Beyond Trace Ratio: Weighted Harmonic Mean of Trace Ratios for Multiclass Discriminant Analysis

Zhihui Li, Feiping Nie, Xiaojun Chang, and Yi Yang

Abstract—Linear discriminant analysis (LDA) is one of the most important supervised linear dimensional reduction techniques which seeks to learn low-dimensional representation from the original high-dimensional feature space through a transformation matrix, while preserving the discriminative information via maximizing the between-class scatter matrix and minimizing the within class scatter matrix. However, the conventional LDA is formulated to maximize the arithmetic mean of trace ratios which suffers from the domination of the largest objectives and might deteriorate the recognition accuracy in practical applications with a large number of classes. In this paper, we propose a new criterion to maximize the weighted harmonic mean of trace ratios, which effectively avoid the domination problem while did not raise any difficulties in the formulation. An efficient algorithm is exploited to solve the proposed challenging problems with fast convergence, which might always find the globally optimal solution just using eigenvalue decomposition in each iteration. Finally, we conduct extensive experiments to illustrate the effectiveness and superiority of our method over both of synthetic datasets and real-life datasets for various tasks, including face recognition, human motion recognition and head pose recognition. The experimental results indicate that our algorithm consistently outperforms other compared methods on all of the datasets.

Index Terms—Multiclass discriminant analysis, weighted harmonic mean, trace ratio

1 Introduction

The increasing amounts of high-dimensional data in many scientific domains [1], [2], [3], [4], [5] requires dimensional reduction techniques to recover the underlying low-dimensional structures in the data [6], [7], [8], [9], [10], [11]. Some researchers have employed feature selection to select the most informative features [11], [12], [13]. The other important dimensionality reduction technique is linear discriminant analysis (LDA) which was pioneered by Fisher [14] and then extended by Rao et al. [15] to multicalss case and have been widely used in machine learning research and applications. According to Fisher criterion which maximize the total scatter versus average within-class scatter is maximized [16], LDA seeks to learn an transformation matrix from high-dimensional feature space to a low-dimensional space while preserving as much of the class discriminatory information as possible [17], [18], [19], [20]. However, since the typical Fish criterion is equivalent to maximizing the arithmetic mean of all pairwise distances, it is evidently suboptimal and inevitably leads to class separation problem when the reduced number of dimensionality is strictly lower than the class number. These critical drawbacks significantly deteriorates the recognition accuracy of FLDA based methods [21], [22], [23], [24], [25].

For this issue, many efforts have been devoted to exploit weighting scheme instead of arithmetic mean to improve the performance of FLDA [26], [27]. For example, Loog et al. [22] propose weighted pairwise Fisher criteria such that the contribution of each class pair depends on the Bayes error rate between the classes. This method inherits the inexpensive computation of traditional LDA; however, the approximate pairwise weights may not be the optimal ones because it is calculated in the original high-dimensional space. Tao et al. [25] assume that all the classes are sampled from homoscedastic Gaussians and developed three new criteria based on the geometric mean of Kullback-Leibler (KL) divergences between different pairs of classes. Bian et al. [28] replaced the geometric mean with harmonic mean and proposed to maximize the harmonic mean of all pairs of symmetric KL divergences under the homoscedastic Gaussian assumption. However, on one hand, gradient method is adopted to solve the proposed challenging problems in [25], [28], which converge very slow in some cases; on the other hand, the criterion established on the basis of maximizing the weighted sum of ratios usually suffers from the domination of the largest ratio.

Considering the fact that the conventional LDA criterion is formulated from the perspective of average-case view

- *Z. Li is with Beijing Etrol Technologies Co., Ltd, Beijing 100095, China. E-mail: zhihuilics@gmail.com.*
- *F. Nie is with the Center for OPTical Imagery Analysis and Learning, Northwestern Polytechnical University, Shaanxi 710065, China. E-mail: feipingnie@gmail.com.*
- *X. Chang is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: cxj273@gmail.com.*
- *Y. Yang is with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Sydney, NSW 2070, Australia. E-mail: yi.yang@uts.edu.au.*

Manuscript received 13 Feb. 2017; revised 6 June 2017; accepted 13 July 2017. Date of publication 18 July 2017; date of current version 8 Sept. 2017. (Corresponding author: Feiping Nie.) Recommended for acceptance by M. Wang. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2017.2728531

and ignores much information about the class separability [29], Wang et al. [30] propose to measure the class separability via maximizing the minimal trace ratio of class paries; Bian et al. [31] assume that the samples are drawn form homoscedastic Gaussians and exploited to maximize the squared minimum distance between all class pairs in low-dimensional subspace. Yang et al. proposed to use trace ratio to learn a discriminative representation from relevance feedback [32]. Zha et al. exploited discriminative information for video indexing in an unsupervised way [33]. Zhang et al. [34] incorporated a worst-case view to define a new between-class scatter measure as the minimum of the pairwise distances between class means, and a new withinclass scatter measure as the maximum of the within-class pairwise distances over all classes. Han et al. However, these challenging optimizations should be solved using Semi-definite programming (SDP) which is very time consuming and fail to handle data set in large scale.

Additionally, Wang et al. [35] argued the ratio trace approximation for trace ratio optimization which is naturally used in conventional LDA and pointed out the approximated solution might lead to uncertainty for the subsequent classification and clustering. It has been investigated theoretically in [36] that a global optimum solution can be achieved for trace ratio problem according to eigenvalue perturbation theory. Extensive empirical results also show the superiority of trace ratio based LDA [37], [38].

In this paper, we extend the original trace ratio framework for LDA and leverage weighted harmonic mean to maximize the multiple objectives of pairwise classes. This method effectively avoids the domination problem of the largest objective via focusing more on the worst objectives and guarantees all of the objectives can not be too small. Thus, it can be more appropriate to handle the situations where a large number of classes is available in practice. Thanks to the advantages of trace ratio form, we convert the maximization of weighted harmonic mean for trace ratios into a same form of minimization problem, which raise no difficulty for the procedure of optimization. We summarize the main contributions of this work in three folds:

1) To reduce dimensionality of data with a large number of classes, we propose a new criterion to maximize the weighted harmonic mean of trace ratios, which effectively avoid the domination problem of the largest objectives while did not raise any difficulties in the formulation.
2) An efficient algorithm is exploited to solve the proposed challenging problems, which might always find the globally optimal solution. Different from the SDP-based algorithm, our algorithm updates the transformation matrix just using eigenvalue decomposition, and thus converges fast with low time complexity.
3) To evaluate the performance of our model, we conduct extensive experiments on both of synthetic datasets and real-life datasets with respect to various tasks, including face recognition, human motion recognition and head pose recognition. The experimental results indicate that our algorithm consistently outperforms other compared methods on all of the datasets.

The rest of the paper is organized as follows: Section 2 revisits the trace ratio criterion in conventional LDA and reformulate the conventional objective function as its equivalent form to maximize the sum over all class pair's between-class distances and minimize the sum over all class pair's within-class distances. In Section 3, we regard to the domination problem of largest objective in dimensionality reduction with a large number of classes and exploit a novel criterion to extend the traditional LDA on the basis of weighted harmonic mean of trace ratio instead of arithmetic mean. An efficient algorithm is developed to solve the proposed challenging problems with fast convergence. Section 4 focus on introduction of the related work and the discussion of significant difference between our method and the previous ones. In Section 5, extensive experiments are conducted over synthetic datasets and real-life datasets to illustrate the effectiveness and superiority of our method. Conclusion is given in Section 6.

