This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

Towards Robust Discriminative Projections Learning via Non-greedy $\ell_{2,1}$-Norm MinMax

Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, and Xuelong Li, *Fellow, IEEE*

Abstract—Linear Discriminant Analysis (LDA) is one of the most successful supervised dimensionality reduction methods and has been widely used in many real-world applications. However, the ℓ_2-norm is employed as the distance measure in objective of LDA, which is sensitive to outliers. Many previous works improve the robustness of LDA by using ℓ_1-norm distance. However, the robustness against outliers is limited and the solver of ℓ_1-norm are mostly based on greedy search strategy, which is time-consuming and easy to get stuck in a local optimum. In this paper, we propose a novel robust LDA measured by $\ell_{2,1}$-norm to learn robust discriminative projections. Since the proposed model needs to minimize and maximize (minmax) $\ell_{2,1}$-norm terms simultaneously, it is challenging to solve. As a result, we first systematically derive an efficient iterative optimization algorithm to solve a general ratio minimization problem, and rigorously prove its convergence. More importantly, an efficient non-greedy iterative re-weighted optimization algorithm is developed based on preceding approach for solving proposed $\ell_{2,1}$-norm minmax problem. Besides, an optimal weighted mean mechanism is driven according to designed objective and solver, which can be applied to other approach for robustness improvement. Extensive experimental results on several real-world datasets show the effectiveness of proposed method.

Index Terms—Robust dimensionality reduction, $\ell_{2,1}$-norm minmax problem, non-greedy iterative re-weighted solver, optimal weighted mean, outlier.

1 INTRODUCTION

Dimensionality Reduction (DR), one of the most fundamental problems in machine learning, has been widely used in many real-world applications, such as face recognition [1], information retrieval [2], medical image processing [3] *etc*. Recently, some dimensionality reduction methods combine with Deep Learning (DL) model to explore nonlinear features for improving the performance of recognition [4], [5]. However, training a Deep Neural Network (DNN) requires a large number of labeled data points, which is resource-consuming and not suitable for small scale datasets. As a result, we prefer to focus on optimizing conventional machine learning model which can satisfy more practical demands such as low-rank [6], sparse [7] and so on. Linear Discriminant Analysis (LDA) [8] and Principal Component Analysis (PCA) [9] are the most popular linear dimensionality reduction methods and have been widely used in data preprocessing stage.

- *Feiping Nie is corresponding author, he at School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R.China. E-mail: feipingnie@gmail.com*
- *Zheng Wang and Rong Wang are with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R.China. E-mail: zhengwangml@gmail.com, wangrong07@tsinghua.org.cn*
- *Zhen Wang is with Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China, and also with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: zhenwang0@gmail.com*
- *Xuelong Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R.China. E-mail: li@nwpu.edu.cn*

Owning to LDA utilizes label information in training stage, it often outperforms PCA in classification task. However, prototype LDA always suffers from some drawbacks such as Small Sample Size problem (SSS) [10], [11], worst-case problem [12], [13], non-Gaussian issue [14] and sensitive to outliers [15]. This paper concentrates on solving the last one for improving robustness against outliers. Measuring the data distance by using ℓ_2-norm enlarges the influence of outliers on objective function value. Besides, the estimation of class mean deviates from the distribution of normal data samples as well as the calculation of within-class and between-class scatter are inaccurate.

In the past decades, one of the major topics to be investigated in this field is to develop robust model for alleviating impact of outliers. Wherein, one way is to make use of class median sample to represent each class for improving the robustness of model [16]–[18], which is however difficult to guarantee convergence. In the early time, another popular manner is to use ℓ_1-norm to improve model's robustness, and plenty of robust PCA algorithms based on ℓ_1-norm [19]–[24] is proposed. Motivated by great success on ℓ_1-PCA model, more and more robust LDA based on ℓ_1-norm have been developed for handling the outlier issue [25], [26]. For instance, Zhong, F *et al.*, [27] directly apply ℓ_1-norm on LDA (LDA-ℓ_1) to improve robustness. However, the solver in LDA-ℓ_1 is built on a greedy search strategy so that the projections need to be sought one by one, which is time-consuming and prone to fall into local optimum. A robust Distance Metric Learning using ℓ_1-norm (ℓ_1-DML) [28] is derived by minimizing and maximizing a number of non-smooth ℓ_1-norm terms, the optimal solution however is hard to be achieved as well. To address this issue, an Improved LDA based on ℓ_1-norm (ILDA-ℓ_1) [29] constructs a successive concave approximation to the original

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

objective and optimizes all projections simultaneously by using the classic projected subgradient method with Armijo line search. However, it only can obtain an approximate solution, namely above issues existed in ℓ_1-norm method still unsolved completely. Recently, a non-greedy iterative optimization algorithm is proposed in [30] (ℓ_1-LDA) that can acquire a closed-form solution of all projections.

Although the aforementioned ℓ_1-norm based methods improve the robustness of model to some extent, they still have several disadvantages: 1) The objective based on ℓ_1-norm is not invariant to rotation, which will degrade the subsequent classification performance [31]. 2) ℓ_1-norm is difficult to achieve the desirable robustness, especially when the number of outliers in training set is large [32]. To alleviate the first drawback, C. Ding *et al.,* [33] propose the rotation invariant ℓ_1-norm (R_1-norm), and several extensive works incorporate R_1-norm to improve robustness, such as robust tensor factorization [34], robust R_1-PCA [20] and so on. Analogously, Nie *et al.,* [35] attempt to use $\ell_{2,1}$-norm based loss function to overcome outliers and a $\ell_{2,1}$-norm regularization to fulfill feature selection. After that, plenty of $\ell_{2,1}$-norm based robust models [36]–[40] have aroused great attention among researchers.

Motivated by above works, in this paper, we propose a novel supervised dimensionality reduction method named $\ell_{2,1}$-LDA which learns robust discriminative projections by minimizing and maximizing $\ell_{2,1}$-norm problem simultaneously. Our model not only calculates the within-class scatter based on $\ell_{2,1}$-norm but also uses $\ell_{2,1}$-norm to calculate the covariance of total samples. In other words, the influence of outliers on estimation of within-class scatter and total samples scatter can be alleviated. Besides, proposed model takes advantage of trace ratio criterion to learn the projections, which is beneficial to subsequent classification task. However, the major challenge of proposed model is to minimize and maximize $\ell_{2,1}$-norm problem simultaneously, which has not yet been addressed in other works. Fortunately, we first develop an efficient optimization algorithm to solve the general ratio minimization problem with rigorously prove convergent. Then, a non-greedy iterative re-weighted optimization algorithm is presented to solve proposed non-convex $\ell_{2,1}$-norm minmax problem. Experimental results demonstrate that proposed $\ell_{2,1}$-LDA is more robust to outliers than conventional LDA and other SOTA robust LDA varieties. The contributions of our paper can be summarized as following three aspects:

- A novel $\ell_{2,1}$-norm LDA model with trace ratio criterion is proposed for robust discriminative projections learning. In proposed model, the within-class scatter and total sample scatter are measured by $\ell_{2,1}$-norm, which learns row or column sparsity loss matrix so that the outliers can be efficiently suppressed. Additionally, we also provide a strict proof to verify that conventional LDA in trace ratio formulation exists a trivial solution, which is first proposed in this paper.
- A novel iterative optimization algorithm is provided to solve general ratio minimization problem with strict convergence proofs. Meanwhile, an efficient non-greedy iterative re-weighted optimization algorithm is proposed to solve the non-convex $\ell_{2,1}$-norm minmax problem, which can obtain more discriminative projections so as to improve classification performance and is timesaving than greedy search strategy.
- An optimal weighted mean mechanism is driven by designed root loss function and non-greedy re-weighted optimization algorithm, which guarantees the correctness on distribution estimation of each class. Moreover, proposed optimal weighted mean mechanism can be extensively applied to original LDA model for improving its robustness against outliers.

The rest of paper is organized as follows. In Section 2, we will innovatively proof that prototype LDA always has a trivial solution, then several related robust LDA model will be introduced. Our novel $\ell_{2,1}$-LDA model will be presented in Section 3. In Section 4, the iterative optimization algorithm to solve general ratio minimization problem and the corresponding non-greedy iterative re-weighted optimization algorithm are provided. Extensive experiments conducted on toy and several real-world gray and RGB image datasets will be put in Section 5. Finally, Section 6 concludes all paper. *Matlab code of our algorithm and partial open datasets can be downloaded from this address.*[1]

Notation. For a matrix $A \in \mathbb{R}^{d\times m}$, ℓ_1-norm and $\ell_{2,1}$-norm of A are respectively defined as

$$\|A\|_1 = \sum_{i=1}^{d}\sum_{j=1}^{m}|a_{ij}|, \quad \|A\|_{2,1} = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{m}a_{ij}^2}. \tag{1}$$

2 RELATED WORK

In this section, we provide an innovative discussion on the solution of LDA and prove that prototype trace ratio LDA exists trivial solution. Then, we review a classic robust LDA model based on ℓ_1-norm.

2.1 Discussion on Solution of Prototype LDA

Given the dataset $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d\times n}$, where d and n denote the dimensionality of each data and the number of all data points respectively. In general, the data X should be centralized, i.e., $\sum_{i=1}^{n} x_i = 0$ in preprocessing stage. It is well-known that the task of Linear Discriminant Analysis (LDA) is to learn a transformation matrix $W \in \mathbb{R}^{d\times m}$ so as to project high-dimensional data into low-dimensional subspace while maximizing the total samples scatter and minimizing the within-class samples scatter simultaneously. Concretely, prototype LDA is to solve following ratio trace problem:

$$\min_{W} Tr\left[(W^T S_t W)^{-1}(W^T S_w W)\right], \tag{2}$$

where $Tr(\cdot)$ denotes the trace of matrix, $S_w = \sum_{k=1}^{c}\sum_{j=1}^{n_k}(x_j - m_k)(x_j - m_k)^T$ is the within scatter matrix, wherein c is the number of classes, n_k represents the number of data points in k-th class and $m_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_j^k$ denotes the mean sample of k-th class, $S_t = \frac{1}{n}XX^T$ is the total scatter matrix. [41] claims that problem (2) can be efficiently solved by using generalized eigenvalue decomposition method, however its solution is invariant so that the subsequent classification and clustering performance are

1. https://github.com/StevenWangNPU/Robust-L21-LDA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

not stable. Another trace ratio version of LDA is developed, which however has trivial solution, which has never been mentioned in previous efforts.

Here, we propose following **Theorem 1** to prove the trivial solution issue in trace ratio LDA, which is first presented in our paper.

Theorem 1 *Supposing* $w^* = arg\min_w \frac{w^T S_w w}{w^T S_t w}$, *then* $W = [w^*, ..., w^*]$ *is a trivial solution for trace ratio LDA.*

Before we prove **Theorem 1**, following **Lemma 1** should be introduced firstly.

Lemma 1 *For any non-negative real numbers* $a_i \geq 0$, $b_i \geq 0$ $(1 \leq i \leq m)$, *if* $\frac{a_1}{b_1} \leq \frac{a_2}{b_2} \leq \cdots \leq \frac{a_m}{b_m}$, *then* $\frac{a_1}{b_1} \leq \frac{a_1+a_2+\cdots+a_m}{b_1+b_2+\cdots+b_m}$.

