute G^* for RGFST (1.6) with $\mu = 0$; and $\mu > 0$ when S_w is singular. However, such strategy is still inefficient in terms of computational complexity and storage for undersampled problems since we have to work on N-by-N matrices to compute the l-largest eigenvalues of a sequence of N-by-N matrices according to (4.1). Fortunately, we will show that either a reduced QR decomposition or a reduced SVD preprocessing of the data matrix $A \in \mathbb{R}^{N\times n}$ can, equivalently, reduce RGFST (1.6) from order N to n, and hence, significantly reduce the computational complexity and storage. In this section, we will establish the equivalent reduced RGFST under the assumption $\mathrm{rank}(A) = n$, which frequently holds with applications (see Section 7, and [20, 27]) such as face recognition.

Let $Q_1 \in \mathbb{R}^{N\times n}$ be any *orthonormal basis* for $\mathrm{span}(A)$; i.e., $\mathrm{span}(Q_1) = \mathrm{span}(A)$, and $Q_1^T Q_1 = I_n$. It is obvious that Q_1 can be computed from either the reduced QR decomposition or the 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. From the definition of S_t (1.3), we then have

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

where $\hat{R} = \frac{1}{n} R(I_n - \frac{1}{n} ee^T)^2 R^T$, and $e = (1, \cdots, 1)^T \in \mathbb{R}^n$. Note 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 (semi-)positive definiteness of $S_w,\ S_b$, it then follows that $S_w Q_2 = S_b Q_2 = 0$. Moreover, under the assumption $\mathrm{rank}(A) = n$, the following lemma holds (Proposition 3, [20]).

LEMMA 5.1. *When* $\mathrm{rank}(A) = n$ *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.*

Using the notations

$$\hat{S}_b = Q_1^T S_b Q_1,\ \hat{S}_w = Q_1^T S_w Q_1 \in \mathbb{R}^{n\times n},\ F_{2,\mu}(G) = \frac{\mathrm{tr}(G^T S_b G)}{\mathrm{tr}(G^T S_w G) + \mu l},\ \hat{F}_{2,\mu}(U) = \frac{\mathrm{tr}(U^T \hat{S}_b U)}{\mathrm{tr}(U^T \hat{S}_w U) + \mu l}, \quad (5.1)$$

we establish the reduced model (5.2) of RGFST (1.6) and have Theorem 5.2.

$$\max_{U\in St(l,n)} \hat{F}_{2,\mu}(U) \equiv \max_{U\in St(l,n)} \frac{\mathrm{tr}(U^T \hat{S}_b U)}{\mathrm{tr}(U^T \hat{S}_w U) + \mu l}. \quad (5.2)$$

THEOREM 5.2. *Suppose* $rank(A) = n$ *and* $l \le c - 1$, *then for any* $\mu > 0$ *and any orthonormal basis* $Q_1 \in \mathbb{R}^{N\times n}$ *for* $\mathrm{span}(A)$, *it follows that*

$$\max_{U\in St(l,n)} \hat{F}_{2,\mu}(U) = \max_{G\in St(l,N)} F_{2,\mu}(G). \quad (5.3)$$

Moreover for any global solution U^* *to (5.2),* $Q_1 U^*$ *is a global solution to (1.6); while for any global solution* G^* *to (1.6),* $Q_1^T G^*$ *is a global solution to (5.2).*

Proof. It is trivial for $S_b = 0$. 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 $F^*_{2,\mu} > 0$. By Theorem 3.3, any global solution G^* to (1.6) is an orthonormal eigenbasis associated with the l-largest eigenvalues of the matrix $D^*_\mu = S_b - F^*_{2,\mu}(S_w + \mu I_N) \in \mathbb{R}^{N\times N}$.