2 Trace Ratio Criterion Revisited

Suppose we are given n training data points $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, each data point x_i belongs to one of the classes $\{l_1, l_2, \ldots, l_c\}$. In the traditional Linear Discriminant Analysis (LDA), the within-class scatter matrix S_w and the between-class scatter matrix S_b are defined as follows:

$$S_w = \sum_{k=1}^{c} \sum_{x_i \in l_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T \tag{1}$$

$$S_b = \sum_{k=1}^{c} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T, \tag{2}$$

where n_k is the number of data points belong to class l_k, $\bar{x}_k = \frac{1}{n_k}\sum_{x_i \in l_k} x_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Under a projection matrix $W \in \mathbb{R}^{d\times m}(m < d)$, the data points are projected onto a lower dimensional subspace, and the d-dimensional data point x_i is projected to be a m-dimensional data point $W^T x_i$. In the lower dimensional subspace, we have the following results:

$$Tr(W^T S_w W) = \sum_{k=1}^{c} \sum_{x_i \in l_k} \left\| W^T (x_i - \bar{x}_k) \right\|_2^2 \tag{3}$$

$$Tr(W^T S_b W) = \sum_{k=1}^{c} n_k \left\| W^T (\bar{x}_k - \bar{x}) \right\|_2^2. \tag{4}$$

From Equations (3) and (4) we can see, under the projection matrix W, the $Tr(W^T S_w W)$ measures the Euclidean distances within the same class, and $Tr(W^T S_b W)$ measures the Euclidean distances between different classes. To maximize the discriminative power under the projection matrix W, it is preferable to minimize $Tr(W^T S_w W)$ and maximize $Tr(W^T S_b W)$ simultaneously. A natural choice is to use the ratio of $Tr(W^T S_b W)$ and $Tr(W^T S_w W)$ as the objective function. Since W is a projection matrix, a natural constraint of W is the orthogonal constraint such that the obtained W is an orthogonal projection matrix. The orthogonal constrained trace ratio problem is

$$\max_{W^T W = I} \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)}. \tag{5}$$

This trace ratio problem has been well studied recently, and the global optimal solution can be efficiently obtained by an iterative algorithm with quadratic convergence. Extensively empirical results show the trace ratio objective outperforms the traditional ratio trace objective.

For the binary class problem, according to the definition of Equation (1), the within-class scatter matrix S_w^{jk} for the jth and kth class is defined as follows:

$$S_w^{jk} = \sum_{h\in\{j,k\}} \sum_{x_i\in l_h} (x_i - \bar{x}_h)(x_i - \bar{x}_h)^T. \tag{6}$$

According to the definition of Equation (2), the between-class scatter matrix S_b^{jk} for the jth and kth class is defined as follows:

$$S_b^{jk} = \sum_{h\in\{j,k\}} n_h(\bar{x}_h - \bar{x}_{jk})(\bar{x}_h - \bar{x}_{jk})^T, \tag{7}$$

where $\bar{x}_{jk} = \frac{1}{n_k+n_k}(\sum_{x_i\in l_j} x_i + \sum_{x_i\in l_k} x_i)$ is the mean of all the data from class j and k. So the orthogonal constrained trace ratio problem in the binary class case is

$$\max_{W^TW=I} \frac{Tr(W^T S_b^{jk} W)}{Tr(W^T S_w^{jk} W)}. \tag{8}$$

According to the definitions of S_w and S_w^{jk}, it can be easily verified that

$$S_w = \frac{1}{c-1}\sum_{k=1}^{c-1}\sum_{j=k+1}^{c} S_w^{jk}. \tag{9}$$

To analyze the relationship between S_b and S_b^{jk}, we need the following lemma:

Lemma 1. *For $k = 1, 2, \ldots, c$, suppose the weight $p_k \geq 0$ and $\sum_{k=1}^{c} p_k = 1$, then we have*

$$\begin{aligned}&\sum_{k=1}^{c} p_k \left(x_k - \sum_{j=1}^{c} p_j x_j\right)\left(x_k - \sum_{j=1}^{c} p_j x_j\right)^T \\ &\quad = \sum_{k=1}^{c-1}\sum_{j=k+1}^{c} p_j p_k (x_j - x_k)(x_j - x_k)^T.\end{aligned} \tag{10}$$

Proof. For the left hand of Equation (10), we have

$$\begin{aligned}&\sum_{k=1}^{c} p_k \left(x_k - \sum_{j=1}^{c} p_j x_j\right)\left(x_k - \sum_{j=1}^{c} p_j x_j\right)^T \\ &\quad = \sum_{k=1}^{c} p_k x_k x_k^T - \sum_{k=1}^{c} p_k x_k \sum_{j=1}^{c} p_j x_j^T \\ &\quad - \sum_{k=1}^{c} p_k \sum_{j=1}^{c} p_j x_j x_k^T + \sum_{k=1}^{c} p_k \sum_{j=1}^{c} p_j x_j \sum_{i=1}^{c} p_i x_i^T.\end{aligned} \tag{11}$$

With simple mathematical deduction, we arrive at

$$\begin{aligned}&\sum_{k=1}^{c} p_k \left(x_k - \sum_{j=1}^{c} p_j x_j\right)\left(x_k - \sum_{j=1}^{c} p_j x_j\right)^T \\ &\quad = \sum_{k=1}^{c} p_k x_k x_k^T - \sum_{k=1}^{c}\sum_{j=1}^{c} p_k p_j x_k x_j^T.\end{aligned} \tag{12}$$

For the right hand of Equation (10), we have

$$\begin{aligned}&\sum_{k=1}^{c-1}\sum_{j=k+1}^{c} p_j p_k (x_j - x_k)(x_j - x_k)^T \\ &\quad = \frac{1}{2}\sum_{k=1}^{c}\sum_{j=1}^{c} p_j p_k (x_j - x_k)(x_j - x_k)^T \\ &\quad = \frac{1}{2}\sum_{k=1}^{c}\sum_{j=1}^{c} p_j p_k x_j x_j^T - \frac{1}{2}\sum_{k=1}^{c}\sum_{j=1}^{c} p_j p_k x_j x_k^T \\ &\quad - \frac{1}{2}\sum_{k=1}^{c}\sum_{j=1}^{c} p_j p_k x_k x_j^T + \frac{1}{2}\sum_{k=1}^{c}\sum_{j=1}^{c} p_j p_k x_k x_k^T \\ &\quad = \sum_{k=1}^{c} p_k \sum_{j=1}^{c} p_j x_j x_j^T - \sum_{k=1}^{c}\sum_{j=1}^{c} p_k p_j x_j x_k^T \\ &\quad = \sum_{j=1}^{c} p_j x_j x_j^T - \sum_{k=1}^{c}\sum_{j=1}^{c} p_k p_j x_j x_k^T.\end{aligned} \tag{13}$$

According to Equations (11) and (13), we get Equation (10). □

The following lemma reveals the relationship between S_b and S_b^{jk}, which are defined in Equations (2) and (7) respectively.