Proof: Supposing $\frac{a_1}{b_1} = r$, for any $a_i \geq 0$, $b_i \geq 0$ $(1 \leq i \leq m)$, we have $rb_i \leq a_i$, then $\frac{a_1+a_2+\cdots+a_m}{b_1+b_2+\cdots+b_m} \geq \frac{rb_1+rb_2+\cdots+rb_m}{b_1+b_2+\cdots+b_m} = r = \frac{a_1}{b_1}$. □

Next, we present the proof of **Theorem 1** as follows:

Proof: Trace ratio LDA can be decomposed as: $\frac{Tr(W^T S_w W)}{Tr(W^T S_t W)} = \frac{\sum_{i=1}^m w_i^T S_w w_i}{\sum_{i=1}^m w_i^T S_t w_i}$. Without loss of generality, we suppose $\frac{w_{i_1}^T S_w w_{i_1}}{w_{i_1}^T S_t w_{i_1}} \leq \frac{w_{i_2}^T S_w w_{i_2}}{w_{i_2}^T S_t w_{i_2}} \leq \cdots \leq \frac{w_{i_m}^T S_w w_{i_m}}{w_{i_m}^T S_t w_{i_m}}$. According to **Lemma 1**, we can obtain $\frac{w_{i_1}^T S_w w_{i_1}}{w_{i_1}^T S_t w_{i_1}} \leq \frac{Tr(W^T S_w W)}{Tr(W^T S_t W)}$. Additionally, according to the assumption in **Theorem 1**, we have $\frac{w^{*T} S_w w^*}{w^{*T} S_t w^*} \leq \frac{w_{i_1}^T S_w w_{i_1}}{w_{i_1}^T S_t w_{i_1}}$. Therefore, the trivial solution is $W = [w^*, ..., w^*]$ where all columns of W are equal to w^*. □

To the best of our knowledge, **Theorem 1** is firstly proposed in our paper, which is very important to deepen our understanding of LDA's solution. Consequently, an efficient but simple strategy to avoid trivial solution in trace ratio LDA is to use orthogonal constraint to keep difference between all columns of W. Without loss of generality, a more reasonable formulation of LDA is to solve following constrained trace ratio problem:

$$\min_{W^T W = I} \frac{Tr(W^T S_w W)}{Tr(W^T S_t W)}. \tag{3}$$

Note that the Eq.(3) can be rewritten as following vector-based problem:

$$\min_{W^T W = I} \frac{\sum_{k=1}^c \sum_{x_i \in \pi_k} \left\| W^T (x_i - m_k) \right\|_2^2}{\frac{1}{n} \sum_{i=1}^n \left\| W^T x_i \right\|_2^2}, \tag{4}$$

where π_k denotes the set of samples in k-th class. However, due to the orthogonal constraint, such a constrained trace ratio LDA model is a non-convex optimization problem which does not have closed-form solution. Previously, two efforts [41], [58] are exclusively proposed to solve such a non-convex optimization problem as well as we will provide a novel iterative optimization algorithm with strict convergence proof to solve problem in Eq.(4).

2.2 Revisit of Robust LDA Based on ℓ_1-norm

In [27], the robust LDA based on ℓ_1-norm (LDA-ℓ_1) aims to solve following problem:

$$\max_{W^T W = I} \frac{\sum_{k=1}^c n_k \left\| W^T (m_k - \overline{x}) \right\|_1}{\sum_{k=1}^c \sum_{x_i \in \pi_k} \left\| W^T (x_i - m_k) \right\|_1}, \tag{5}$$

which is difficult to be solved. A greedy search method simply transforms it to following problem in which the projections can be found one by one

$$\max_{w^T w = 1} \frac{\sum_{k=1}^c n_k \left| w^T (m_k - \overline{x}) \right|}{\sum_{k=1}^c \sum_{x_i \in \pi_k} \left| w^T (x_i - m_k) \right|}. \tag{6}$$

As shown in [30], a non-greedy strategy optimization algorithm is proposed to solve following problem instead of above ratio problem

$$G(W, \lambda) = \arg \max_{W^T W = I} H(W) - \lambda M(W), \tag{7}$$

where $H(W) = \sum_{k=1}^c n_k \left\| W^T (m_k - \overline{x}) \right\|_1$ and $M(W) = \sum_{k=1}^c \sum_{x_i \in \pi_k} \left\| W^T (x_i - m_k) \right\|_1$. In problem (7), λ relates to W, and they need to be optimized iteratively. Specifically, in the k-th iteration, λ^k is calculated by the objective function value of problem (5) in $(k-1)$-th iteration, e.g.

$$\lambda^k = \frac{\sum_{k=1}^c n_k \left\| \left(W^{k-1}\right)^T (m_k - \overline{x}) \right\|_1}{\sum_{k=1}^c \sum_{x_i \in \pi_k} \left\| \left(W^{k-1}\right)^T (x_i - m_k) \right\|_1}. \tag{8}$$

After some algebraic calculations, all projections W can be obtained by projected subgradient method with Armigo line search at once.

In next section, we will propose a novel robust discriminative projections learning method which focuses on solving the $\ell_{2,1}$-norm minmax problem.

3 Robust Discriminative Projections Learning via $\ell_{2,1}$-norm Minmax

It is well-known that the ℓ_2-norm in Eq.(4) will enlarge the influences of outliers on objective function value. One strategy is to use ℓ_1-norm to replace ℓ_2-norm, which has proven to be not a wise approach. The first reason, according to Eq.(1), the distance in spatial dimensions and the summation between different data samples are both measured by using ℓ_1-norm. In other words, the subtle distinction between spatial dimensions and data points is lost. Besides, in Eq.(5), measuring the distance between data points and centroids in ℓ_1-norm violates model's assumption that the data distribution is Gaussian. In detail, from Fig. 1a we can observe that in ℓ_1-norm, the equidistance surface $\|x - y\|_1 = const$ is a simplex surface wherein the longest direction and shortest direction are p and $\sqrt{p}$ respectively when each data point has p-dimensional attributes, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

widens distance gaps among data points within same class and degrades the subsequent classification performance when dealing with high-dimensional data. Therefore, it is more reasonable to leverage ℓ_2-norm to measure the distance in spatial dimensions as the equidistance surface $\|x-y\|^2 = const$ is a sphere which is closer to Gaussian distribution. Last one, most of optimization algorithms for ℓ_1-norm problem are based on greedy search strategy, which is time-consuming and prone to fall into local optimum [30].

As shown in Eq.(1), $\ell_{2,1}$-norm is able to measure the distance of spatial dimensions in ℓ_2 and enforce sparsity over different data points for improving robustness against outliers in ℓ_1. Intuitively, in Fig. 1b, $\ell_{2,1}$-norm is designed to learn row sparsity structure (each row is a data point), which can reward the robustness against to outliers. In comparison, ℓ_1-norm concentrates on suppressing the anomaly over all values without considering the distinction between row and column, which can not efficiently eliminate the influence of outliers.

Based on aforementioned reasons, we design to employ $\ell_{2,1}$-norm instead of using ℓ_1-norm and ℓ_2-norm to minimize and maximize the within-class scatter and total sample scatter simultaneously for improving the discriminability of model at projected space as much as possible. Concretely, the robust model can be simply formulated as

$$\begin{cases} \min \sum_{k=1}^{c} \sum_{x_i \in \pi_k} \left\| W^T (x_i - m_k) \right\|_2 \\ \max \frac{1}{n} \sum_{i=1}^{n} \left\| W^T x_i \right\|_2 = \frac{1}{n} \left\| X^T W \right\|_{2,1}. \end{cases} \tag{9}$$

Additionally, it is worth noting that the outliers not only enlarge the objective function value but also affect the estimation of class mean and result in deviation of distribution of each class. To address this issue, we abandon using fixed class arithmetic mean to represent each class, instead setting class mean m_k as variable to be optimized for finding better class centroids. Then, above $\ell_{2,1}$-norm minmax problem in Eq.(9) can be unified to following ratio minimization problem:

$$\min_{W^T W = I, m_k} \frac{\sum_{k=1}^{c} \sum_{x_i \in \pi_k} \left\| W^T (x_i - m_k) \right\|_2}{\frac{1}{n} \left\| X^T W \right\|_{2,1}}. \tag{10}$$

Note that, since proposed minimization problem in Eq.(10) is a non-convex optimization problem, and there exists two variables, i.e., W and m_k need to be optimized, solving such a problem is challenging.

Comparing to existing robust dimensionality reduction methods, the superiorities of proposed model in Eq.(10) can be summarized as following three points:

- **Robustness**: The $\ell_{2,1}$-norm measurement used in proposed model not only can capture the subtle distinction between spatial dimensions and data points but also promote the sparsity at data points level rather than at feature level in ℓ_1-norm. Thus, the learned row or column sparsity structure is able to efficiently alleviate the influence of outliers.
- **Discriminability**: In Eq.(10), simultaneously minimizing within-class scatter and maximizing total samples scatter contributes to improve separability of all classes. In addition, unlike most existing robust LDA based on ℓ_1-norm, our elaborately designed non-greedy optimization algorithm in Section 3 updates all projections simultaneously, which can obtain much better solution than greedy search solvers. These above two facts can reward the discriminability of model and provide some cues for supporting the performance improvement on classification task of our method.
- **Extensibility**: The root loss driven optimal weighted mean m_k adaptively assigns weights to each sample (see in Eq.(26)) for weakening the deviation in class mean estimation, which can not be achieved in ℓ_1-norm based methods. Moreover, our optimal weighted mean can be extensively applied to classical dimension reduction method to improve its robustness against outliers, and some theoretical and practical evidences are shown in Section 4.4 and Section 5.5.4 respectively.

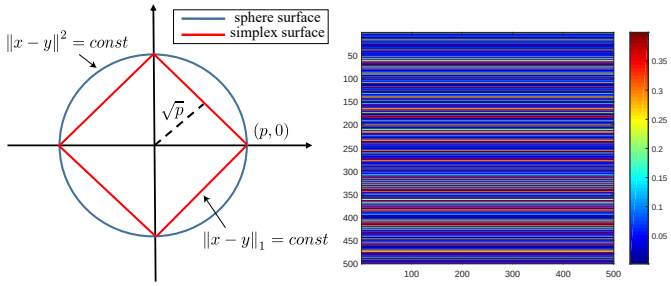


Fig. 1: (a) Illustration example of sphere surface and simplex surface in two-dimensional space. (b) Visualization of row sparsity structure learned by using $\ell_{2,1}$-norm. Each row denotes a data point and dark blue denotes the values elements are close to zero.

4 NON-GREEDY $\ell_{2,1}$-NORM MINMAX ITERATIVE RE-WEIGHTED OPTIMIZATION ALGORITHM

To our best knowledge, problem (10) involving simultaneously minimizing and maximizing $\ell_{2,1}$-norm problem that has not been solved perfectly in previous works. Thus, the following $\ell_{2,1}$-norm minmax iterative re-weighted optimization algorithm of is a major contribution of our paper. To solve problem (10), we first propose an efficient iterative optimization algorithm to solve a more general ratio minimization problem in what follows.

4.1 Algorithm to Solve the General Ratio Minimization Problem

We firstly consider solving following general ratio minimization problem:

$$\min_{v \in \mathcal{C}} \frac{\sum_i f_i(g_i(v))}{\sum_i h_i(q_i(v))}, \tag{11}$$

where $f_i(v)$ is an arbitrary concave function, while $h_i(v)$ is an arbitrary convex function in the domain of $g_i(v)$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

$q_i(v)$ respectively. $v \in \mathcal{C}$ is an arbitrary constraint on v, and the v, $g_i(v)$ and $q_i(v)$ can be scalar, vector or matrix. Besides, supposing $\forall v \in \mathcal{C}$, we always have $h_i(q_i(v)) \geq 0$.

Regrettably, directly finding the minimum ratio value of objective function in problem (11) is challenging. We develop a novel technique that can solve general problem (11) iteratively according to following Theorem 2.

