ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/325459717

A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction

Article *in* IEEE Transactions on Knowledge and Data Engineering · May 2018

DOI: 10.1109/TKDE.2018.2842023

CITATIONS
7

READS
261

3 authors, including:

Zheng Wang
Northwestern Polytechnical University
12 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Feiping Nie
University of Texas at Arlington
402 PUBLICATIONS **11,001** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Video Understanding View project

Hyperspectral Images Clustering View project

All content following this page was uploaded by Zheng Wang on 17 July 2018.

The user has requested enhancement of the downloaded file.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction

Haifeng Zhao, Zheng Wang, and Feiping Nie

Abstract—Dimensionality reduction is a critical technology in the domain of pattern recognition, and linear discriminant analysis (LDA) is one of the most popular supervised dimensionality reduction methods. However, whenever its distance criterion of objective function uses L_2-norm, it is sensitive to outliers. In this paper, we propose a new formulation of linear discriminant analysis via joint $L_{2,1}$-norm minimization on objective function to induce robustness, so as to efficiently alleviate the influence of outliers and improve the robustness of proposed method. An efficient iterative algorithm is proposed to solve the optimization problem and proved to be convergent. Extensive experiments are performed on an artificial data set, on UCI data sets and on four face data sets, which sufficiently demonstrates the efficiency of comparing to other methods and robustness to outliers of our approach.

Index Terms—Robust linear discriminant analysis, dimensionality reduction, $L_{2,1}$-norm minimization

1 Introduction

DIMENSIONALITY reduction is a critical technology to solve the problem of curse of dimensionality [1], whose goal is to map high-dimensional data into lower-dimensional subspace without losing discriminant information. Numerous dimensionality reduction approaches have been proposed in the past several decades. Taking the class labels of samples into account or not, dimensionality reduction algorithms will fall into three categories: unsupervised algorithms [2], semi-supervised algorithms [3] and supervised algorithms [4]. Among them, Principal Component Analysis (PCA) is a classical unsupervised dimensionality reduction method which learns a projection matrix such that the variance of low-dimensional data is maximized. The most well-known disadvantage of unsupervised algorithms is that the label information is not used in the task of classification. Since the learning data sets are labeled, it makes sense to use labeled information to create more efficient methods, namely supervised dimensionality reduction on many applications such as face recognition [5], handprint classification [6], text classification [7] and so on [8]. R.A.FISHER *et al*. propose a discriminant analysis based on Fisher criterion (FDA) [4] which maximizes between-class scatter and minimizes within-class scatter. Yet, FDA is only suitable for two-class classification tasks, while most of classification problems are multi-class classification in the real world. Therefore, Rao C R *et al.* propose an extension of FDA, i.e., LDA [9] to tackle multi-class issue.

However, LDA suffers from several drawbacks, wherein, the first drawback is that conventional LDA is incompetent to deal with multi-modal data whose distribution is more complex than Gaussian. To overcome this issue, F. Nie *et al.* present a pairwise formulation of LDA, namely Neighborhood MinMax Projections (NMMP) [10], which attempts to pull the considered pairwise points within the same class as close as possible and push those between different classes separate. In this way, the points within same class are no longer gathered to class mean point, but close to each other, which is sensitive to local structure of data manifold. Additionally, most of dimensionality reduction methods involve solving Trace Ratio (TR) problem which does not have a closed-form global optimum solution as usual. Y. Jia *et al.* [11] develop a Decomposed Newton's Method (DNM) to efficiently find the global optimum of the Trace Ratio (TR) problem and prove the foundation and superiority of the new approach by the theoretical analysis. Furthermore, LDA also requires sufficient train data to avoid the Small Sample Size problem (SSS) [12], which makes it difficult to deal with the small-scale date with high dimensionality. To overcome the SSS problem, numerous extensions of LDA have been proposed, such as Pseudo-inverse LDA [13], Two-stage LDA [14], Regularized LDA (RLDA) [15], Orthogonal LDA (OLDA) [16], Null space LDA (NLDA) [17], Direct LDA (DLDA) [18], Maximum Margin Criterion (MMC) [19], Angle Linear Discriminant Embedding (ALDE) [20] and so on.

Pseudo-inverse LDA [13] exploits a positive pseudo-inverse matrix $\boldsymbol{S}_{\boldsymbol{w}}^{\dagger}$ instead of calculating the matrix $\boldsymbol{S}_{\boldsymbol{w}}^{-1}$. The two-stage LDA [14] utilizes PCA to reduce the dimension of original data to m ($m = rank(\boldsymbol{S}_{\boldsymbol{t}})$, where $\boldsymbol{S}_{\boldsymbol{t}}$ is the total-class scatter matrix), then the generated within-class scatter matrix $\boldsymbol{S}_{\boldsymbol{w}}$ is nonsingular and its regarding inverse matrix can be obtained. However, some discriminative information may be discarded in the PCA stage as well. For the purpose of avoiding the defect of two-stage LDA, Null