Lemma 2. *The between-class scatter matrix S_b can be rewritten as*

$$S_b = \frac{1}{n}\sum_{k=1}^{c-1}\sum_{j=k+1}^{c} (n_j + n_k) S_b^{jk}. \tag{14}$$

Proof. According to Equation (2) and Lemma 1, for the left hand of Equation (14), we have

$$\begin{aligned}S_b &= n\sum_{k=1}^{c}\frac{n_k}{n}(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T \\ &= n\sum_{k=1}^{c-1}\sum_{j=k+1}^{c}\frac{n_j n_k}{n^2}(\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T.\end{aligned} \tag{15}$$

According to Equation (7) and Lemma 1, we have

$$\begin{aligned}S_b^{jk} &= (n_j + n_k)\sum_{h\in\{j,k\}}\frac{n_h}{n_j + n_k}(\bar{x}_h - \bar{x}_{jk})(\bar{x}_h - \bar{x}_{jk})^T \\ &= (n_j + n_k)\frac{n_j n_k}{(n_j + n_k)^2}(\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T.\end{aligned} \tag{16}$$

So for the right hand of Equation (14), we have

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{c-1}\sum_{j=k+1}^{c}(n_j + n_k) S_b^{jk} \\ &\quad = \frac{1}{n}\sum_{k=1}^{c-1}\sum_{j=k+1}^{c} n_j n_k (\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T.\end{aligned} \tag{17}$$

According to Equations (15) and (17), we get Equation (14). □

According to Equations (14) and (9), the orthogonal constrained trace ratio problem in Equation (5) can be rewritten as

TABLE 1
A Simple Example

	J_1	J_2	J_3	J_4	$\sum_i J_i$	$\sum_i \frac{1}{J_i}$
Case 1	100	50	10	0.1	160.1	10.13
Case 2	50	40	30	30	150	0.1117

$$\max_{W^TW=I} \frac{\sum_{k=1}^{c-1}\sum_{j=k+1}^{c}(n_j+n_k)Tr(W^TS_b^{jk}W)}{\sum_{k=1}^{c-1}\sum_{j=k+1}^{c}Tr(W^TS_w^{jk}W)}. \quad (18)$$

3 Weighted Harmonic Mean of Trace Ratios

From Equation (18) we can see, traditional trace ratio problem maximizes the sum of between-class distances of all binary classes, and minimizes the sum of within-class distances of all binary classes at the same time. Therefore, traditional trace ratio problem does not explicitly maximize the ratios of between-class distances and within-class distances for all the binary classes. It is possible there are binary classes that totally overlapped. This case could happen especially when the number of classes is very large.

To focus the separability of every binary classes, one would solve the following problem to explicitly maximize the weighted sum of ratios of between-class and within-class distances for all the binary classes

$$\max_{W^TW=I} \sum_{k=1}^{c-1}\sum_{j=k+1}^{c}(n_j+n_k)\frac{Tr(W^TS_b^{jk}W)}{Tr(W^TS_w^{jk}W)}. \quad (19)$$

However, we will see from the analysis of the next section that maximizing the weighted sum of ratios is still problematic. The largest ratios could dominate the sum of ratios, and thus other ratios are ignored and would be very small although the sum of ratios is maximized.

3.1 Arithmetic Mean versus Harmonic Mean

Suppose we need to maximize multiple objectives $J_1, J_2, J_3, \ldots$, a simplest method is to maximize the weighted sum (i.e., the weighted arithmetic mean) of the objectives. The problem is to solve

$$\max_x \sum_i p_i J_i(x), \quad (20)$$

where p_i is the weight for the objective J_i. However, in Equation (20), the largest objective J_i could dominate the sum of the objectives, which might make some other objectives very small. Let's take a simple example to see it. As shown in Table 1, suppose there are four objectives J_1, J_2, J_3, J_4 to be maximized with the same weight. Consider the following two cases: in the first case, the objective values J_1, J_2, J_3, J_4 are $100, 50, 1, 0.1$, respectively, and in the second case, the objective values J_1, J_2, J_3, J_4 are $50, 40, 30, 30$, respectively. Obviously, the second case is better than the first case since there are two very small objectives in the first case. However, as can be seen in Table 1, the sum of the objectives in the first case is larger than that of in the second case. If we solve problem Equation (20), the first case will be selected other than the second case, which is not a preferred solution.

Therefore, maximizing the arithmetic mean is not a good choice for maximizing the multiple objectives. To solve this issue, a natural method is to optimize the worst case, i.e., maximize the minimal of the objectives. The problem is to solve

$$\max_x \min_i J_i(x). \quad (21)$$

This problem can be equivalently rewritten as

$$\max_{x,t\le J_i(x)} t. \quad (22)$$

Usually, the problem (Equation (22)) is difficult to solve and is very time consuming.

In this paper, in order to maximize multiple objectives, we propose to maximize the weighted harmonic mean of the objectives. According to the definition of weighted harmonic mean, we solve the following problem:

$$\max_x \frac{1}{\sum_i p_i \frac{1}{J_i(x)}}. \quad (23)$$

The problem (23) is equivalent to

$$\min_x \sum_i p_i \frac{1}{J_i(x)}. \quad (24)$$

We can see that in the problem (Equation (24)), we minimize the weighted sum of the objectives' reciprocals. If an objective is too small, then reciprocal of the objective will be very large. Therefore, the problem (Equation (24)) focus more on the worst objectives, which is similar to the problem (Equation (22)), but is usually more easier to solve.

From the above analysis we can see, maximizing the arithmetic mean focus more on the best objectives, which will make that some objectives could be too small. In contrast, maximizing the harmonic mean (i.e., minimizing the sum of reciprocals) focus more on the worst objectives, which guarantee that all the objectives can not be too small. Therefore, maximizing the harmonic mean, i.e., solving (Equation (24)), is a good choice for maximizing multiple objectives.

3.2 Weighted Harmonic Mean of Trace Ratios

Motivated by the above analysis, instead of maximizing the weighted arithmetic mean of the trace ratios of all the binary classes as in Equation (19), we propose to maximize the harmonic mean of the trace ratios. According to Equation (24), we propose to solve the following problem for discriminant analysis:

$$\min_{W^TW=I} \sum_{k=1}^{c-1}\sum_{j=k+1}^{c}(n_j+n_k)\frac{Tr(W^TS_w^{jk}W)}{Tr(W^TS_b^{jk}W)}. \quad (25)$$

It is worth noting that, thanks to the trace ratio form in the objective of class pairs, the problem Equation (25) has the same form as in the problem Equation (19). Therefore, compared with maximizing the arithmetic mean of the trace ratios, maximizing the harmonic mean of the trace ratios does not raise the difficulty. Solving problem Equation (25) and solving problem Equation (19) would have similar algorithms.