Theorem 2 *The global minimal objective function value of problem (11) is the root of following function:*

$$J(\lambda) = \min_{v \in \mathcal{C}} \sum_i f_i(g_i(v)) - \lambda \sum_i h_i(q_i(v)), \tag{12}$$

where λ is objective function value of problem (11).

Proof: Suppose v^* denotes the global solution of problem (11), and its corresponding global minimal objective function value is λ^*, i.e., $\frac{f_i(g_i(v^*))}{h_i(q_i(v^*))} = \lambda^*$. As a result, $\forall v \in \mathcal{C}$, we can obtain $\frac{f_i(g_i(v))}{h_i(q_i(v))} \geq \lambda^*$. Combining the assumption, i.e., $h_i(q_i(v)) \geq 0$, we can achieve $f_i(g_i(v)) - \lambda^* h_i(q_i(v)) \geq 0$. It is evident that problem (12) is lower bounded: $J(\lambda^*) = 0$. In conclusion, the global minimal objective function value of problem (11) λ^* is the root of function (12). □

According to the Theorem 2 and supposing the global minimal objective function value of problem (11) is λ^*, then the optimal global solution v^* in problem (11) can be obtained by solving following problem:

$$v^* = \arg\min_{v \in \mathcal{C}} \sum_i f_i\left(g_i(v)\right) - \lambda^* \sum_i h_i\left(q_i(v)\right). \tag{13}$$

In a word, updating two variable v and λ iteratively until reaches convergence, and the proposed general minmax problem can be solve globally. Denote $f_i^{'}\left(g_i(v)\right)$ and $h_i^{'}\left(q_i(v)\right)$ are the supergradient of function f_i and h_i at point $g_i(v)$ and $q_i(v)$ respectively. Combining the Re-weighted optimization algorithm proposed in [35], an efficient iterative algorithm to solve proposed general ratio minimization problem can be summarized in following Algorithm 1.

Algorithm 1 An efficient iterative optimization algorithm to solve the general ratio minimization problem in Eq.(11).

Initialization: $v \in \mathcal{C}$, $t = 1$.
While not converge **do**
1. For each i, calculate the supergradient of function f_i and h_i at points $g_i(v^t)$ and $q_i(v^t)$: $F_i = f_i^{'}\left(g_i(v^t)\right)$, $H_i = h_i^{'}\left(q_i(v^t)\right)$
2. Calculate $\lambda^t = \dfrac{\sum_i f_i(g_i(v^t))}{\sum_i h_i(q_i(v^t))}$.
3. Calculate $v^{t+1} = \arg\min_{v \in \mathcal{C}} \sum_i Tr(F_i^T g_i(v)) - \lambda^t \sum_i Tr(H_i^T q_i(v))$.
4. $t = t + 1$.
end while

4.2 Convergence Analysis of Algorithm 1

Theorem 3 *The Algorithm 1 will decrease the objective function value of problem (11) in each iteration until convergent.*

Proof: For simplicity, we denote v^{t+1} as $\widetilde{v}$. In the t-th iteration of Algorithm 1 and according to the step 3, we have:

$$\begin{aligned} &\sum_i Tr(F_i^T g_i(\widetilde{v})) - \lambda^t \sum_i Tr(H_i^T q_i(\widetilde{v})) \\ &\leq \sum_i Tr(F_i^T g_i(v^t)) - \lambda^t \sum_i Tr(H_i^T q_i(v^t)), \end{aligned} \tag{14}$$

where the equality holds when the algorithm converges. Since the function $f_i(\cdot)$ and $h_i(\cdot)$ are the concave and convex function respectively, and we can obtain following two inequations according to the definition of supergradient [42]

$$f_i\left(g_i(\widetilde{v})\right) - f_i\left(g_i(v^t)\right) \leq Tr(F_i^T g_i(\widetilde{v})) - Tr(F_i^T g_i(v^t)) \tag{15}$$

$$h_i\left(q_i(\widetilde{v})\right) - h_i\left(q_i(v^t)\right) \geq Tr(H_i^T q_i(\widetilde{v})) - Tr(H_i^T q_i(v^t)). \tag{16}$$

Combining Eq.(15) and Eq.(16), we can get

$$\begin{aligned} &\sum_i \Big(f_i(g_i(\widetilde{v})) - f_i(g_i(v^t)) + Tr(H_i^T q_i(\widetilde{v})) - Tr(H_i^T q_i(v^t)) \Big) \\ &\leq \sum_i \Big(Tr(F_i^T g_i(\widetilde{v})) - Tr(F_i^T g_i(v^t)) + h_i(q_i(\widetilde{v})) - h_i(q_i(v^t)) \Big). \end{aligned} \tag{17}$$

Substituting Eq.(14) into Eq.(17), we have

$$\begin{aligned} &\sum_i \Big(f_i(g_i(\widetilde{v})) - f_i(g_i(v^t)) + Tr(H_i^T q_i(\widetilde{v})) - Tr(H_i^T q_i(v^t)) \Big) \\ &\leq \lambda^t \sum_i \Big(Tr(H_i^T q_i(\widetilde{v})) - Tr(H_i^T q_i(v^t)) \Big) + \sum_i \big(h_i(q_i(\widetilde{v})) \\ &\quad - h_i(q_i(v^t)) \big). \end{aligned} \tag{18}$$

According to the step 2 in Algorithm 1, we have $\sum_i f_i(g_i(v^t)) = \lambda^t \sum_i h_i(q_i(v^t))$, $\sum_i f_i(g_i(\widetilde{v})) = \lambda^{t+1} \sum_i h_i(q_i(\widetilde{v}))$, and substitute them into Eq.(18), then we can obtain

$$\begin{aligned} &(\lambda^{t+1} - 1) \sum_i h_i(q_i(\widetilde{v})) \leq (\lambda^t - 1) \sum_i \big(Tr(H_i^T q_i(\widetilde{v})) - \\ &Tr(H_i^T q_i(v^t)) + h_i(q_i(v^t)) \big) \leq (\lambda^t - 1) \sum_i h_i(q_i(\widetilde{v})) \\ &\Rightarrow \lambda^{t+1} \leq \lambda^t. \end{aligned} \tag{19}$$

Thus the Algorithm 1 will monotonically decrease the objective function value of general problem in Eq.(11) in each iteration until it converges. □

4.3 Non-greedy Optimization Algorithm to Solve $\ell_{2,1}$-norm MinMax Problem in Eq.(10)

In this section, we will present an efficient non-greedy iterative re-weighted optimization algorithm to solve $\ell_{2,1}$-norm minmax problem in Eq.(10). Beforehand, we define the function $f_i(\cdot)$ as

$$f_i(g_i(W, m_k)) = \sqrt{g_i(W, m_k)}, \tag{20}$$

where $g_i(W, m_k) = \left\| W^T(x_i - m_k) \right\|_2^2$, meanwhile, since the denominator of problem (10) is always larger than 0, the function $h_i(\cdot)$ can be absolute value function. It is evident that function $f_i(\cdot)$ and $h_i(\cdot)$ are concave and convex function respectively, thus the problem (10) is a special case of general

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

problem (11). Therefore, we can solve proposed problem (11) according to Algorithm 1.

Actually, the key point of Algorithm 1 is to solve the problem in the third step. In each iteration, the third step of Algorithm 1 aims to solve following root loss problem:

$$\min_{W^TW=I,m_k} \sum_{k=1}^{c} \sum_{x_i\in\pi_k} \left\|W^T(x_i-m_k)\right\|_2 - \lambda \sum_{i=1}^{n} \left\|W^T x_i\right\|_2 . \tag{21}$$

There have two variables need to be optimized in problem (21), thus an efficient iterative optimization algorithm is proposed in what follows.

When W is fixed, the problem (21) is reduced to solve following problem in each class

$$\min_{m_k} \sum_{x_i\in\pi_k} \left\|W^T(x_i-m_k)\right\|_2 . \tag{22}$$

Here, we use the re-weighted method to solve problem (22), and above problem can be rewritten as

$$\min_{m_k} \sum_{x_i\in\pi_k} p_{ik} \left\|W^T(x_i-m_k)\right\|_2^2, \tag{23}$$

where $p_{ik} = \frac{1}{2\|W^T(x_i-m_k)\|_2}$ is the weights. Taking the partial derivative of problem (23), and setting it to zero, we can obtain

$$\sum_{x_i\in\pi_k} p_{ik}(WW^Tm_k - WW^Tx_i) = 0. \tag{24}$$

Supposing $h_k = \sum_{x_i\in\pi_k} p_{ik}(x_i - m_k)$, we can have $WW^Th_k = 0$, and $\{h_k | w_i^T h_k^i = 0, (i = 1, 2, ..., m)\}$ is actually an orthogonal complements of W which can be generalized to $h_k = \alpha_1 W + \alpha_2 \overline{W}$. Substituting h_k into $WW^Th_k = 0$, we can get

$$\alpha_1 WW^TW + \alpha_2 WW^T\overline{W} = 0. \tag{25}$$

Obviously, the second term of Eq.(25) is 0, and $WW^TW \neq 0$ is always set up, thus the value of α_1 can only be equal to 0. Simplify, it is easy to obtain $h_k = \sum_{x_i\in\pi_k} p_{ik}(m_k - x_i) = \alpha_2\overline{W}$, combining with $WW^Th_k = 0$, we can get $\alpha_2 WW^T\overline{W} = 0$. Fortunately, $WW^T\overline{W} = 0$ is always set up, namely, α_2 can be set to arbitrary value. For simplicity, supposing $\alpha_2 = 0$, then $\sum_{x_i\in\pi_k} p_{ik}(m_k - x_i) = 0$, and the problem (23) has following closed-form solution

$$m_k = \frac{\sum_{x_i\in\pi_k} p_{ik}x_i}{\sum_{x_i\in\pi_k} p_{ik}}, \tag{26}$$

which is a weighted class mean, and the p_{ik} denotes the weight of i-th sample in k-th class. Intuitively, if the data point x_i in k-th class is an outlier point, the weight p_{ik} will be small, then the outlier point has less effect on the class mean. After several iterations, the weighted class mean obtained by Eq.(26) will reach to the optimal weighted mean which is an important factor in terms of model robustness improvement.

When m_k is fixed, the problem (21) can be reduced to solve

$$\min_{W^TW=I} \sum_{k=1}^{c} \sum_{x_i\in\pi_k} \left\|W^T(x_i-m_k)\right\|_2 - \lambda \sum_{i=1}^{n} \left\|W^T x_i\right\|_2 , \tag{27}$$

where the second term is actually a $\ell_{2,1}$-norm maximization problem [31]. According to the third step of Algorithm 1, problem (27) can be written as

$$\min_{W^TW=I} \sum_{k=1}^{c} \sum_{x_i\in\pi_k} p_{ik} \left\|W^T(x_i-m_k)\right\|_2^2 - \lambda \sum_{i=1}^{n} \mu_i^T W^T x_i, \tag{28}$$

where the weights $p_{ik} = \frac{1}{2\|W^T(x_i-m_k)\|_2}$ and the vector $\mu_i \in \mathbb{R}^{m\times 1}$ is defined as:

$$\mu_i = \begin{cases} \frac{W^Tx_i}{\|W^Tx_i\|_2} & if \quad \left\|W^Tx_i\right\|_2 \neq 0, \\ \mathbf{0} & if \quad \left\|W^Tx_i\right\|_2 = 0. \end{cases} \tag{29}$$

Therefore, the matrix form of problem (28) is

$$\min_{W^TW=I} Tr(W^TAW) - 2Tr(W^TB), \tag{30}$$

where positive semi-definite symmetric matrix $A = \sum_{k=1}^{c} \sum_{x_i\in\pi_k} p_{ik}(x_i - m_k)(x_i - m_k)^T$ is the weighted within-class scatter matrix, and matrix $B = \frac{\lambda}{2}\sum_{i=1}^{n} x_i\mu_i^T \in \mathbb{R}^{d\times m}$.