We first assume that $G^* = [g_1, \cdots, g_l] \in \mathbb{R}^{N\times l}$ is such a solution that for each i $(1 \leq i \leq l)$, $g_i \in \mathbb{R}^N$ is the orthonormal eigenvector of D^*_μ corresponding to the ith largest eigenvalue $\lambda_i(D^*_\mu)$, i.e.,

$$D^*_\mu g_i = [S_b - F^*_{2,\mu}(S_w + \mu I_N)]g_i = \lambda_i(D^*_\mu)g_i, \quad i = 1, \cdots, l. \tag{5.4}$$

Let $[Q_1, Q_2] \in \mathbb{R}^{N\times N}$ be orthogonal, then any $g_i \in \mathbb{R}^N$ can be expressed as $g_i = Q_1 g_i^{(1)} + Q_2 g_i^{(2)}$, for $i = 1, \cdots, l$. We next prove by contradiction that $g_i^{(2)} = 0, \quad \forall i = 1, \cdots, l$.

Suppose there exists $g_j^{(2)} \neq 0$ for some $1 \leq j \leq l$, then from (5.4) and $S_w Q_2 = S_b Q_2 = 0$, one has

$$Q_2^T[S_b - F^*_{2,\mu}(S_w + \mu I_N)](Q_1 g_j^{(1)} + Q_2 g_j^{(2)}) = -F^*_{2,\mu}\mu g_j^{(2)} = \lambda_j(D^*_\mu)g_j^{(2)}$$

which leads to $-F^*_{2,\mu}\mu = \lambda_j(D^*_\mu)$, and hence $\lambda_j(D^*_\mu) < 0$ since $F^*_{2,\mu} > 0$. Let $\hat{\lambda}_j$ be the jth largest eigenvalue of the matrix $S_b - F^*_{2,\mu}S_w$, and from Lemma 5.1 and $j \leq l \leq c-1$, one has $\hat{\lambda}_j > 0$. On the other hand, it is clear that $\lambda_j(D^*_\mu) = \hat{\lambda}_j - F^*_{2,\mu}\mu$, and $-F^*_{2,\mu}\mu = \lambda_j(D^*_\mu)$ implies $\lambda_j(D^*_\mu) = \hat{\lambda}_j + \lambda_j(D^*_\mu)$, which gives $\hat{\lambda}_j = 0$, a contradiction. Therefore, we claim that $g_i = Q_1 g_i^{(1)}$, for all $i = 1, \cdots, l$; or

$$G^* = Q_1[g_1^{(1)}, \cdots, g_l^{(1)}] = Q_1\bar{G}^*, \quad \bar{G}^* := [g_1^{(1)}, \cdots, g_l^{(1)}] \in \mathbb{R}^{n\times l};$$

hence $I_l = (G^*)^T G^* = (\bar{G}^*)^T Q_1^T Q_1 \bar{G}^* = (\bar{G}^*)^T \bar{G}^*$, indicating $\bar{G}^* = Q_1^T G^*$ is a feasible solution to (5.2) and

$$F^*_{2,\mu} = \frac{\mathrm{tr}((\bar{G}^*)^T Q_1^T S_b Q_1 \bar{G}^*)}{\mathrm{tr}((\bar{G}^*)^T Q_1^T S_w Q_1 \bar{G}^*) + \mu l} = \frac{\mathrm{tr}((\bar{G}^*)^T \hat{S}_b \bar{G}^*)}{\mathrm{tr}((\bar{G}^*)^T \hat{S}_w \bar{G}^*) + \mu l} \leq \hat{F}^*_{2,\mu} \leq F^*_{2,\mu},$$

which implies that $\bar{G}^* = Q_1^T G^*$ is a global solution to (5.2), and also proves (5.3).