For notation simplicity, we need to solve the following problem in order to solve Equation (25):

$$\min_{W^TW=I} \sum_{i=1}^{n} \frac{Tr(W^TA_iW)}{Tr(W^TB_iW)}. \tag{26}$$

The Lagrangian function of Equation (26) is

$$f(W,\Lambda) = \sum_{i=1}^{n} \frac{Tr(W^TA_iW)}{Tr(W^TB_iW)} - Tr(\Lambda(W^TW-I)). \tag{27}$$

Note that

$$\begin{aligned} &\frac{\partial}{\partial W}\sum_{i=1}^{n} \frac{Tr(W^TA_iW)}{Tr(W^TB_iW)} \\ &\quad = \sum_{i=1}^{n} \frac{1}{tr(W^TB_iW)}\left(A_i - \frac{tr(W^TA_iW)}{tr(W^TB_iW)}B_i\right). \end{aligned} \tag{28}$$

By taking the derivative of Equation (27) w.r.t. W, we have

$$\sum_{i=1}^{n} \frac{1}{tr(W^TB_iW)}\left(A_i - \frac{tr(W^TA_iW)}{tr(W^TB_iW)}B_i\right)W = W\Lambda, \tag{29}$$

which can be rewritten as

$$MW = W\Lambda, \tag{30}$$

where the matrix M is

$$M = \sum_{i=1}^{n} \frac{1}{tr(W^TB_iW)}\left(A_i - \frac{tr(W^TA_iW)}{tr(W^TB_iW)}B_i\right). \tag{31}$$

Note that the M in Equation (31) also depends on W, which is unknown. Therefore, we propose an iterative algorithm to find the optimal solution W. First, we guess a solution W, based on which we can calculate M by Equation (31). Then we can update W according to Equation (30), and iteratively update M and W until the algorithm converges. The algorithm is summarized in Algorithm 1.

Algorithm 1. Algorithm to Solve the Problem Equation (26)

Input: $A_i|_{i=1}^n$ and $B_i|_{i=1}^n$, m.
1 Initialize $W \in \mathbb{R}^{d\times m}$ such that $W^TW = I$;
2 **while** *not converge* **do**
3 1. Calculate M according to Equation (31);
4 2. Update W, which is formed by the m eigenvectors of M corresponding to the m smallest eigenvalues;
Output: $W \in \mathbb{R}^{d\times m}$.

From Algorithm 1 we can see, in Equation (26), when $n = 1$, then the algorithm is reduced to the algorithm proposed in [39] for the traditional trace ratio problem.

We have the following theorem for Algorithm 1.

Theorem 1. *The converged solution of Algorithm 1 is a local optimal solution to the problem Equation (26).*

Proof. According to steps 1 and 2 in Algorithm 1, the converged solution will satisfy Equation (29), which is the KKT condition [40] of the problem Equation (26). Therefore, the converged solution of Algorithm 1 is a local optimal solution to the problem Equation (26). □

In the experiments, we find that Algorithm 1 always converges, and it is more interesting to see that, the Algorithm 1 with different initial solutions W always converges to the same solution, which indicates the Algorithm 1 might always find the globally optimal solution to the problem (Equation (26)).

Based on the Algorithm 1, we summarize the discriminant analysis method with weighted harmonic mean of trace ratios in Algorithm 2.

Algorithm 2. Weighted Harmonic Mean of Trace Ratios for Discriminant Analysis

Input: The training data $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}, \ldots, m$.
1 1. Calculate S_w^{jk} according to Equation (6) ;
2 2. Calculate S_b^{jk} according to Equation (7);
3 3. Find the optimal solution W to the problem (25) with Algorithm 1;
Output: $W \in \mathbb{R}^{d\times m}$.

4 Related Work

Traditional LDA is to solve the following problem

$$\min_{W} Tr(W^TS_wW)^{-1}(W^TS_bW), \tag{32}$$

where S_w and S_b are defined by Equations (3) and (4), respectively.

It is known that in the binary-class problem, if both classes are sampled from homoscedastic Gaussians, i.e., Gaussians with an identical covariance, traditional LDA criterion is the Bayes optimal criterion. However, when the data are sampled from heteroscedastic Gaussians, or there are more than two classes, traditional LDA criterion is suboptimal.

Denote

$$S_{jk} = (\bar{x}_j - \bar{x}_k)(\bar{x}_j - \bar{x}_k)^T. \tag{33}$$

In [41], it shows that $S_b = \sum_{k=1}^{c-1}\sum_{j=k+1}^{c} \frac{n_jn_k}{n^2} S_{jk}$, thus the LDA problem can be rewritten as

$$\min_{W} \sum_{k=1}^{c-1}\sum_{j=k+1}^{c} \frac{n_jn_k}{n^2} Tr(W^TS_wW)^{-1}(W^TS_{jk}W). \tag{34}$$

To focus more on the closer pairwise classes, [41] propose a weighted LDA method to solve the following weighted problem:

$$\min_{W} \sum_{k=1}^{c-1}\sum_{j=k+1}^{c} \frac{n_jn_k}{n^2} a(\Delta_{jk}) Tr(W^TS_wW)^{-1}(W^TS_{jk}W), \tag{35}$$

where a is a weighting function depends on $\Delta_{jk} = \sqrt{(\bar{x}_j - \bar{x}_k)^T S_w^{-1}(\bar{x}_j - \bar{x}_k)}$. This method is efficient since the optimal solution to the problem (Equation (35)) can be calculated by eigen-decomposition as in traditional LDA. However, the weights in Equation (35) is simply calculated according to the distances in the original space, but not calculated according to the distances in the optimal subspace

W. Therefore, the calculated weights might not the optimal weights, especially when the data distribution in the optimal subspace changes largely from the original space.

As shown in [42], the KL divergence between the densities of class j and k in the subspace W can be written as

$$D_W(l_j\|l_k) = \frac{1}{2}\big(\ln|W^TS_w^kW| - \ln|W^TS_w^jW| + Tr((W^TS_w^kW)^{-1}W^T(S_w^j + S_{jk})W)\big). \tag{36}$$

When all the classes are sampled from homoscedastic Gaussians, [42] proved the KL divergence in Equation (36) can be written as

$$D_W(l_j\|l_k) = \frac{1}{2}Tr((W^TS_wW)^{-1}W^TS_{jk}W) + constant. \tag{37}$$

Then the LDA problem in Equation (32) or (34) can be rewritten as

$$\min_W \sum_{k=1}^{c-1}\sum_{j=k+1}^{c}\frac{n_jn_k}{n^2}D_W(l_j\|l_k). \tag{38}$$

Therefore, under the assumption that all the classes are sampled from homoscedastic Gaussians, traditional LDA can be seen as maximizing the weighted arithmetic mean of the KL divergences between all pairs of classes [42]. To focus more on the closer pairs of classes, [41] proposed to maximize the weighted geometric mean of the KL divergences between all pairs of classes, which is to solve the following problem:

$$\min_W \prod_{k=1}^{c-1}\prod_{j=k+1}^{c}\big(D_W(l_j\|l_k)\big)^{\frac{n_jn_k}{n^2}}. \tag{39}$$

Since this problem is difficult to solve, [41] uses gradient method to solve it, which converges very slow in some cases.