It is well-known that above problem (30) is the Quadratic Problem on Stiefel Manifold (QPSM) [43]. Fortunately, a iteration optimization algorithm, named Generalized Power Iteration method (GPI) [44] can solve above QPSM and obtain a local optimum.

According to the above analysis, the detailed iterative non-greedy optimization algorithm to solve the robust $\ell_{2,1}$-norm LDA problem (10) can be summarized in Algorithm 2. Additionally, the proof of convergence of Algorithm 2 can be easily derived according to the convergent proof of Algorithm 1, and the convergence condition of Algorithm 2 used in our experiments is $|\lambda^t - \lambda^{t+1}| \leq 10^{-4}$.

Algorithm 2 An efficient non-greedy iterative re-weighted optimization algorithm to solve problem in Eq.(10).

Input: Data: $X \in \mathbb{R}^{d\times n}$ where X is centralized, label: $Y \in \mathbb{R}^{n\times 1}$, reduced dimension: m
Initialization: $W^1 \in \mathbb{R}^{d\times m}$ such that $W^TW = I$, $p_{ik}^1 = 1$, $t = 1$.
repeat
1. Calculate m_k^t according to Eq.(26) by using p_{ik}^t.
2. Calculate $\lambda^t = \frac{\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|(W^t)^T(x_i-m_k^t)\right\|_2}{\sum_{i=1}^{n}\left\|(W^t)^Tx_i\right\|_2}$.
3. Calculate μ_i^t according to Eq.(29) with W^t.
4. Calculate $A^t = \sum_{k=1}^{c}\sum_{x_i\in\pi_k} p_{ik}^t(x_i - m_k^t)(x_i - m_k^t)^T$ and $B^t = \frac{\lambda^t}{2}\sum_{i=1}^{n} x_i(\mu_i^t)^T$
5. Update W^{t+1} by using GPI with A^t and B^t.
6. Update $p_{ik}^{t+1} = \frac{1}{2\left\|(W^{t+1})^T(x_i-m_k^t)\right\|_2}$.
7. $t = t + 1$.
until convergence
Output: $W^* \in \mathbb{R}^{d\times m}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

4.4 Discussion

In this section, we provide an interesting discussion, i.e., *"whether original LDA using the optimal weighted mean m_k^* in Eq.(26) can generate robust projections?"* Before answering this question, we first introduce Lemma 2:

Lemma 2 *For $\forall\ 0 < r < 1$, solving the r-th root loss problem $\min_x \sum_{i=1}^n f_i(x)^r$ is equivalent to solving an adaptively weighted problem $\min_{x,\alpha} \sum_{i=1}^n \alpha_i f_i(x)$, where $\alpha_i = \frac{r}{f_i(x)^{(1-r)}}$.*

The proofs of Lemma 2 are shown in [45]. According to Lemma 2, we can infer that if the optimal solution α^* is given, the original r-th root loss problem can be viewed as general problem $\min_x \sum_{i=1}^n \alpha_i^* f_i(x)$ where the root issue has been removed. In fact, proposed problem in Eq.(10) is a special case of $\frac{1}{2}$-th root loss problem when the optimal weighted mean is given. In other words, according to Lemma 2, proposed method is to solve ℓ_2-norm trace ratio problem which essentially is LDA in Eq.(4) when m_k^* is given. Therefore, the answer of the question mentioned in the beginning of this section is *True*. Besides, two groups of experiments conducted on real-world datasets in Section 5.5.4 verify the correctness of our conclusion.

It is worth noting that although proposed optimal weighted mean m_k^* plays an important role in improving the robustness of model, it is unknown before model training. In other words, excluding manual intervention, we have no idea about which samples are outliers, and the weights p can not be obtained in advance. Proposed method is capable of learning the weights p directly from data itself, which is driven by the designed root loss function and thanks to the proposed non-greedy iterative re-weighted optimization algorithm.

5 Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of proposed $\ell_{2,1}$-LDA compared to conventional LDA [8], ℓ_1-LDA [30], LDA-ℓ_1 [27], ℓ_1-DML and ILDA-ℓ_1 [29] and RLDA [38] on synthetic data and several gray or color image datasets.

5.1 Synthetic experiment

In this synthetic experiment, we evaluate the robustness of proposed method with comparison to conventional LDA and ℓ_1-LDA. In Fig. 2, we create two classes 2D points specified by blue and red dot respectively, and each class contains 100 samples. Besides, for comparing the robustness, we also construct another four outliers as the training data which are placed in the lower left of Figure. Then, we learn the projections using $\ell_{2,1}$-LDA (green line), LDA (cyan dashed), and ℓ_1-LDA (black dot line). Obviously, the projections generated by LDA are severely deviated from X-axis, which indicates that LDA is absolutely unable to handle outliers. Moreover, the ℓ_1-LDA's projections slightly deviated from optimal projections. Whereas, the projections learnt by proposed method are closer to optimal projections than ℓ_1-LDA, which indicates that proposed $\ell_{2,1}$-LDA is more robust to outliers than ℓ_1-LDA.

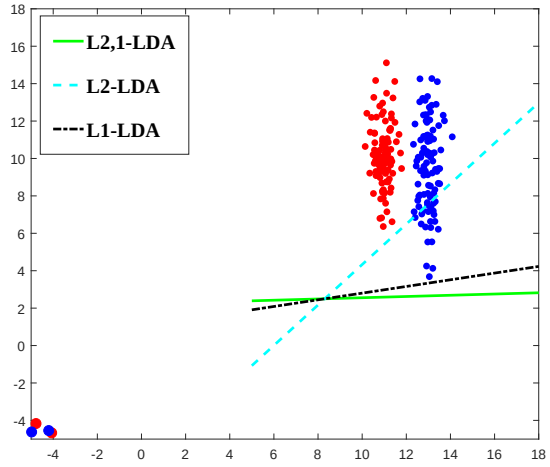


Fig. 2: Projections learned by proposed $\ell_{2,1}$-LDA, ℓ_2-LDA, and ℓ_1-LDA on synthetic dataset with outliers.

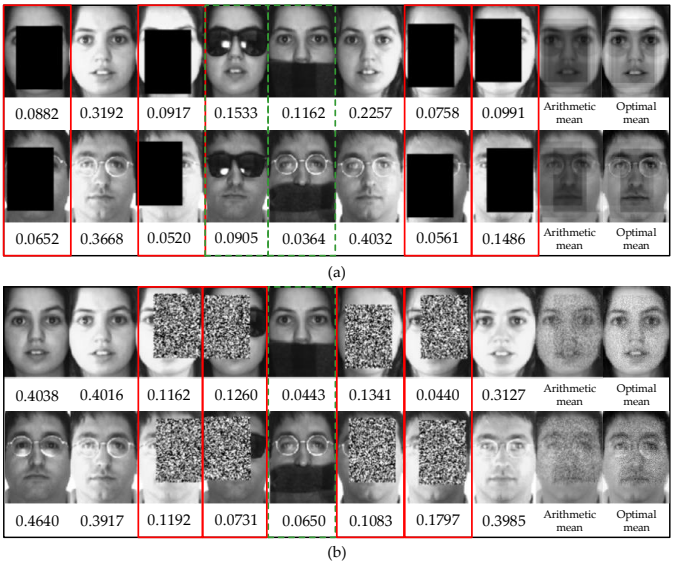


Fig. 3: Visualization of class arithmetic mean sample and optimal weighted mean generated by proposed method on AR dataset, (a). occluded by black block, (b). occluded by salt and pepper block.

5.2 Visualization of optimal weighted mean

In this section, in order to demonstrate the superiority of proposed model in aspect of optimal weighted mean, we present an visualization experiment conducted on AR dataset. In Fig. 3, we randomly choose 8 images in each class for training, then randomly add black block occlusion (a) and salt and pepper occlusion (b) in training images. The first eight columns in Fig. 3 are the training images, the ninth column denotes the arithmetic mean sample and the last column exhibits the optimal weighted mean sample generated by proposed method. Besides, the weights learned by our method are shown below the images as well. It is evident that the weights of occluded images (framed by a red rectangle) and the faces with sunglass or scarf (framed by a green dotted rectangle) are much smaller than the weights of normal face images. Intuitively, our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

optimal weighted mean samples can recover more details such as facial area and myopic lens than arithmetic mean samples. It is demonstrated that our optimal weighted mean mechanism authentically generate positive effects on the estimation of class mean.

5.3 t-SNE visualization on MNIST digits

Mapping high-dimensional data samples onto 2-D subspace is an intuitive way of analyzing the structure of learned subspace. Therefore, to analyze the property of robustness to the outlier sample in proposed method, we employ an unsupervised dimensionality reduction method t-SNE [46] technique to visualize the 2-D mappings of learned low-dimensional data. For data preparation, we choose a standard benchmark MNIST handwritten digits database [47], and in order to keep class balance, we randomly select 200 samples in each class for training model. Moreover, to simulate the circumstance of robust learning, we randomly choose 20 samples as outlier points in each class and add black block occlusions on them with different locations. In training stage, we first employ some related SOTA robust dimensionality reduction methods (introduced in Section 5.4.2) and proposed $\ell_{2,1}$-LDA to mapping those data into different subspace with highest classification accuracy. Then, t-SNE algorithm is used to map those projected data points into 2-D scatterplot for visualization. This is necessary because that since the optimal reduced dimension of all methods are different, it is unfair to directly evaluate the performance of dimensionality reduction algorithms uniformly in 2-D subspace.

Fig. 4 shows the 2-D scatterplot obtained by t-SNE algorithm conducted on low-dimensional MNIST digital numbers learned by using different robust dimension reduction algorithms and proposed method. In the context of visualization, "**outlier isolation**" and "**class separation**" are two important measurements for evaluating the quality of mappings in low-dimensional subspace. Concretely, in terms of "**outlier isolation**" criterion, we use "dashed box" to mark outliers in Fig. 4a-4f, it is clear that compared to other methods, the outlier samples stay further away from normal ones when they lie on the 2-D subspace generated by LDA-ℓ_1 and proposed $\ell_{2,1}$-LDA. It demonstrates their superiority on robustness to outliers than other competitors. Additionally, in Fig. 5 we show some image samples from MNIST digits (number "4" and "9") that look very similar and easily be misclassified. That is, the arithmetic mean of these two classes stay close, which causes that these two classes samples often overlapped in subspace because samples within same class have a tendency to be close to class mean. Intuitively, we use "solid box" to mark the completely overlapped classes in Fig. 4a, 4d, and inapparent overlap occurred in other competitors as well, which results from they calculate arithmetic mean in each class and it is prone to be influenced by outlier samples. In contrast, in Fig. 4f, the subspace generated by proposed $\ell_{2,1}$-LDA is capable of discriminative power that can separate those two class well, which is thanks to the optimal weighted mean introduced in Section 5.2. Above experimental analysis demonstrate the superiority on "**class separation**" of proposed method compared to other related approaches.

5.4 Robust classification on gray image datasets

In this section, we evaluate the robustness of proposed method on several real-world gray image datasets with comparison to six competitors. We randomly select 50% of all samples for training and the rest of samples for testing. For convenience, we downsample each image to suitable size with original scale, and PCA algorithm is used for maintaining 98% principle information of each data in all of our experiments. The k-nearest neighbor (kNN) classifier is used for classification. All of experiments are repeated 10 times, the average recognition rate and standard deviation are recorded as the final performance.