Now, in general, we assume $G^* \in \mathbb{R}^{N\times l}$ is an arbitrary global solution to (1.6). It is clear that there exists an orthogonal matrix $Q \in \mathbb{R}^{l\times l}$ such that the columns of G^*Q are orthonormal eigenvectors of D^*_μ, and from the previous argument, it then follows that $Q_1^T G^* Q$ is a global solution to (5.2). Therefore,

$$\hat{F}^*_{2,\mu} = \frac{\mathrm{tr}(Q(G^*)^T Q_1 \hat{S}_b Q_1^T G^* Q)}{\mathrm{tr}(Q(G^*)^T Q_1 \hat{S}_w Q_1^T G^* Q) + \mu l} = \frac{\mathrm{tr}((G^*)^T Q_1 \hat{S}_b Q_1^T G^*)}{\mathrm{tr}((G^*)^T Q_1 \hat{S}_w Q_1^T G^*) + \mu l}.$$

Noting that $I_l = (Q_1^T G^* Q)^T (Q_1^T G^* Q) = Q^T (Q_1^T G^*)^T (Q_1^T G^*) Q$, or $I_l = (Q_1^T G^*)^T (Q_1^T G^*)$, we then know $Q_1^T G^*$ is a global solution to (5.2).

Furthermore, if $U^* \in \mathbb{R}^{n\times l}$ is a global solution to (5.2), by noting (5.3) and $(Q_1 U^*)^T (Q_1 U^*) = I_l$, $Q_1 U^* \in \mathbb{R}^{N\times l}$, it follows that $Q_1 U^*$ is a global solution to (1.6). This completes the proof. □

Theorem 5.2 then yields an efficient algorithm for (1.6), Algorithm 3, for undersampled problems.

Algorithm 3 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 linear independent, this algorithm computes a global solution $G^* \in \mathbb{R}^{N\times l}$ for $l \leq c-1$, of (1.6) based on its equivalent reduced problem (5.2).

1. Compute the reduced QR decomposition of A, i.e., $A = Q_1 R, \quad Q_1 \in \mathbb{R}^{N\times n}, \quad R \in \mathbb{R}^{n\times n}$.
2. Form $\hat{S}_b = Q_1^T H_b H_b^T Q_1 \in \mathbb{R}^{n\times n}, \quad \hat{S}_w = Q_1^T H_w H_w^T Q_1 \in \mathbb{R}^{n\times n}$.
3. Compute a global solution $U^* \in \mathbb{R}^{n\times l}$ to (5.2) by using Algorithm 2.
4. $G^* = Q_1 U^*$.

Obviously, step 3 in Algorithm 3 can also be realized by alternatively employing Algorithm 1, and we refer such implementation as Algorithm 3-B.

6. Practical implementations. For a practical implementation of Algorithm 3, we should concern how to implement step 2 in Algorithm 2 efficiently. It involves computing the first l-largest eigenvalues together with the associated orthonormal eigenvectors of a sequence of matrices, say

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

where U_k is generated successively by Algorithm 2. A naive way is to compute the full eigensystem of $\hat{E}_k(\mu)$ which requires then $\mathcal{O}(n^3)$ flops and $\mathcal{O}(n^2)$ storage. However, it is usually impossible for (1.6) directly in which the eigensystem computation involves $\mathcal{O}(N^3)$ 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 scaled problems with special structure, and has been successfully incorporated into the MATLAB platform (**eigs.m**). IRLM can produce the l-largest eigenvalues and associated eigenvectors numerically orthogonal to working precision with $n \cdot \mathcal{O}(l) + \mathcal{O}(l^2)$ 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. See [23, 11, 18, 19] and the references therein for a detailed discussion.

To sum up, therefore, Algorithm 3 requires $\mathcal{O}(Nn)$ storage and has computational complexity as summarized in Table 1, where $\hat{I}$ denotes the length of the sequence $\{\hat{E}_k(\mu)\}_{k=1}^{\hat{I}}$ till convergence and $\hat{T}$ the maximal iteration number of IRLM for solving the orthonormal l-largest eigenbasis of $\hat{E}_k(\mu),\ 1 \leq k \leq \hat{I}$.

Table 1. Summary of orders of flops for Algorithm 3.

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

It should be noted that the orders of computational complexity and storage of Algorithm 3 are the same to the fast LDA algorithms, LDA/QR-GSVD, LDA/QR-regGSVD, in [22].