Further, it is shown that under the homoscedastic Gaussian assumption, the symmetric KL divergence between class j and k in the subspace W can be written as

$$SD_W(l_j\|l_k) = Tr((W^TS_wW)^{-1}W^TS_{jk}W). \tag{40}$$

So under the homoscedastic Gaussian assumption, traditional LDA can be also seen as maximizing the weighted arithmetic mean of the symmetric KL divergences between all pairs of classes. Based on this motivation, it is straightforward proposed to maximize the weighted harmonic mean of the symmetric KL divergences between all pairs of classes, which is to solve the following problem:

$$\min_W \sum_{k=1}^{c-1}\sum_{j=k+1}^{c}\frac{n_jn_k}{n^2}\big(SD_W(l_j\|l_k)\big)^{-1}. \tag{41}$$

According to Equation (40), it is rewritten as

$$\min_W \sum_{k=1}^{c-1}\sum_{j=k+1}^{c}\frac{n_jn_k}{n^2}\Big(Tr((W^TS_wW)^{-1}W^TS_{jk}W)\Big)^{-1}. \tag{42}$$

Maximizing the harmonic mean in Equation (42) is much more difficult than maximizing the arithmetic mean in Equation (34) since the ratio trace is used as objective for class pairs. A conjugate gradient method is used to solve the problem Equation (42), which also converges very slow in some cases.

In [43], the following problem is proposed to maximize the closest class pair:

$$\max_{W^TW=I}\min_{j,k}\frac{Tr(W^TS_b^{jk}W)}{Tr(W^TS_w^{jk}W)}. \tag{43}$$

It has been shown in [44] that the convex hull of the set $\{M_W|M_W = W^TW, W^TW = I, W\in\mathbb{R}^{d\times m}\}$ is the following set: $\{M|TrM = m, 0\preceq M\preceq I, M\in R^{d\times d}\}$. So the convex relaxation of the problem Equation (43) can be rewritten as

$$\begin{aligned}\max_{W,\delta}\quad & \delta\\ s.t.\quad & \forall j,k, Tr(S_b^{jk}Z)\geq\delta Tr(S_w^{jk}Z),\\ & 0\preceq Z\preceq I, Tr(Z) = m,\end{aligned} \tag{44}$$

which is a Semi-definite programming (SDP) and can be solved with optimal solution. However, SDP is very time consuming and can only handle small scale data set.

In [45], the following problem is proposed to maximize the worst case:

$$\max_{W^TW=I}\frac{\min_{j,k}Tr(W^TS_{jk}W)}{\max_kTr(W^TS_w^kW)}. \tag{45}$$

Using the similar trick as in [39], [46], the problem Equation (45) can be solved by iteratively solving the following problem:

$$\max_{W^TW=I}\min_{j,k}Tr(W^TS_{jk}W) - \lambda\max_kTr(W^TS_w^kW), \tag{46}$$

where λ is the objective value of Equation (45) with the current solution W. Similarly, the convex relaxation of the problem Equation (46) can be rewritten as

$$\begin{aligned}\max_{W,s,t}\quad & s-\lambda t\\ s.t.\quad & \forall j,k, Tr(S_{jk}Z)\geq s, \forall k, Tr(S_w^kZ)\leq t,\\ & 0\preceq Z\preceq I, Tr(Z) = m,\end{aligned} \tag{47}$$

which can be solve with SDP but it is very time consuming.

Recently, [47] proposed to solve the following problem:

$$\max_{W^TW=I}\min_{j,k}\frac{n^2}{n_jn_k}Tr(W^TS_{jk}W). \tag{48}$$

Similarly, the convex relaxation of the problem (48) can be rewritten as

$$\begin{aligned}\max_{W,\delta}\quad & \delta\\ s.t.\quad & \forall j,k, \frac{n^2}{n_jn_k}Tr(S_{jk}Z)\geq\delta,\\ & 0\preceq Z\preceq I, Tr(Z) = m.\end{aligned} \tag{49}$$

In [47], a further local SDP relaxation is introduced and sequential SDP is used to solve the relaxed problem of Equation (48), which is also very time consuming and can only handle data set with very small scale.

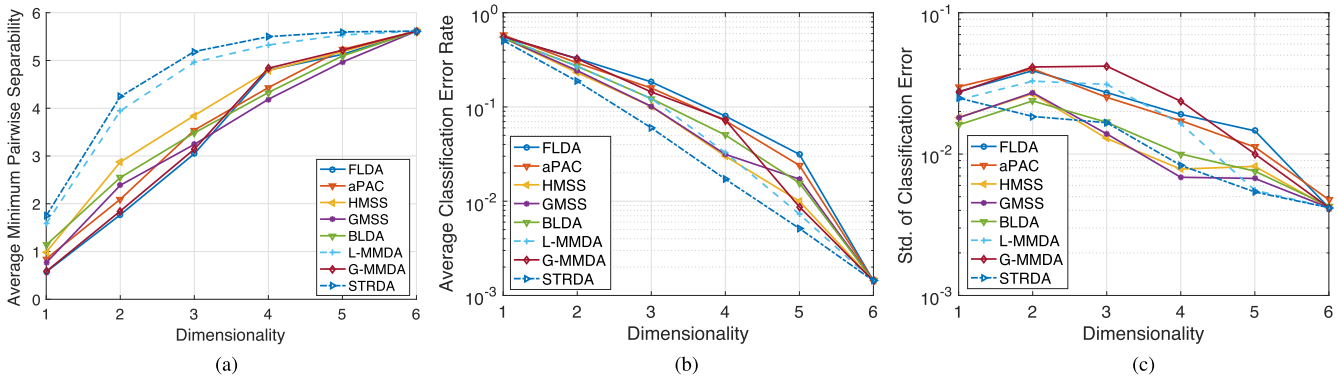


Fig. 1. We compare the performance of FLDA, aPAC, HMSS, GMSS, BLDA, L-MMDA, G-MMDA, and STRDA on the uniformly distributed dataset. The evaluation metric we used here are (a) the average minimum pairwise distance; (b) average classification error rate w.r.t. reduced dimensionality; and (c) the standard deviation of classification error rate w.r.t. reduced dimensionality.

5 Experiments

In this section, statistical experiments are conducted to evaluate the effectiveness of the proposed algorithm on two synthetic datasets, six face datasets.

5.1 Synthetic Data Test

We first conduct statistical experiments on synthetic data to show the effectiveness of the proposed algorithm in terms of discriminative dimension reduction. Following [47], we consider a sever-class classification problem represented by seven 10-dimensional homoscedastic Gaussians. The common covariance matrix is $\mathbf{I}_{10}$. We randomly sample the distinct class means from a 10-dimensional Guassian distribution with zero mean and a covariance matrix of $2\mathbf{I}_{10}$. The class means are sampled 500 times. For each time of the realizations, we generate 200 samples, 100 for training and the remaining for test, for all of the seven different classes. By doing so, we have 500 independent groups of training and test samples. This dataset is called uniformly distributed dataset. For the other dataset, we start with the same procedure, following by adding a bias of 15 to the first dimension of the means of the first three classes while sampling the means of the seven classes from the Gaussian distribution. This dataset is named as the nonuniformly distributed dataset, which is used to evaluate whether the dimension methods will be affeted by the nonuniform distribution of classes.