5.4.1 Data preparation and experimental setups

Eight gray face datasets are conducted on our experiments, including PIX[2], JAFFE[2], Yale[2], GTdb[3], YaleB [48], ORL [49], AR[4] and PIE [50] wherein the PIE data used in our experiments is a subset that named POSE27 and contains 3329 samples. The description of used datasets are shown in Table 1. To evaluate the robustness against outliers of proposed method and compared methods, we randomly select some samples in training set as outliers which randomly occluded by fixed-sized black block and salt and pepper occlusion (shown in Fig. 6). To test the robustness of all methods under different level of outliers, we consider selecting different number of images as outliers in each class. Table 2 shows the robust classification performance, i.e., maximal average recognition rate and standard deviations with optimal dimensions on different number of outliers per class in training. Wherein, symbol O_1 and O_2 denote the number of outliers occluded by black block and salt and pepper occlusion in training set per class respectively.

TABLE 1: Descriptions of gray image datasets.

Data	# of Samples	Features	Classes
PIX	100	100×100	10
Yale	165	243×320	15
JAFFE	200	128×128	10
ORL	400	32×32	40
GTdb	750	180×120	50
YaleB	2414	32×32	38
AR	2600	165×120	100
PIE	11554	64×64	68

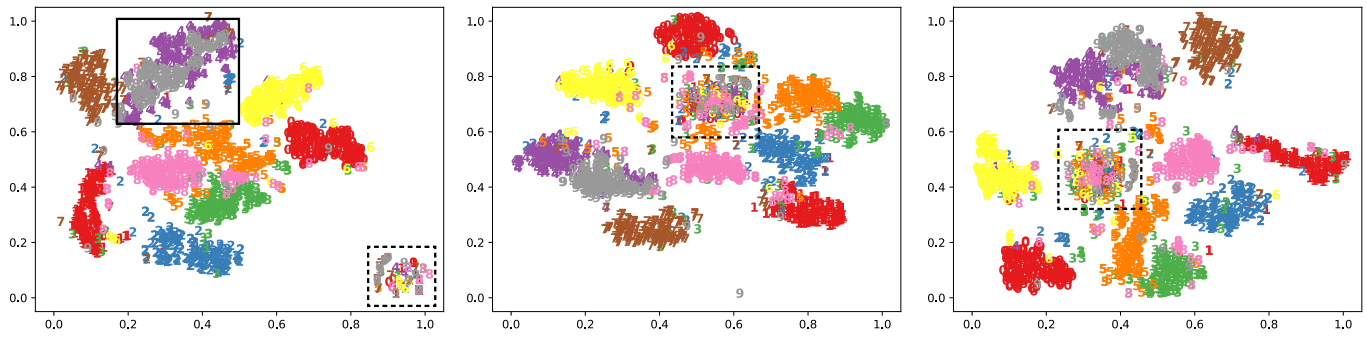
5.4.2 Compared methods and parameters setting

In our experiments, we use 5-fold cross-validation to tune the parameters in following competitors and record the best average recognition accuracy and standard deviation with optimal reduced dimensions on all methods.

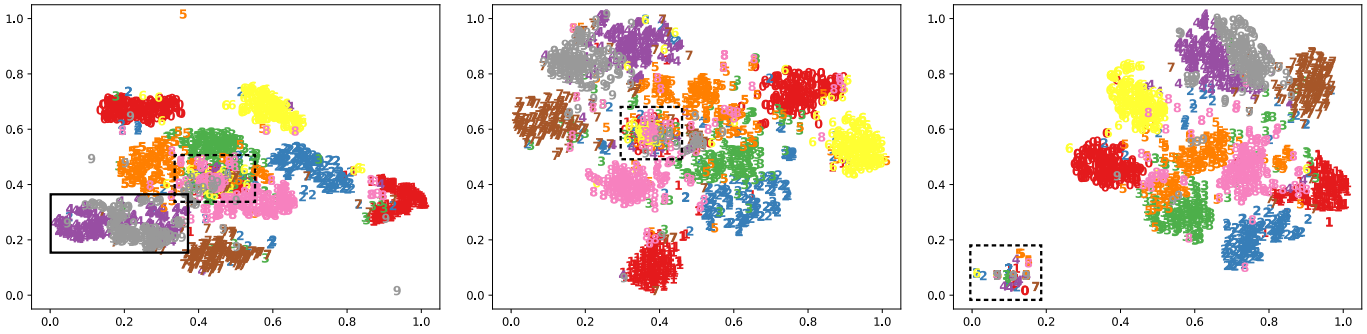
- **LDA** [8]: For fair comparison, the reduced dimensions of LDA is set to $c-1$, and the latent *SSS* problem can be removed by using PCA preprocessing.
- **LDA-ℓ_1** [27]: The parameters in LDA-ℓ_1, i.e., learning rate β and speed of convergence are set as default value 10^{-2}, 10^{-4} respectively.

2. http://www.escience.cn/system/file?fileId=82035
3. http://www.anefian.com/research/face_reco.htm
4. http://www.face-rec.org/databases/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) LDA-ℓ_1 (b) ℓ_1-LDA (c) ℓ_1-DML



(d) ILDA-ℓ_1 (e) RLDA (f) proposed $\ell_{2,1}$-LDA

Fig. 4: t-SNE 2-D mappings visualization on MNIST digits after dimension reduction by using (a) LDA-ℓ_1, (b) ℓ_1-LDA, (c) ℓ_1-DML, (d) ILDA-ℓ_1, (e) RLDA and (f) proposed $\ell_{2,1}$-LDA, respectively.

Fig. 5: Misclassified MNIST images, the first row is digital number "4" and and the second row is digital number "9" respectively.

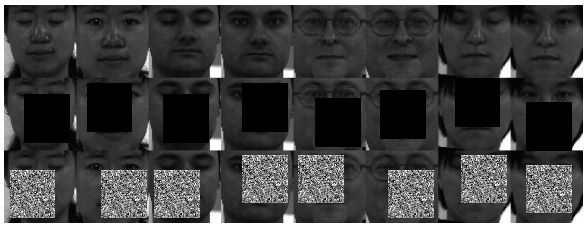


Fig. 6: Some original (first row) and outlier image samples with black block occlusion (second row) and salt and pepper occlusion (third row) in PIE database.

- ℓ_1**-LDA** [30]: This method solves the problem by using projected subgradient method with Armigo line search, and the general Armijo step size is set to 10^{-4} as usual.
- ℓ_1**-DML** [28]: It is a robust Mahalanobis distance metric learning method which can be transformed to robust dimension reduction.
- **RLDA** [38]: As suggested in paper, the initialization of weights d_{ik} for each sample per class are set to 1.
- **ILDA-**ℓ_1 [29]: The parameter λ used to balance the within-class compactness and between-class separability and it is select by grid search in $[0, 1]$.

5.4.3 *Experimental results analysis*

Here, we compare the results of proposed method with above SOTA methods. From the experimental results shown in Table 2, we have following observations:

- First, generally speaking, in comparison of LDA and other robust approaches, it must be pointed out that as the number of outliers increases, the performance of LDA decreases drastically, meanwhile robust algorithms outperform LDA in most cases. These basic findings verify that $\ell_{2,1}$-norm and ℓ_1-norm are more robust to outliers than ℓ_2-norm, which is consistent with previous conclusion.
- Second, regrading the limitations of greedy search optimization algorithm, the performance of LDA-ℓ_1 seriously fall behind other methods on Yale, YaleB and AR datasets in terms of maximal average recognition rate. This findings are directly in line with previous conclusion that greedy search optimization algorithm is prone to get stuck in local optimum. On the contrary, the competitors based on non-greedy optimization algorithm such as ℓ_1-LDA, RLDA and proposed method achieve stable performance on all datasets. As a result, the non-greedy optimization algorithm is beneficial to classification task.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

- Third, proposed $\ell_{2,1}$-LDA obtains over 2.2%, 4.4%, 5.06% and 4.3% higher accuracies than RLDA in JAFFE, PIX, Yale and ORL datasets respectively with various outlier number, which indicates that proposed $\ell_{2,1}$-norm minmax objective is capable of producing more discriminative projections than simplex minimization objective function.
- Last, our method always obtain around 2.0%, 3.46% and 2.36% higher accuracies than the second highest results produced by ℓ_1-LDA, ℓ_1-DML and ILDA-ℓ_1 on PIX, Yale and YaleB datasets respectively, when the number of outliers reaches the maximum. The results demonstrate two facts: 1) $\ell_{2,1}$-norm possesses more robustness than ℓ_1-norm, when the number of outliers is large. 2) The optimal weighted mean produced by proposed method is beneficial to estimate true data distribution as well as improve recognition rate. Last one, it is worth noting that our method always outperforms original LDA even though there are no occluded images in training set, which is because the data itself is not pure enough and contains outlier samples possibly, meanwhile our proposed minmax objective function and corresponding solver facilitate the solution be equipped with more discriminative power.

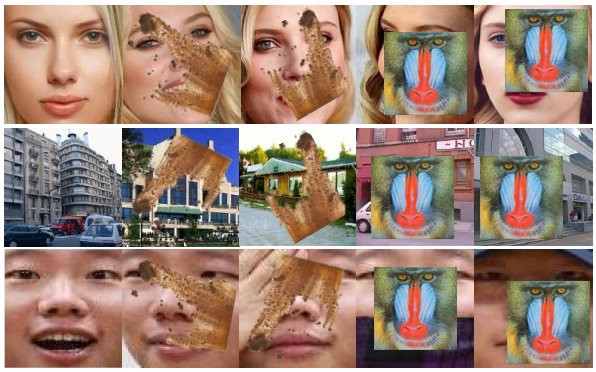


Fig. 7: Sample images of Pubfig, OSR and NUST-RF datasets (from top to bottom) with normal type, mud and baboon occlusions (from left to right) respectively.

5.5 Robust classification on RGB real-world image sets

We evaluate our model on another two groups of occlusion experiments conducted on RGB real-world star face and outdoor scenes image sets for validating the robustness of proposed model. Some normal and outlier examples occluded by baboon and mud in PubFig, OSR and NUST-RF datasets are shown in Fig. 7. The experimental results histogram are shown in Fig. 8 and Fig. 9 where O_3 and O_4 denote the number of outlier images occluded by baboon and mud in each class of training set respectively.

5.5.1 Data preparation and experimental setups

- **PubFig** - We use a subset of public figure face image set named PubFig [51], [52] which contain 772 images from 8 face categories, and each image is a 256×256 RGB image. We extract 12580 LOMO features [53] from each image and PCA is further applied to reduce the feature dimension to 100. We randomly choose 20% and 70% images in each class for training where several images are picked as outliers by adding baboon and mud occlusions randomly. The experiments are repeated 20 times, and the best average recognition rates are recorded.
- **OSR** [54] - This dataset contains 2688 RGB images with 8 outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. Each image has 256×256 pixel, we extract 512 dimension gist features [55] for data representation. PCA preprocessing is applied to preserve 98% major information of each data. We randomly choose 10% and 50% of all samples for training and the rest of samples for testing. Other settings are the same to the settings in PubFig dataset.
- **NUST-RF** - NUST Robust Face database consists of various face images under different occlusions [56]. Except occlusion, it also includes variations of illumination, expression and pose. We use two subsets face images of NUST-RF database, which contains 30 subjects captured in two environments (indoor and outdoor) respectively. We manually cropped the face portion of the image, then normalized it to 80×60 pixels and 3,740 LOMO features are extracted for data representation finally.