7. Experiments. We evaluate the performances of the proposed algorithm and the RGFST (1.6) on six public data sets. Table 2 describes in detail the information in our testing.

Table 2. Summary of real world data sets.

	Data set	Dimension (N)	Training (n)	Test	Number of classes (c)	rank(A)
Image	ORL	10304	280	120	40	280
	Yale	16384	105	60	15	105
Gene	Leukaemia	7129	32	40	2	32
	Colon	2000	22	40	2	22
Text	Text A4	1795	240	160	4	240
	Text A4-U	708	208	92	4	201

7.1. Data sets. There are two image data sets in Table 2. The ORL database of faces [29] contains 400 face images taken from 40 distinct subjects with 10 images each person taken at different times, varying the lighting, facial expressions and facial details. For Yale data set, there are 165 images from 15 individuals (11 images with different facial expression or configuration per subject).

Two microarray data sets are colon cancer data [1] and Leukaemia MIT AML/ALL data [10]. 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 dataset contains 7129 genes expression of 72 samples, including 47 AML samples and 25 ALL samples.

The text data was the publicly available *20-Newsgroups* data. The original text data was first preprocessed to strip the news messages from the e-mail headers and special tags, and eliminate the stop words 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 [21] was used in preprocessing. Each category was described by a subset of words. Data set A4 contains semantically different categories; 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 3 to implement (1.6). We set the tolerance $\epsilon = 10^{-6}$ and the initial point $U_0 = [I_l, 0]^T \in \mathbb{R}^{n\times l}$ for Algorithm 2 in step 2 of Algorithm 3. To compare the classification performance, we use various regularization parameters $\mu = 10^i$ for $i = -4, -3, \cdots, 4$ both for (1.7) and (5.2), and list in Table 3 the classification accuracies with different l, in which we give the highest accuracy with specific μ for each l. The classification accuracy is computed by employing the K-Nearest-Neighbor (KNN) procedure (see [15, 25]) with $K = 3$ in all cases. To investigate the performance of Algorithm 3, we also provide the computed optimal objective value $\hat{F}^*_{2,\mu}$ and the iteration numbers $\hat{I}$ used in Algorithm 2 for each case. It is interesting to note that in almost all the tested cases, $\hat{I}$ is less than 10, verifying numerically its superlinear convergence stated in Theorem 4.4. Moreover, the classification accuracies indicate that for most cases tested, (5.2) can generate better classification results than (1.7).

Table 3 also implies that the optimal objective value $\hat{F}^*_{2,\mu}$ of the reduced RGFST (5.2) may change tremendously when either l or μ has a small change, which makes it difficult in selecting the initial interval $[\alpha_l, \alpha_r]$ for Algorithm 1 to perform Algorithm 3-B successfully. Therefore, it is impractical to implement all our testing cases (with $\mu = 10^i$ for $i = -4, \cdots, 4$ and different l in Table 3) by Algorithm 3-B. One kind

of partial implementation of it is to specify the initial lengths $\Delta\alpha_0 = \alpha_r - \alpha_l$ using the value $\hat{F}^*_{2,\mu}$ provided in Table 3 so that $\hat{F}^*_{2,\mu} \in [\alpha_l, \alpha_r]$. By choosing various initial lengths $\Delta\alpha_0$, we can see the performance of Algorithm 3-B and compare it with Algorithm 3. We present in Table 4 the numerical results based on the data set Text A4. The tolerances used in Algorithm 2 (for Algorithm 3) and Algorithm 1 (for Algorithm 3-B) are $\epsilon = 10^{-6}$ and $\epsilon = 10^{-4}$ respectively, and the specific initial intervals $\Delta\alpha_0 = 10^1$, $\Delta\alpha_0 = 10^2$ and $\Delta\alpha_0 = 10^3$ are tested for Algorithm 3-B. For the case $l = 4$ and $\mu = 10^3$, in Figure 7.1, we plot the objective value $\hat{F}_{2,\mu}(U_k)$ at each computed U_k both from Algorithm 2 (for Algorithm 3) and Algorithm 1 (for Algorithm 3-B), in which the fast convergence of Algorithm 2 is depicted.