- *Haifeng Zhao is with Key Lab of Intelligent Computing and Signal Processing of MOE & School of Computer and Technology, Anhui University, HeFei 230039, P.R.China.*
  E-mail: senith@163.com
- *Zheng Wang is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shanxi, P.R.China.*
  E-mail: zhengwangml@gmail.com
- *Feiping Nie is the corresponding author, he is with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shanxi, P.R.China.*
  E-mail: feipingnie@gmail.com

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

space LDA (NLDA) [17] calculates the projections according to the null space of within-class scatter matrix $\boldsymbol{S_w}$, in which some discriminant information are always preserved. Direct LDA (DLDA) [18] tackles the SSS problem by discarding the null space of between-class scatter matrix $\boldsymbol{S_b}$, which is non-information. Additionally, some other methods consider using new optimization criterion to overcome the SSS problem. For instance, Ye J [16] presents a generalized discriminant analysis based on a new optimization criterion, namely Orthogonal LDA which obtains orthogonal projections so as to preserve more discriminant data structure [21], even though the scatter matrix is singular. Li H *et al.* [19] define a maximum margin criterion as discriminant analysis criterion to maximize the distance between any two different classes and the final matrix to be decomposed is $\boldsymbol{S_b}-\boldsymbol{S_w}$ rather than $(\boldsymbol{S_w})^{-1}\boldsymbol{S_b}$. Therefore, it is not necessary to consider whether $\boldsymbol{S_w}$ is singular or not. Recently, Liu S *et al.* [20] propose a method, namely angle linear discriminant embedding, using the cosine of angle to redefine within and between scatter matrices then it can deal with the SSS problem as well.

Another major drawback of LDA is that both LDA and its variants are sensitive to outliers. The reason of this sensitivity is that the L_2-norm is used as distance criterion of objective function, and it magnifies the influence of outliers in the objective function. Taking into consideration the results shown in [22], we point out L_1-norm is more robust to outliers than L_2-norm. Therefore, many robust dimensionality reduction methods via joint L_1-norm have been proposed in the past few years whose pioneered works include R_1-PCA [23], L_1-PCA [24], L_1-LDA [25] etc. Ding C *et al.* [23] propose a rotational invariance L_1-norm PCA (R_1-PCA) which can effectively suppress the outliers while keeping the property of being rotational invariant. However, it is stuck in the defect of high computational complexity. In order to reduce the computational complexity, a robust principal component analysis based on L_1-norm (L_1-PCA [24]) is proposed, and it solves the L_1-norm maximization optimization through an efficient and easy to implement algorithm. Nonetheless, the projections are calculated one by one with a greedy search strategy in which it is easy to get stuck in a local optimum. To overcome this problem, a non-greedy search strategy method has been proposed in [26], in which all projections are optimized simultaneously. Zhong F *et al.* [25] propose another version of L_1-LDA and theoretically prove that it is robust to outliers. However, it is difficult to obtain a global optimal solution and it only obtains an approximate solution by a greedy search method. Actually, the motivation of ALDE [20] is that the optimization model based on angle is similar to L_1-norm model, consequently, ALDE also possesses the property of robustness. All aforementioned methods concentrate on the improvement of the objective function and neglect some limitations about data. In real applications, only a few samples per class are available for training as usual. As a result, it is difficult to precisely estimate the class mean by using the class average sample, especially, in the case that the train samples include outliers. Consequently, Yang J *et al.* [27] propose a median LDA (MLDA) which uses the class median sample instead of the class mean to improve the robustness of model.

As mentioned in [28], a novel robust feature selection method (RFS) via joint the $L_{2,1}$-norm on both loss function and regularization has been proposed, whose main task is to solve the difficulty of being robust to outliers. In addition, an efficient algorithm to solve the $L_{2,1}$-norm optimization problem is presented with proved convergence. The most similar work to RFS is RRC (Rotational invariant norm based Regression for Classification) [29] which uses the $L_{2,1}$-norm regression model to obtain the class labels for classifying test samples directly. In this paper, motivated by [28], we propose a new formulation of linear discriminant analysis for robust dimensionality reduction (RLDA) which uses the $L_{2,1}$-norm to replace L_2-norm based on the LDA model intuitively. As a consequence, the residual of objective function is not squared with the result that the outliers have less effect than the squared residual.

The main contributions of this paper are summarized as follows.

1) A novel and straightforward robust dimensionality reduction algorithm via joint $L_{2,1}$-norm with linear discriminant analysis is proposed to suppress the influence of outliers for improving the robustness of the model.
2) The class central point is no longer class mean point, but class weighted mean point in our model which can avoid the influence of outliers.
3) An efficient iterative algorithm is presented to solve the $L_{2,1}$-norm minimization optimization problem, so as to adaptively assign weights that equipped all points with physical meaning.

Additionally, compared with the previous robust linear discriminant analysis model, our method has the following differences.