We compare the effectiveness of proposed method with FLDA [16], aPAC [41], HMSS [42], GMSS [42], BLDA [48], L-MMDA [47] and G-MMDA [47]. For all the compared algorithms, we first project the original data into the subspace with varying dimensions from one to six. The nearest neighbor (NM) classifier is used in all the experiments. We consider the following performance evaluation metrics. (1) minimum pairwise distance in the projected low-dimensional subspace: the largest minimum pairwise distance indicates the best discriminant ability. (2) average classification error rate with stand deviation: this evaluation metric has been widely used for discriminant dimension reduction methods. (3) two-dimensional graphical representation of classes, which can visualize the separability of the discriminant dimension reduction methods.

For each dimension reduction method, we calcluate the minimum pairwise distance in the projected subspace by averaging 500 indepdent runnings. We report the experimental results on uniformly and nonuniformly distributed dataset in Figs. 1a and 3a, respectively. From the experimental results we can see that the proposed algorithm generally perform much better than the other compared algorithms (except for projecting to 6-dimensional subspace, for which all the compared algorithms obtain the same performance).

The performance on the two synthetic datasets in terms of average classification error rate and stand deviation are reported in Figs. 1b, 1c, 3b, and 3c. These results are plotted in a log scale. From the experimental results we can see that the proposed algorithm obtains the best performance on both datasets.

To further evaluate the discriminant ability of the proposed algorithm, we show two-dimensional plots of all the compared methods. For each synthetic dataset, we raondly pick oen group of training data. The graphical representations are shown in Figs. 2 and 4. From these experimental results we observe that the traditional FLDA can not seperate classes well. We can also see that the proposed algorithm gets the best performance on both uniformly and nonuniformly distributed datasets.

5.2 Real-World Datasets

To further demonstrate the discriminant ability of the proposed algorithm, we conduct additional experiments on real-world datasets.

5.2.1 *Experiments on Object Recognition*

We report the experimental results of the proposed algorithm using a well-know object recognition dataset, the Coil20 dataset [49], in which there are 1,440 size normalized object images divided into 20 classes. The objects have a wide variety of complex geometric and reflectance characteristics. We divide the entire dataset into two parts (a training set and a testing set) using 10-fold cross validation. In our experiments, the entire Coil20 dataset are used to test the classification performance of all the compared algorithms.

We apply all the compared algorithms to the Coil20 dataset, and employ nearest neighbor (NN) as a classifier. We show the average classification error rate versus subspace dimensionality in Table 2 and Fig. 5. From the experimetnal results we can see that the proposed algorithm gives the

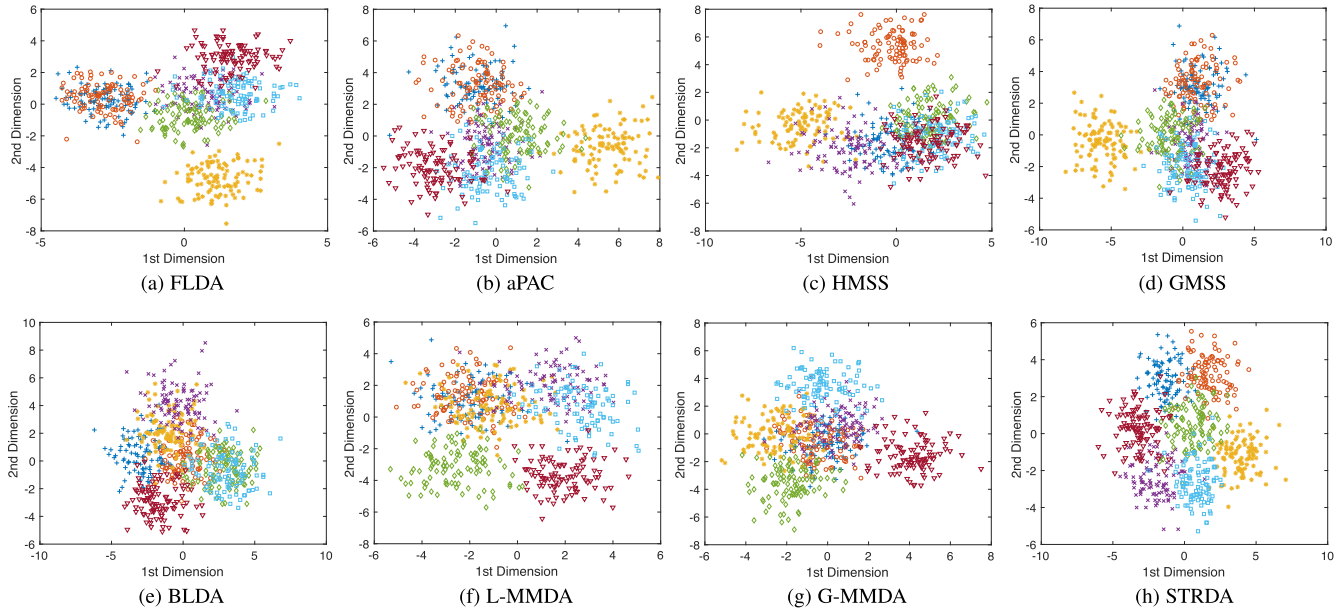


Fig. 2. We randomly select one training set from the uniformly distributed dataset and plot two-dimensional data representation. (a) FLDA, (b) aPAC, (c) GMSS, (d) HMSS, (e) BLDA, (f) L-MMDA, (g) G-MMDA, and (h) STRDA. From the graphical representation, we can see that the proposed algorithm clearly has the best separability.

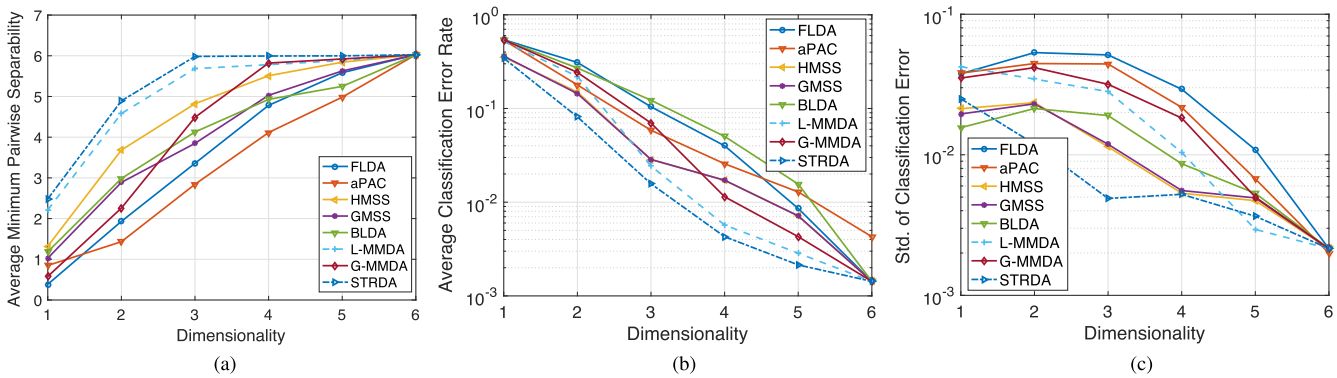


Fig. 3. We compare the performance of FLDA, aPAC, HMSS, GMSS, BLDA, L-MMDA, G-MMDA and STRDA on the nonuniformly distributed dataset. The evaluation metric we used here are (a) the average minimum pairwise distance; (b) average classification error rate w.r.t. reduced dimensionality; and (c) the standard deviation of classification error rate w.r.t. reduced dimensionality.

best performance for all the cases among all the compared algorithms on the object recognition dataset. When the number of the selected features is small, the proposed algorithm gets much better performance than the other compared algorithms. When the number of the selected features is large enough, they perform similarly.