Table 3. Experimental results.

	Reduced RGFST (5.2)					Model (1.7)		
Data set	Accuracy (%)	l	μ	$\hat{I}$	$\hat{F}^*_{2,\mu}$	Accuracy (%)	l	μ
ORL	96.667	10	10^{-4}	9	3.7251×10^9	96.667	10	10^{-4}
	97.500	20	10^4	7	3.4937×10^1	97.500	20	10^4
	97.500	30	10^4	7	2.7668×10^1	96.667	30	10^{-4}
	99.167	40	10^4	6	2.3079×10^1	97.500	40	10^{-2}
Yale	91.667	6	10^4	8	2.3886×10^1	90.000	6	10^4
	95.000	10	10^4	7	1.8439×10^1	93.333	10	10^{-4}
	95.000	14	10^4	7	1.5229×10^1	93.333	14	10^{-4}
Leukaemia	95.000	1	10^{-4}	8	1.7962×10^{12}	95.000	1	10^{-4}
	92.500	3	10^{-4}	6	9.0403×10^0	90.000	3	10^{-3}
Colon	85.000	1	10^{-4}	9	5.9909×10^{10}	85.000	1	10^{-4}
	87.500	3	10^{-4}	8	2.1531×10^1	85.000	3	10^{-3}
Text A4	79.375	2	10^{-3}	4	3.2445×10^2	77.500	2	10^0
	90.625	4	10^3	3	1.5556×10^{-3}	90.625	4	10^0
	91.250	6	10^3	3	1.0378×10^{-3}	90.000	6	10^0
Text A4-U	83.696	2	10^2	2	7.2317×10^{-3}	82.609	2	10^0
	89.130	4	10^1	3	4.4032×10^{-2}	86.957	4	10^0
	90.217	6	10^0	3	1.8179×10^{-1}	84.783	6	10^3

Table 4. Performance comparison of Alg. 3-B and Alg. 3 on Text A4.

Text A4		Algorithm 3-B			Algorithm 3	
l	μ	$\Delta\alpha_0$	$\hat{I}$	CPU(s)	$\hat{I}$	CPU(s)
2	10^{-3}	10^1	17	2.27		
2	10^{-3}	10^2	20	2.66	4	1.56
2	10^{-3}	10^3	24	3.11		
4	10^3	10^1	17	2.16		
4	10^3	10^2	20	2.49	3	0.97
4	10^3	10^3	24	2.94		
6	10^3	10^1	17	2.18		
6	10^3	10^2	20	2.57	3	1.06
6	10^3	10^3	24	3.03		

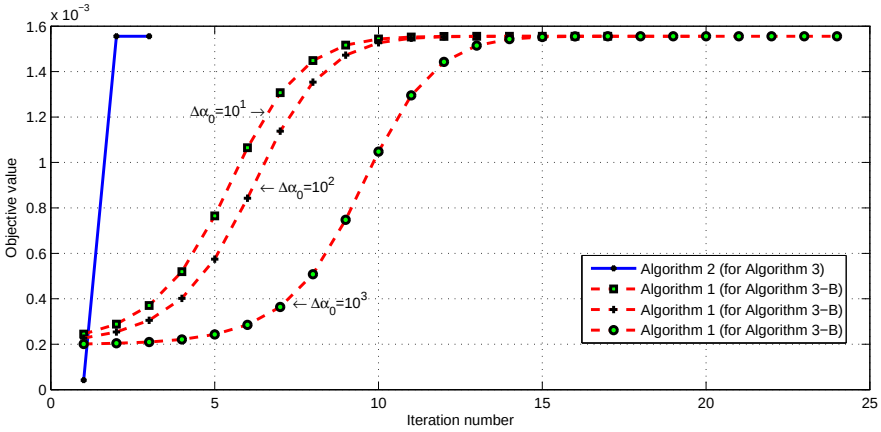

FIG. 7.1. *The value $\hat{F}_{2,\mu}(U_k)$ at each iteration of Algorithm 2 (for Algorithm 3) and Algorithm 1 (for Algorithm 3-B) for the case $l = 4$ and $\mu = 10^3$ of the data set Text A4.*