5.5.2 Experimental results analysis on recognition rate

Fig. 8 and Fig. 9 show the best average recognition accuracy of all competitors on PubFig and OSR datasets respectively with various training sample size and two different kinds of occlusions. From these results, we can conclude following observations:

- Although the performance of all methods have improved to some extent when the number of training samples increases, the gaps of performance between different methods are not varying dramatically, which demonstrates the authenticity of experimental results.
- It is clear that LDA and LDA-ℓ_1 perform not well on Pugfig dataset, because the metric based on ℓ_2-norm and greedy search optimization algorithm are not conducive to improve robust classification recognition rate, and those results are directly in line with previous conclusion in section 5.4.3.
- From Fig. 8b, d and Fig. 9b, d, we can observe that our method always obtain 1.11%, 3.89%, 2.82% and 1.78% higher accuracies than the second highest results produced by ℓ_1-LDA, ℓ_1-DML and ILDA-ℓ_1 on Pugfig and OSR datasets with various training sample size respectively, when the number of outliers in training set reaches to 30. This observation indicates that proposed $\ell_{2,1}$-LDA is more powerful for handling the challenge that the number of outliers in training set is large.

5.5.3 Experimental results analysis on ROC curves

It is well-known that Receiver Operating Characteristics (ROC) is an effective way for evaluating different methods in terms of robust pattern recognition task so as to measure the accuracy of outlier rejection [57]. Experiments are performed on a robust face database, named NUST-RF, and we randomly select 50% samples for training and remaining data for testing. For evaluating the property of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2: Robust classification results on gray image sets (best average accuracy ± standard deviations%) with optimal dimensions on different number of outliers per class in training.

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
JAFFE	O_1 = 0	96.80±1.94	98.40±2.45	99.60±0.80	99.80±0.40	97.40±1.02	98.00±0.40	**100.00±0.00**
	O_1 = 2	71.60±3.06	96.40±1.02	98.20±0.75	98.40±0.49	96.80±1.08	97.60±0.80	**99.00±0.63**
	O_1 = 4	52.60±2.01	90.00±2.53	97.60±1.10	98.00±0.80	95.20±1.44	95.80±1.67	**98.80±0.75**
	O_2 = 0	96.40±1.85	99.60±0.49	**100.00±0.00**	99.80±0.40	97.40±1.20	99.20±1.17	**100.00±0.00**
	O_2 = 2	86.00±2.94	99.40±0.40	99.60±0.49	97.80±1.47	94.20±2.28	98.80±1.85	**99.80±0.40**
	O_2 = 4	74.20±3.10	98.60±1.55	98.40±0.80	96.80±2.79	93.00±1.19	98.20±2.61	**99.60±0.49**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
PIX	O_1 = 0	89.60±4.80	97.60±2.33	98.40±1.96	98.00±2.33	90.00±3.10	99.00±1.79	**99.20±0.98**
	O_1 = 1	86.40±6.12	96.40±3.71	95.20±4.31	94.40±3.25	88.67±0.78	93.60±4.27	**97.20±0.44**
	O_1 = 2	74.00±4.45	94.00±2.19	94.00±2.04	94.80±2.71	87.49±0.67	92.00±1.26	**96.80±2.40**
	O_2 = 0	87.60±4.63	97.20±2.04	97.80±2.33	97.60±2.33	92.40±3.88	98.40±1.96	**98.80±1.60**
	O_2 = 1	85.20±4.12	95.20±1.60	96.20±2.71	94.40±2.04	90.20±1.04	95.20±1.50	**97.60±1.50**
	O_2 = 2	70.00±6.88	93.80±2.19	93.40±2.65	92.40±3.35	89.60±2.18	92.00±2.33	**94.00±2.71**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
Yale	O_1 = 0	81.07±4.66	79.20±2.75	86.93±3.92	86.13±3.33	82.40±3.11	85.60±2.53	**87.73±2.13**
	O_1 = 1	77.87±5.24	74.67±3.86	84.00±4.10	83.73±3.44	81.33±3.71	71.20±1.81	**87.20±2.72**
	O_1 = 2	73.60±4.57	69.87±2.47	81.60±4.38	81.87±4.43	80.27±4.01	64.00±3.73	**85.33±3.04**
	O_2 = 0	80.80±3.44	73.87±3.44	86.47±2.87	85.07±4.49	81.07±2.72	86.13±3.20	**86.67±0.84**
	O_2 = 1	77.60±4.51	74.13±2.99	81.60±4.01	82.40±3.86	77.33±2.00	83.20±2.75	**84.37±2.23**
	O_2 = 2	71.73±3.99	70.67±3.33	80.20±2.75	81.60±2.44	72.00±2.39	80.00±2.99	**82.60±2.85**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
GTdb	O_1 = 0	70.63±1.10	77.23±1.72	**81.89±1.08**	81.51±1.03	72.09±1.26	80.83±1.16	81.83±0.95
	O_1 = 1	68.91±1.65	74.26±1.64	**80.29±1.05**	77.71±1.10	70.91±1.58	69.20±1.93	79.69±1.16
	O_1 = 2	68.17±1.65	71.31±1.47	74.69±1.51	74.40±1.68	69.83±1.40	67.71±1.98	**76.60±1.62**
	O_1 = 3	65.80±2.00	69.17±1.79	72.69±2.43	73.29±2.17	66.54±1.90	65.37±2.20	**75.03±1.58**
	O_2 = 0	71.83±0.85	77.74±1.70	**82.91±1.13**	81.17±1.10	72.89±1.26	81.31±1.50	82.89±0.93
	O_2 = 1	68.86±1.44	75.83±1.17	80.17±0.81	78.71±1.15	69.77±1.57	79.66±0.90	**81.40±0.97**
	O_2 = 2	66.77±1.43	73.66±1.73	78.34±1.58	78.03±1.59	66.74±1.27	77.66±1.41	**79.69±1.93**
	O_2 = 3	64.31±1.54	70.11±1.91	76.26±2.31	**76.37±1.70**	64.29±1.54	73.46±3.63	76.09±1.66

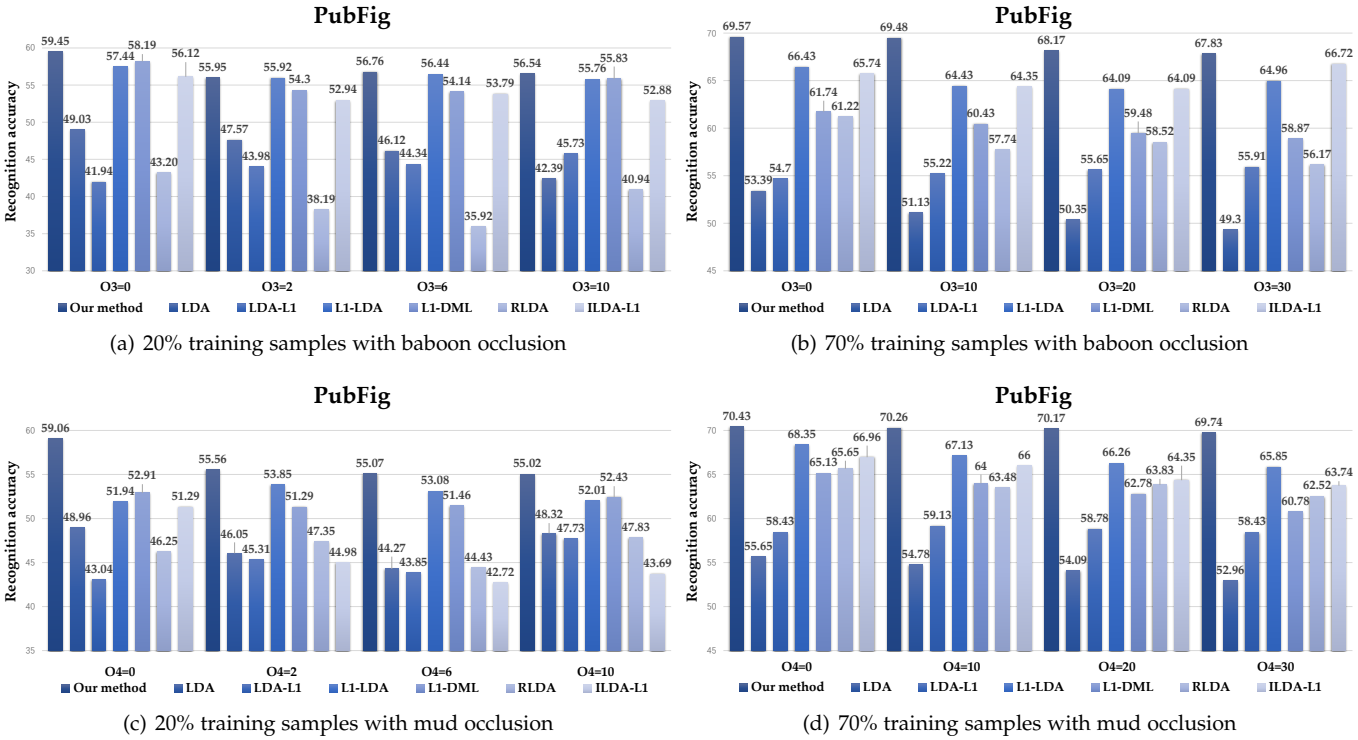
Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
YaleB	O_1 = 0	87.92±0.35	55.05±0.65	91.34±0.57	91.37±0.40	92.27±0.45	85.94±1.62	**93.20±0.23**
	O_1 = 4	85.69±0.65	53.69±0.61	90.31±0.17	90.21±0.51	90.97±0.83	85.98±1.35	**92.30±0.69**
	O_1 = 6	85.06±0.35	51.72±0.61	89.86±0.29	89.66±0.24	90.71±0.47	82.36±0.76	**91.78±0.37**
	O_1 = 8	84.76±0.84	50.95±0.69	90.12±0.33	89.49±0.25	90.21±0.55	81.73±0.46	**91.65±0.26**
	O_1 = 10	83.92±0.55	49.06±0.99	89.38±0.52	89.24±0.54	89.69±0.43	80.00±1.26	**91.60±0.51**
	O_2 = 0	87.32±0.44	56.13±0.96	92.20±0.39	91.57±0.45	92.63±0.51	86.44±1.12	**93.61±0.10**
	O_2 = 4	84.35±0.59	52.88±1.16	91.35±0.54	91.02±0.35	91.57±0.37	86.57±1.24	**92.98±0.33**
	O_2 = 6	84.50±0.45	51.83±0.67	89.99±0.34	90.66±0.38	90.92±0.47	86.31±0.39	**92.61±0.35**
	O_2 = 8	83.07±0.43	50.39±1.01	89.59±0.58	89.23±0.49	90.09±0.67	85.26±1.33	**91.47±0.51**
	O_2 = 10	82.16±0.30	48.46±0.41	88.91±0.63	88.46±0.42	89.54±0.58	84.23±1.38	**90.67±0.41**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
ORL	O_1 = 0	87.70±2.11	94.80±1.03	94.90±0.40	96.00±0.80	91.50±0.93	96.30±0.84	**96.70±0.80**
	O_1 = 1	86.90±1.83	93.60±0.87	95.80±0.66	95.70±0.80	91.20±1.25	95.10±0.87	**96.30±0.51**
	O_1 = 2	87.80±1.39	93.90±1.21	95.60±0.81	95.70±0.89	92.10±2.10	94.80±1.02	**96.40±0.80**
	O_1 = 3	87.70±1.16	94.08±0.86	95.30±0.49	95.60±0.81	91.30±1.03	95.00±1.05	**96.10±0.58**
	O_2 = 0	88.80±0.75	95.10±0.71	96.90±0.75	97.20±0.68	92.00±1.53	96.80±1.05	**97.70±0.68**
	O_2 = 1	89.20±1.56	95.60±0.97	96.40±0.63	96.60±0.86	92.20±1.44	96.20±0.68	**97.50±0.84**
	O_2 = 2	87.00±0.87	95.00±0.81	96.50±0.84	96.80±0.68	91.40±0.86	96.20±0.87	**97.20±0.84**
	O_2 = 3	89.20±1.03	95.40±0.86	95.20±0.51	96.00±0.84	91.60±0.71	96.10±1.07	**97.20±0.95**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
AR	O_1 = 0	83.35±0.40	58.02±0.48	90.77±0.51	88.25±0.50	91.52±0.43	76.92±0.72	**92.46±0.30**
	O_1 = 2	78.98±0.47	48.57±0.85	87.75±0.79	84.09±0.70	89.00±0.33	52.80±1.49	**89.69±0.36**
	O_1 = 4	74.97±0.49	49.15±0.80	82.89±0.93	77.62±0.74	81.92±0.18	42.35±1.53	**84.43±0.62**
	O_1 = 6	70.05±1.21	40.46±0.96	79.06±1.31	73.72±1.18	69.75±0.91	39.18±1.20	**79.71±1.66**
	O_2 = 0	83.15±0.39	54.57±0.85	91.06±0.22	89.08±0.16	92.48±0.40	78.29±1.77	**92.77±0.23**
	O_2 = 2	80.40±0.09	54.80±1.08	88.69±0.32	85.49±0.05	**91.12±0.16**	75.92±0.07	90.62±0.37
	O_2 = 4	77.42±0.23	52.11±0.50	86.60±0.32	81.85±0.34	**88.49±0.47**	69.46±0.69	88.22±0.05
	O_2 = 6	74.42±0.33	44.02±0.74	83.42±1.02	78.35±1.05	84.62±0.37	64.69±0.63	**85.17±1.04**