5.2.2 *Experiments on Face Recognition*

We evaluate the performance of the proposed algorithm in terms of face recognition and compare it with the other state-of-the-art methods. We utlize six benchmark face datasets, Umist, JAFFE, YaleB, FERET, PIE and ORL, in this evaluation. The Umist dataset [50] contains 575 face images from 20 different people. Each image was resized to 28×23. The JAFFE dataset [51] consists of 213 images of 7 facial expressions posed by 10 Japanese female models. The images are cropped to 32×32. The YaleB dataset [52] contains 2,414 near frontal images from 38 persons under different illuminations. Each image is resized to 32×32. The ORL dataset [53] consists of 10 face images from 40 subjects for a total of 400 images, with some variation in pose, facial expression and details. The images were resized to 32×32. The PIE dataset [54] consists of 41,368 images of 68 people. Each person was imaged under 13 different poses, 43 different illumination conditions, and with 4 different expressions. The FERET dataset [55] contains 800 still facial images from 200 classes.

Since there are no official splits for these datasets, we first determine the number of images for each subject for training using 10-fold cross-validation. With the selected images, one class (subject) can be properly represented. We repeated this procedure for each dataset 50 times independently and reported the average performances.

In this experiment, we preprocessed the data for all the compared dimension reduction algorithms. We preserved the complete principal space. To avoid the singularity problem, we added the covariance matrix with a small term. After we apply the dimension reduction methods to the dataset, we use the nearest neighbor classifier to do recognition in the

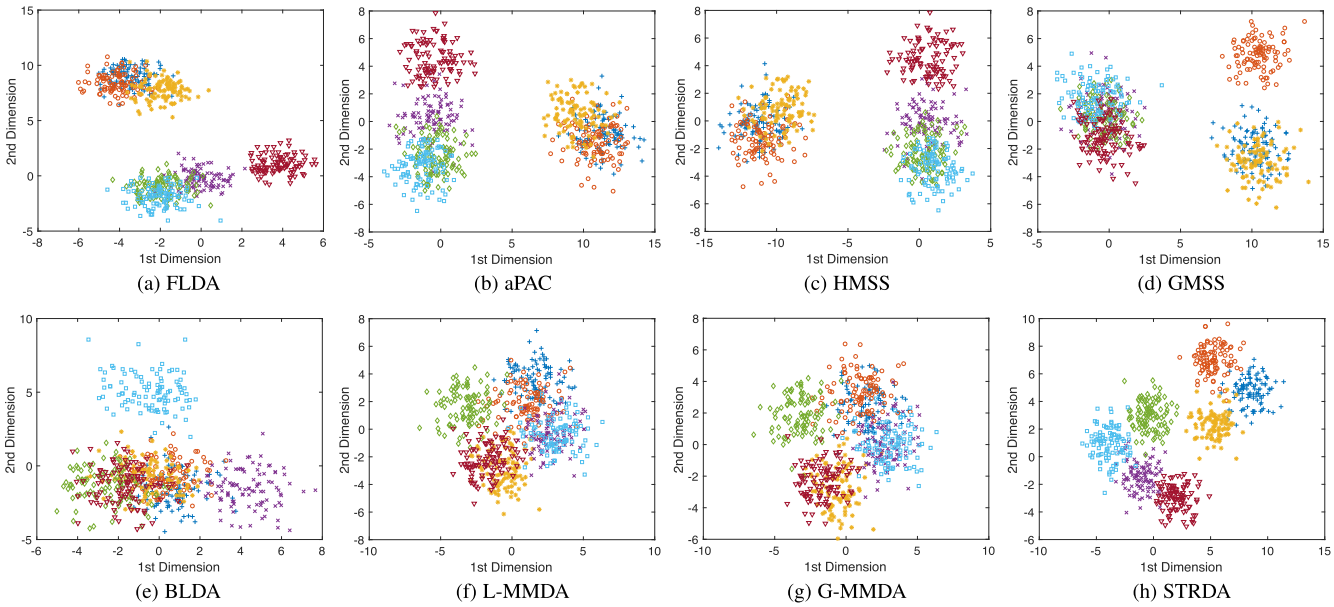


Fig. 4. We randomly select one training set from the nonuniformly distributed dataset and plot two-dimensional data representation. (a) FLDA, (b) aPAC, (c) GMSS, (d) HMSS, (e) BLDA, (f) L-MMDA, (g) G-MMDA, and (h) STRDA. From the graphical representation, we can see that the proposed algorithm clearly has the best separability.

TABLE 2
Performance (Error Rates) of All the Compared Algorithms on the Coil20 Dataset

Dim	1	3	5	7	9	11	13	15	17	Best
FLDA	0.644	0.289	0.172	0.131	0.151	0.127	0.118	0.114	0.117	0.114(15)
aPAC	0.627	0.178	0.117	0.105	0.100	0.103	0.107	0.107	0.109	0.100(9)
HMSS	0.591	0.131	0.101	0.102	0.095	0.099	0.102	0.102	0.106	0.095(9)
GMSS	0.595	0.133	0.106	0.099	0.100	0.103	0.108	0.109	0.110	0.099(7)
BLDA	0.620	0.181	0.119	0.104	0.100	0.102	0.108	0.106	0.110	0.100(9)
L-MMDA	0.682	0.198	0.132	0.114	0.101	0.098	0.110	0.109	0.112	0.098(11)
G-MMDA	0.703	0.267	0.144	0.115	0.105	0.098	0.110	0.109	0.112	0.098(11)
STRDA	0.551	0.111	0.081	0.072	0.075	0.079	0.092	0.092	0.096	0.072(7)

Nearest neighbor is used as a classifier.

TABLE 3
Classification Error Rate of All the Compared Algorithms on the Six Face Datasets (JAFFE, UMIST, ORL, YaleB, PIE, and FERET)

Dataset	JAFFE	UMIST	ORL	YaleB	PIE	FERET
FLDA	0.074	0.095	0.058	0.104	0.123	0.084
aPAC	0.063	0.086	0.074	0.082	0.131	0.091
HMSS	0.064	0.089	0.051	0.094	0.128	0.089
GMSS	0.071	0.093	0.053	0.099	0.135	0.084
BLDA	0.059	0.083	0.049	0.101	0.119	0.077
L-MMDA	0.056	0.076	0.058	0.093	0.084	0.072
G-MMDA	0.064	0.084	0.047	0.069	0.073	0.081
STRDA	0.038	0.061	0.041	0.064	0.068	0.066

Nearest neighbor is used as a classifier.

whitened space. We report the best performance in Table 3, from these experimental results we can see that the proposed algorithm clearly outperforms the other compared dimension reduction algorithms.