8. Conclusions. In this paper, we investigated another generalization of Fisher linear discriminant [6] GFST and its regularization form RGFST (1.6). We discussed the global solutions of RGFST and developed a global and superlinear convergence iterative scheme. For the undersampled problem, we established a reduced model of RGFST (1.6) with order n instead of N, and proved its equivalence. An algorithm (Algorithm 3) was proposed based on the reduced QR decomposition of the data matrix A and practical implementations, computational complexity and storage were discussed, which indicate that our method is very competitive to the most recently developed algorithms for the popular LDA model (1.5). Finally, we observed the superlinear convergence of our algorithm in experimental results, and the advantages of RGFST (1.6) in classification.

REFERENCES

[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, pp. 6745–6750, 1999.

[2] F. Chatelin, *Eigenvalues of Matrices,* John Wiley and Sons, 1993.

[3] M. T. Chu and N. T. Trendafilov, "The orthogonally constrained regression revisited," *J. Comput. Graph. Stat.,* vol. 10, pp. 746–771, 2001.

[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, pp. 303-353, 1998.

[6] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annual of Eugenics,* vol. 7, pp. 179–188, 1936.

[7] D. Foley and J. Sammon, "An optimal set of discriminant vectors," *IEEE Trans Computers,* vol. 24, pp. 281-289, 1975.

[8] J. Friedman, "Regularized Discriminant Analysis," *J. Am. Statistical Assoc.,* vol. 84, no. 405, pp. 165-175, 1989.

[9] K. Fukunaga, *Introduction to Statistical Pattern Classification*, Academic Press, 1990.

[10] 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, 531–537, 1999.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations, 3rd ed.*, Johns Hopkins University Press, Baltimore, MD, 1996.

[12] Y. Guo, S. Li, J. Yang, T. Shu and L. Wu, "A Generalized Foley-Sammon Transform (GFST) Based on Generalized Fisher Discriminant Criterion and Its Application to Face Recognition," *Pattern Recognition Letter,* vol. 24, pp. 147-158, 2003.

[13] Li. H, T. Jiang and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Networks,* vol. 17. no. 1, pp. 157-65, 2006.

[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, no. 1, pp. 165-179, 2003.

[16] P. Howland and H. Park, "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 8, pp. 995-1006, 2004.

[17] I. T. Jolliffe, *Principal Component Analysis,* Springer-Verlag, 1986.

[18] R. B. Lehoucq and D. C. Sorensen, "Deflation Techniques for an Implicitly Re-Started Arnoldi Iteration," *SIAM J. Matrix Analysis and Applications,* vol. 17, pp. 789-821, 1996.

[19] 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.

[20] J. Liu, S. Chen, and X. Tan, "A study on three linear discriminant analysis based methods in small sample size problem", *Pattern Recognition,* vol. 41(1), pp. 102-116, 2008.

[21] A. K. McCallum, *A toolkit for statistical language modeling, text retrieval, classification and clustering*, 1996.

[22] 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.

[23] G. W. Stewart, *Matrix Algorithms: Vol. II, Eigensystems,* SIAM, Philadelphia, PA, 2001.

[24] D. L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, 1996.

[25] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, 1999.

[26] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research,* vol. 6, pp. 483-502, 2005.

[27] J. Ye and T. Xiong, " Computaional and theoretical analysis of null space and orthogonal linear discriminant analysis," *Journal of Machine Learning Research,* vol. 7, pp. 1183–1204, 2006.

[28] J. 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, no. 8, pp. 982-994, 2004.

[29] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html