Data	Occlusion	LDA	LDA-ℓ_1	ℓ_1-LDA	ℓ_1-DML	RLDA	ILDA-ℓ_1	$\ell_{2,1}$-LDA
PIE	O_1 = 0	96.19±0.17	93.62±0.61	98.06±0.09	97.87±0.14	98.12±0.13	96.37±0.44	**98.53±0.09**
	O_1 = 4	95.71±0.18	92.52±0.70	97.85±0.13	97.47±0.09	98.03±0.18	95.32±0.37	**98.20±0.13**
	O_1 = 6	95.65±0.13	91.89±0.75	97.49±0.09	97.24±0.09	97.84±0.16	94.86±0.18	**97.92±0.08**
	O_1 = 8	94.97±0.29	90.42±1.06	97.46±0.18	96.90±0.19	97.61±0.22	93.50±0.73	**97.72±0.12**
	O_1 = 10	94.92±0.07	89.97±1.03	97.23±0.14	96.71±0.22	97.56±0.17	91.50±0.64	**97.65±0.14**
	O_2 = 0	95.89±0.19	94.56±0.45	97.96±0.18	97.88±0.26	98.05±0.15	96.59±0.57	**98.28±0.21**
	O_2 = 4	96.15±0.16	94.06±0.51	97.63±0.41	97.39±0.33	**97.96±0.13**	95.71±0.57	97.83±0.44
	O_2 = 6	96.60±0.20	90.53±0.80	97.46±0.26	97.42±0.43	97.90±0.13	95.16±0.52	**98.00±0.17**
	O_2 = 8	94.52±0.29	91.50±0.67	96.63±0.18	96.31±0.42	**97.50±0.38**	95.19±0.92	97.35±0.37
	O_2 = 10	94.30±0.20	92.08±0.74	96.63±0.46	96.22±0.59	**97.08±0.23**	94.20±0.81	96.95±0.47

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence



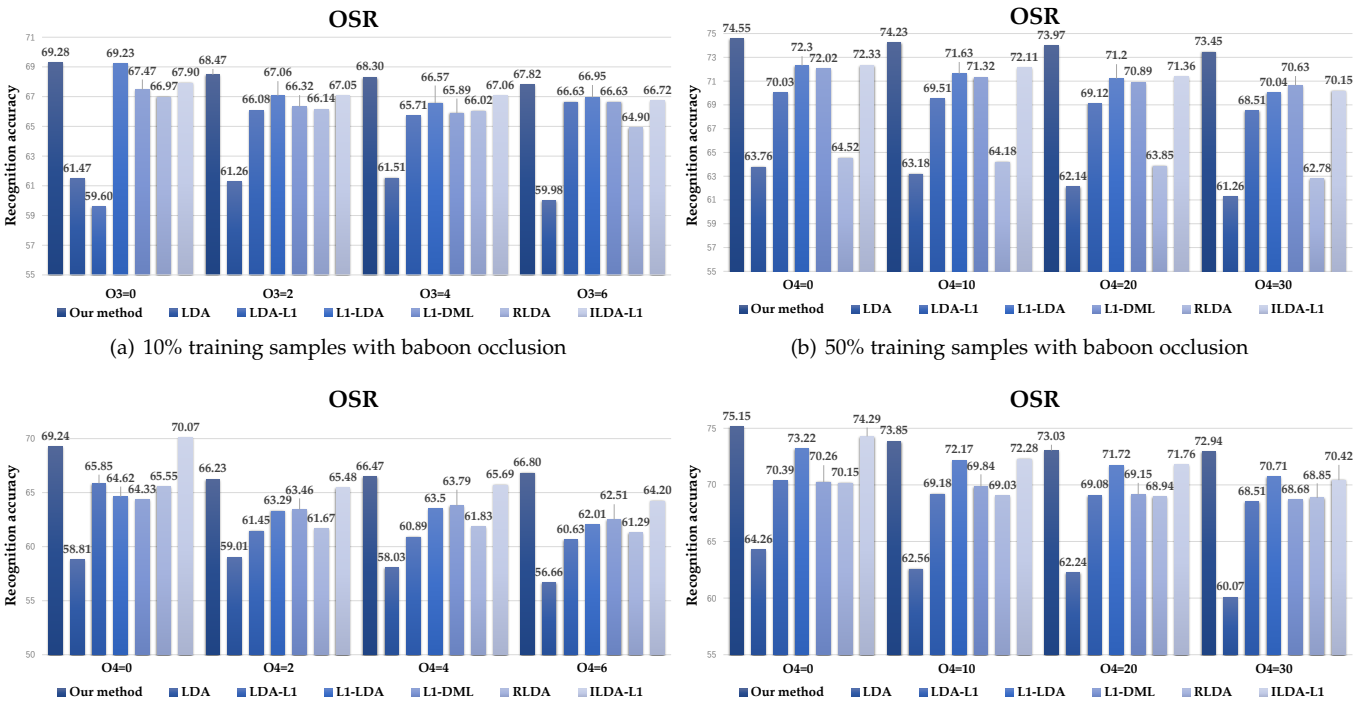
(a) 20% training samples with baboon occlusion

(b) 70% training samples with baboon occlusion

(c) 20% training samples with mud occlusion

(d) 70% training samples with mud occlusion

Fig. 8: The best average recognition accuracy (%) of all competitors with different number of training samples and occlusions on PubFig dataset.



(a) 10% training samples with baboon occlusion

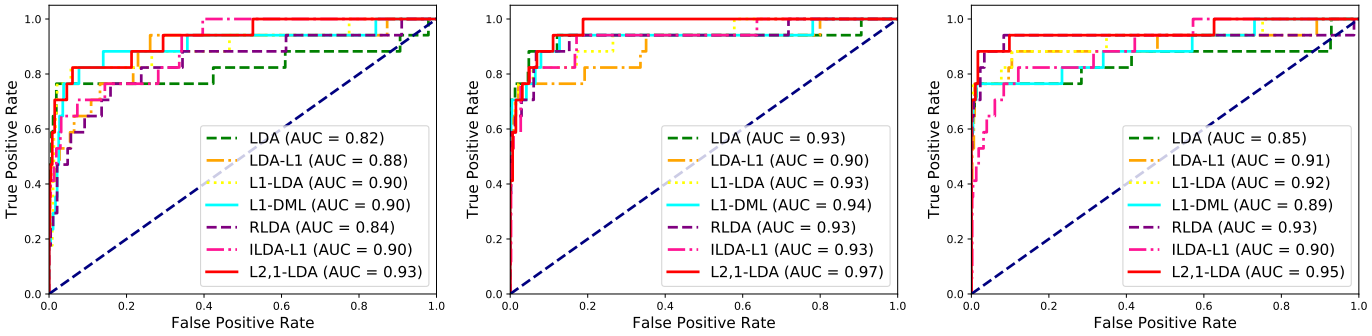
(b) 50% training samples with baboon occlusion

(c) 10% training samples with mud occlusion

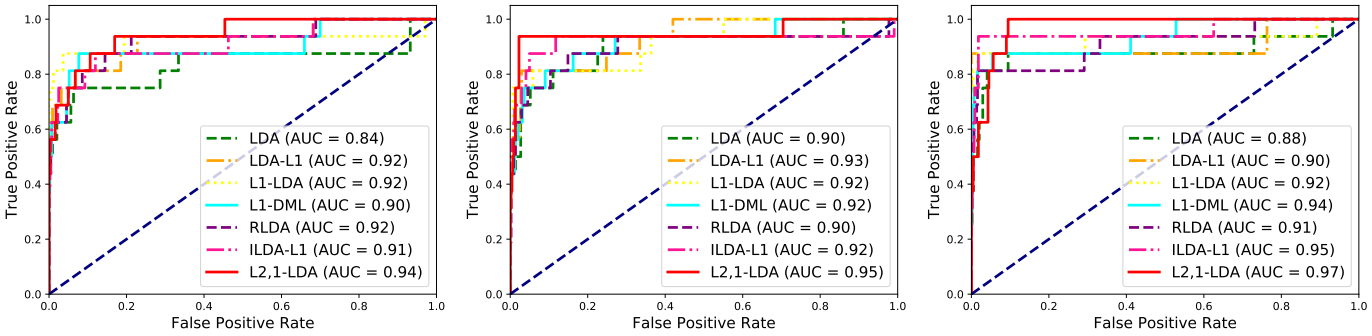
(d) 50% training samples with mud occlusion

Fig. 9: The best average recognition accuracy (%) of all competitors with different number of training samples and occlusions on OSR dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Indoor data with black block, baboon and mud occlusions (from left to right).



(b) Outdoor data with black block, baboon and mud occlusions (from left to right).

Fig. 10: Receiver operating characteristics (ROC) of proposed $\ell_{2,1}$-LDA in comparison to the ROC of the other SOTA methods on NUST-RF dataset including indoor (a) and outdoor (b), respectively.

robustness to outliers in proposed method, 10 images that added three types of occlusions are randomly selected in training set as outliers. Then, we seek best low-dimensional representation of all data samples in terms of recognition rate via kNN classifier by using aforementioned dimension reduction algorithms. However, the prediction label information still can not be obtained directly so far. Thus, we resort Linear Regression model (LR) to achieve label probability prediction matrix of testing data. A threshold is used to produce a multiclass classifier and each threshold value generates a different point in ROC curve.