6 Conclusions

In this paper, a new criterion is exploited to extent the conventional trace ratio based LDA via maximizing the weighted harmonic mean of trace ratios, which effectively avoid the domination problem while did not raise any difficulties. We propose an efficient algorithm to solve the proposed challenging problems with fast convergence, which might always find the globally optimal solution just using eigenvalue decomposition in each iteration. Extensive experimental results illustrate the effectiveness and superiority of the proposed method over both of synthetic datasets and real-life datasets for various tasks in comparison with other compared methods on all of the datasets. In the future, we will deploy the proposed

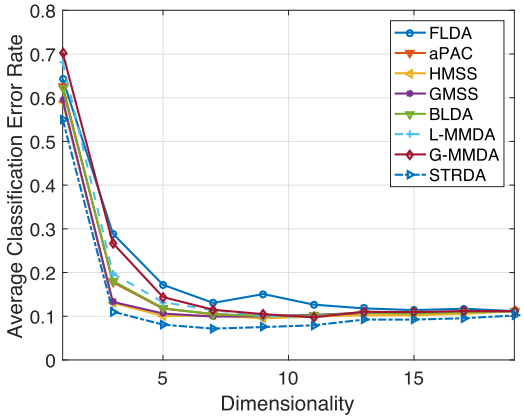


Fig. 5. Object recognition on Coil20 dataset. The nearest neighbor classifier is used as an evaluation metric.

algorithm to other real-world applications, i.e., person re-identification.

ACKNOWLEDGMENTS

This work was in part supported by the Data to Decisions Cooperative Research Centre (www.d2dcrc.com.au).

REFERENCES

[1] Y. Han, F. Wu, Q. Tian, and Y. Zhuang, "Image annotation by input-output structural grouping sparsity," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 3066–3079, Jun. 2012.

[2] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[3] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.

[4] M. Wang, W. Fu, S. Hao, H. Liu, and X. Wu, "Learning on big graph: Label inference and regularization with anchor hierarchy," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1101–1114, May 2017.

[5] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1864–1877, Jul. 2016.

[6] M. Wang, X. Liu, and X. Wu, "Visual classification by l1-hypergraph modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2564–2574, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2015.2415497

[7] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, Jan. 2014.

[8] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[9] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.

[10] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2494–2502, Dec. 2016.

[11] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 252–264, Feb. 2015.

[12] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1171–1177.

[13] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1302–1308.

[14] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[15] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *J. Roy. Statist. Soc. B: Methodological*, vol. 10, no. 2, pp. 159–203, 1948.

[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. 2nd ed. New York, NY, USA: Academic Press, 1990.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[18] S. Yuan, X. Mao, and L. Chen, "Multilinear spatial discriminant analysis for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2669–2681, Jun. 2017.

[19] Y. Zhou and S. Sun, "Manifold partition discriminant analysis," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 830–840, Apr. 2017.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[21] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 623–627, Jun. 2000.

[22] M. Loog, R. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.

[23] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[24] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ, USA: Wiley, 2004.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[26] Y. Zhao and S. Zhang, "Generalized dimension-reduction framework for recent-biased time series analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 231–244, Feb. 2006.

[27] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[28] W. Bian and D. Tao, "Harmonic mean for subspace selection," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[29] Y. Han, Y. Yang, F. Wu, and R. Hong, "Compact and discriminative descriptor inference using multi-cues," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5114–5126, Dec. 2015.

[30] P. Wang, C. Shen, H. Zheng, and Z. Ren, "A variant of the trace quotient formulation for dimensionality reduction," in *Proc. Asian Conf. Comput. Vis.*, 2009, pp. 277–286.

[31] W. Bian and D. Tao, "Max-min distance analysis by using sequential SDP relaxation for dimension reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 1037–1050, May 2011.

[32] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[33] Z. Zha, M. Wang, Y. Zheng, Y. Yang, R. Hong, and T. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, Feb. 2012.

[34] Y. Zhang and D.-Y. Yeung, "Worst-case linear discriminant analysis," in *Proc. Advances Neural Inf. Process. Syst.*, 2010, pp. 2568–2576.

[35] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio versus ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[36] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, Apr. 2009.

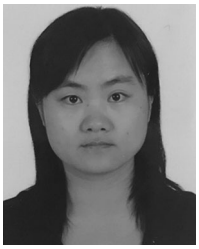
[37] M. Zhao, R. H. Chan, P. Tang, T. W. Chow, and S. W. Wong, "Trace ratio linear discriminant analysis for medical diagnosis: A case study of dementia," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 431–434, May 2013.

[38] L.-H. Zhang, "Uncorrelated trace ratio linear discriminant analysis for undersampled problems," *Pattern Recognit. Lett.*, vol. 32, no. 3, pp. 476–484, 2011.

[39] H. Wang, S. Yan, D. Xu, X. Tang, and T. S. Huang, "Trace ratio versus ratio trace for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[41] M. Loog, R. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.
[42] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.
[43] P. Wang, C. Shen, H. Zheng, and Z. Ren, "A variant of the trace quotient formulation for dimensionality reduction," in *Proc. Asian Conf. Comput. Vis.*, 2009, pp. 277–286.
[44] M. L. Overton and R. S. Womersley, "Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices," *Math. Program.*, vol. 62, no. 1–3, pp. 321–357, 1993.
[45] Y. Zhang and D.-Y. Yeung, "Worst-case linear discriminant analysis," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 2568–2576.
[46] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, Apr. 2009.
[47] W. Bian and D. Tao, "Max-min distance analysis by using sequential SDP relaxation for dimension reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 1037–1050, May 2011.
[48] O. C. Hamsici and A. M. Martínez, "Bayes optimality in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 647–657, Apr. 2008.
[49] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Tech. Rep. CUCS-005-96, Feb. 1996.
[50] T. J. Millar, P. R. A. Farquhar, and K. Willacy, "The UMIST database for astrochemistry 1995," *Astronomy and Astrophysics Supplement Series*, vol. 121, no. 1, pp. 139–185, 1997.
[51] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd Int. Conf. Face Gesture Recognit.*, 1998, pp. 200–205.
[52] 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 Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[53] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, 1994, pp. 138–142.
[54] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[55] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.



Zhihui Li received the BS degree from the Beijing University of Posts and Telecommunications, in 2008. She is currently working as a data analyst with Beijing Etrol Technologies Co., Ltd. Her research interests include artificial intelligence, machine learning, and computer vision.

Feiping Nie received the PhD degree in computer science from Tsinghua University, Beijing, China, in 2009. He is currently a professor with the Center for OPTical Imagery Analysis and Learning, Northwestern Polytechnical University, Shaanxi, China. His research interests include machine learning and its applications fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He has published more than 100 papers in the prestigious journals and conferences like the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Knowledge and Data Engineering*, ICML, NIPS, KDD, and etc. He is now serving as an associate editor or a PC member for several prestigious journals and conferences in the related fields.

Xiaojun Chang received the PhD degree from the Centre for Quantum Computation and Intelligent Systems (QCIS), University of Technology Sydney, Australia, in 2016. After that, he has been working as a postdoctoral research associate in the Language Technologies Institute, Carnegie Mellon University. His research interests include machine learning, data mining, and computer vision.

Yi Yang received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. He was a post-doctoral research in the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video semantics understanding.