As illustrated in Fig. 10, experimental results analysis of ROC curves can be expanded from following aspects: 1) All ROC curves generated by using learned low-dimensional data outperform diagonal line (blue dash line), which demonstrates that the classifier results are obviously better than the results produced by strategy of randomly guessing a class. 2) In general, a good algorithm should achieve high true positive rate (TPR) at a low false positive rate (FPR), that is, the larger the area under ROC curve (AUC value), the more discriminative information the data contains. Obviously, proposed method achieves best AUC value compared to other competitors based on ℓ_1-norm, which caused by ℓ_1-norm is not invariant to rotation and the performance is suboptimal as well. 3) Our method achieves superior and stable performance in the context of diverse occlusions that is because of that $\ell_{2,1}$-norm can achieve the desirable robustness compared to ℓ_1-norm.

5.5.4 *Further Applications*

In the previous Section 4.4, we have theoretically verified that combining optimal weighted mean m_k^* with original LDA can improve the robustness to outliers of model. In this subsection, we provide two groups of experiments that respectively conducted on gray and RGB image datasets to verify our conclusion in practice. Classification results are shown in Table 3, wherein RT denotes the LDA model used is the ratio trace formulation in Eq.(2), and TR(1), TR(2) represent trace ratio LDA in Eq.(3) with different optimization algorithms that proposed in [41] and [58] respectively. Experimental settings are the same as the previous section. From the experimental results, we can conclude two facts: 1) Three types of conventional LDA combined optimal weighted mean obtain around 5% and 2% higher accuracies than LDA in Table 2 on AR and YaleB datasets respectively. Moreover, they all achieve comparable performance against robust LDA based on ℓ_1-norm shown in Fig. 7 and Fig. 8 when training set contains outliers. This fact demonstrates the effectiveness of proposed optimal weighted mean on robustness to outliers. 2) Our method still outperforms these LDA using optimal weighted mean mechanism in most cases, which is attributed to proposed minmax objective function and non-greedy iterative re-weighted optimization algorithm that can find more discriminative projections.

6 Conclusion

In this paper, we propose a novel LDA model based on $\ell_{2,1}$-norm for robust dimensionality reduction. Different from most existing related algorithms, proposed model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2961877, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3: Experimental results of ratio trace (RT) and trace ratio (TR) LDA with optimal weighted mean.

Dataset	YaleB			PIE		
	RT	TR(1)	TR(2)	RT	TR(1)	TR(2)
$O_1 = 4$	91.10	91.05	91.19	97.28	97.46	97.63
$O_1 = 6$	90.16	90.51	90.31	96.74	96.89	97.18
$O_1 = 8$	89.78	88.76	88.16	95.88	96.04	96.09
$O_2 = 4$	92.61	90.87	91.20	97.92†	96.87	97.00
$O_2 = 6$	90.71	91.54	90.95	97.06	96.93	96.87
$O_2 = 8$	91.62†	91.36	89.47	96.93	96.26	96.32

Dataset	Pubfig			OSR		
	RT	TR(1)	TR(2)	RT	TR(1)	TR(2)
$O_3 = 10$	67.39	68.70	66.96	70.90	69.98	69.48
$O_3 = 20$	66.52	65.22	66.09	70.16	69.73	68.73
$O_3 = 30$	64.35	63.91	64.78	69.03	69.11	68.36
$O_4 = 10$	70.43†	70.00	69.57	69.73	69.60	69.98
$O_4 = 20$	68.26	68.70	69.13	68.86	68.36	68.61
$O_4 = 30$	67.96	67.83	67.39	67.95	67.99	68.49

† denotes the accuracy outperforms proposed method.

needs to minmax $\ell_{2,1}$-norm simultaneously, which is never been solved ideally. For solving proposed non-convex $\ell_{2,1}$-norm minmax problem, we first design an iterative optimization algorithm to solve general minimization ratio problem with rigorous proof of convergence. Furthermore, an efficient non-greedy iterative re-weighted optimization algorithm is derived to solve proposed $\ell_{2,1}$-norm minmax problem. Experiments conducted on several real-world datasets illustrate the superiority of proposed method over the other SOTA competitors. In the future works, we will expand our algorithm to solve more general minmax problem such as ℓ_p-LDA or ℓ_p-PCA, etc.

7 Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the National Natural Science Foundation of China under Grant 61772427 and Grant 61751202, and in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501.

References

[1] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng, "Discriminant analysis of principal components for face recognition," in *Face Recognition*. Springer, 1998, pp. 73–85.

[2] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, 2012.

[3] H. Wang, F. Nie, H. Huang, S. Risacher, A. J. Saykin, L. Shen *et al.*, "Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2011, pp. 115–123.

[4] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080–2088.

[7] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," Yale University New Haven United States, Tech. Rep., 1997.

[9] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[10] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise lda for solving the small sample size problem in face verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 506–519, 2007.

[11] F. Nie, S. Xiang, and C. Zhang, "Neighborhood minmax projections." in *IJCAI*, 2007, pp. 993–998.

[12] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2100–2110, 2017.

[13] Y. Zhang and D.-Y. Yeung, "Worst-case linear discriminant analysis," in *Advances in Neural Information Processing Systems*, 2010, pp. 2568–2576.

[14] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu, "Fast and orthogonal locality preserving projections for dimensionality reduction," *IEEE Trans. Image Process*, vol. 26, no. 10, pp. 5019–5030, 2017.

[15] C. Croux and C. Dehon, "Robust linear discriminant analysis using s-estimators," *Canadian Journal of Statistics*, vol. 29, no. 3, pp. 473–493, 2001.

[16] J. Yang, D. Zhang, and J.-y. Yang, "Median lda: a robust feature extraction method for face recognition," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, vol. 5. IEEE, 2006, pp. 4208–4213.

[17] X.-j. Wang, "Modular pca based on within-class median for face recognition," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 1. IEEE, 2010, pp. 52–56.

[18] E. Zyad, C. Khalid, and B. Mohammed, "Combination of r1-pca and median lda for anomaly network detection," in *2017 Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2017, pp. 1–5.

[19] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse pca by l1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[20] J. Gao, "Robust l1 principal component analysis and its bayesian variational inference," *Neural computation*, vol. 20, no. 2, pp. 555–572, 2008.

[21] N. Kwak, "Principal component analysis based on l1-norm maximization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.

[22] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy l1-norm maximization," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1433.

[23] P. P. Markopoulos, S. Kundu, S. Chamadia, and D. A. Pados, "Efficient l1-norm principal-component analysis via bit flipping," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4252–4264, 2017.

[24] R. Wang, F. Nie, X. Yang, F. Gao, and M. Yao, "Robust 2dpca with non-greedy ell_1-norm maximization for image analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 1108–1112, May 2015.

[25] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with l1-norm," *IEEE transactions on cybernetics*, vol. 44, no. 6, pp. 828–842, 2014.

[26] W. Zheng, Z. Lin, and H. Wang, "L1-norm kernel discriminant analysis via bayes error bound optimization for robust feature extraction," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 4, pp. 793–805, 2014.

[27] F. Zhong and J. Zhang, "Linear discriminant analysis based on l1-norm maximization," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3018–3027, 2013.

[28] H. Wang, F. Nie, and H. Huang, "Robust distance metric learning via simultaneous l1-norm minimization and maximization," in *International Conference on Machine Learning*, 2014, pp. 1836–1844.

[29] X. Chen, J. Yang, and Z. Jin, "An improved linear discriminant analysis with l1-norm for robust feature extraction," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1585–1590.

[30] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A non-greedy algorithm for l1-norm lda," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 684–695, 2017.

[31] F. Nie and H. Huang, "Non-greedy l21-norm maximization for principal component analysis," *arXiv preprint arXiv:1603.08293*, 2016.

[32] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, "Lp-and ls-norm distance based robust linear discriminant analysis," *Neural Networks*, 2018.

[33] C. Ding, D. Zhou, X. He, and H. Zha, "R1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288.

[34] H. Huang and C. Ding, "Robust tensor factorization using r1 norm," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[35] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[36] C.-X. Ren, D.-Q. Dai, and H. Yan, "Robust classification using 2, 1-norm based regression model," *Pattern Recognition*, vol. 45, no. 7, pp. 2708–2718, 2012.

[37] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *International conference on machine learning*, 2014, pp. 1062–1070.

[38] H. Zhao, Z. Wang, and F. Nie, "A new formulation of linear discriminant analysis for robust dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[39] X. Shi, F. Nie, Z. Lai, and Z. Guo, "Robust principal component analysis via optimal mean by joint 2, 1 and schatten p-norms minimization," *Neurocomputing*, vol. 283, pp. 205–213, 2018.

[40] R. Zhang, F. Nie, and X. Li, "Auto-weighted two-dimensional principal component analysis with robust outliers," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6065–6069.

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

[42] K. C. Border, Supergradients, Caltech Div. Humanities Social Sci., California Inst. Technol., Pasadena, CA, USA, 2001, pp. 116.

[43] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, no. Feb, pp. 517–553, 2010.

[44] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the stiefel manifold," *Science China Information Sciences*, vol. 60, no. 11, p. 112101, 2017.

[45] F. Nie, J. Li, X. Li *et al.*, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification." in *IJCAI*, 2016, pp. 1881–1887.

[46] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[47] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[48] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2691–2698.

[49] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[50] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 53–58.

[51] J. Xu, L. Luo, C. Deng, and H. Huang, "Bilevel distance metric learning for robust image recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 4198–4207.

[52] D. Parikh and K. Grauman, "Relative attributes," in *IEEE International Conference on Computer Vision (ICCV)*, Nov 2011.

[53] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

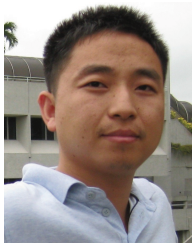
[54] J. Xu, L. Luo, C. Deng, and H. Huang, "New robust metric learning model using maximum correntropy criterion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2555–2564.

[55] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[56] S. Chen, J. Yang, L. Luo, Y. Wei, K. Zhang, and Y. Tai, "Low-rank latent pattern approximation with applications to robust image classification," *IEEE transactions on image processing*, vol. 26, no. 11, pp. 5519–5530, 2017.

[57] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2010.

[58] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.



Feiping Nie Feiping Nie received the Ph.D. degree in Computer Science from Tsinghua University, China in 2009, and currently is full professor in Northwestern Polytechnical University, China. His research interests are machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing and information retrieval. He has published more than 100 papers in the following journals and conferences: TPAMI, IJCV, TIP, TNNLS, TKDE, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, ACM MM. His papers have been cited more than 10000 times and the H-index is 57. He is now serving as Associate Editor or PC member for several prestigious journals and conferences in the related fields.

Zheng Wang is currently pursuing the Ph.D. degree at School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University. His research interests include dimensionality reduction, deep cross-modal learning, medical image processing and so on.

Rong Wang received the B.S. degree in information engineering, the M.S. degree in signal and information processing, and the Ph.D. degree in computer science from Xian Research Institute of Hi-Tech, Xi'an, China, in 2004, 2007 and 2013, respectively. During 2007 and 2013, he also studied in the Department of Automation, Tsinghua University, Beijing, China for his Ph.D. degree. He is currently an associate professor at the School of Cybersecurity and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests focus on machine learning and its applications.

Zhen Wang received the Ph.D. degree from Hong Kong Baptist University, Hong Kong, in 2014. From 2014 to 2016, he was a JSPS Senior Researcher with the Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka, Japan. Since 2017, he has been a Full Professor with Northwestern Polytechnical University, Xi'an, China. Thus far, he has published more than 100 scientific papers and obtained over 9400 citations. His current research interests include network science, complex system, big data, evolutionary game theory, behavior decision, and behavior recognition. He was a recipient of the National 1000 Talent Plan Program of China. He serves as an editor or an academic editor for seven journals.

Xuelong Li (M'02-SM'07-F'12) is a full professor with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P.R. China.