1) The objective function is different from previous works. Specifically, L_1-LDA [25] solves the L_1-norm maximization optimization problem rather than $L_{2,1}$-norm minimization optimization problem. Robust feature selection (RFS) [28] and RRC [29] are established on regression model rather than linear discriminant analysis model.
2) The objective function is more reasonable and suitable to the treatment of data with outliers which sheds light on the robustness of our method.
3) The proposed efficient iterative algorithm can naturally and effectively solve the $L_{2,1}$-norm minimization optimization problem without high computational complexity.

In our experiments, we evaluate the performance of the proposed method, comparing methods that have become quite popular in the field of dimensionality reduction, namely LDA, MMC, NMMP, Trace Ratio and ALDE on both artificial toy data set, UCI data sets and four face data sets. Experimental results demonstrate the robustness to outliers of our approach. Besides, we also evaluate the effect of different initializations of the solution, as a consequence, the experimental results show that the objective function always converges to constant values. Hence, the solution of the proposed iterative algorithm is a global optimum even though the optimization problem is a non-convex problem.

The remainder of this paper is organized as follows: in Section 2, some related works are introduced. The proposed

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

$L_{2,1}$-norm LDA is presented and the convergence of the algorithm is proved in Section 3. The experimental results are described in Section 4. Lastly, Section 5 gives the conclusion.

2 RELATED WORK

2.1 Notations and Definitions

We summarize some notations and definitions of norms used in this paper. We denote vector and matrix by boldface lowercase letters and boldface uppercase letters respectively. The $L_{2,1}$-norm of vector $\mathbf{u} \in \mathbb{R}^n$ is defined as: $||\mathbf{u}||_{2,1} = \sqrt{\sum_{i=1}^{n} u_i^2}$ and the $L_{2,1}$-norm of matrix $\mathbf{U} \in \mathbb{R}^{n\times m}$ is defined as:

$$||\mathbf{U}||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} u_{ij}^2} \tag{1}$$

where the i-th element of vector $\mathbf{u}$ is denoted by u_i and the (i,j)-th element of matrix $\mathbf{U}$ is denoted as u_{ij} as well. Besides, the Frobenius norm of vector $\mathbf{u}$ is denoted by $||\mathbf{u}||_F$ which is equal to $||\mathbf{u}||_2^2$.

2.2 A Brief Review of Linear Discriminant Analysis

Given the data matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where d and n denote the dimensionality of the original space and the number of samples respectively, the goal of LDA is to learn a linear transformation matrix $\mathbf{W} \in \mathbb{R}^{d\times m} (m \ll d)$ to map the high-dimensional data $\mathbf{x} \in \mathbb{R}^d$ into a low-dimensional data $\mathbf{y} \in \mathbb{R}^m$ by:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{2}$$

$\mathbf{X}$ is grouped as $\mathbf{X} = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 ... \boldsymbol{\pi}_c]$, where c denotes the number of classes, $\boldsymbol{\pi}_i \in \mathbb{R}^{d\times n_i}$ denotes the data set of class i and n_i denotes the number of data samples in class i.

In order to get discriminant information, the optimal projection matrix can be obtained by solving the following problem [12]:

$$\min Tr((\mathbf{W}^T\boldsymbol{S_t}\mathbf{W})^{-1}(\mathbf{W}^T\boldsymbol{S_w}\mathbf{W})) \tag{3}$$

where the $Tr(\cdot)$ denotes the trace of matrix, the within-class scatter matrix is defined by $\boldsymbol{S_w} = \sum_{i=1}^{c}\sum_{\mathbf{x}\in\boldsymbol{\pi}_i}(\mathbf{x}-\bar{\mathbf{x}}_i)(\mathbf{x}-\bar{\mathbf{x}}_i)^T$ (where $\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{\mathbf{x}_j\in\boldsymbol{\pi}_i}\mathbf{x}_j$ denotes the mean of the samples in class i) and the total-class scatter matrix $\boldsymbol{S_t} = \sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$ (where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ denotes the mean of all the samples). The solution of the objective function can be reduced to solving the following eigenvalues decomposition problem:

$$\boldsymbol{S_t}^{-1}\boldsymbol{S_w}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda} \tag{4}$$

where $\boldsymbol{\Lambda}$ is the eigenvalue matrix of $\boldsymbol{S_t}^{-1}\boldsymbol{S_w}$. Lastly, the transformation matrix $\mathbf{W}$ is composed by of the eigenvectors of $\boldsymbol{S_t}^{-1}\boldsymbol{S_w}$ corresponding to the first m smallest eigenvalues.

2.3 Robust Feature Selection Based on $L_{2,1}$-norm

Given the data sets $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, the class indicator matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathbb{R}^{c\times n}$ is defined as follows: if the i-th ($i = 1, 2, ..., n$) sample belongs to j-th ($j = 1, ..., c$) class, then $\mathbf{y}_i = [0, ..., 0, 1, 0, ..., 0]^T \in \mathbb{R}^c$ contains only the j-th element equal to one. The main task of least square regression is to solve the following optimization problem to obtain the projection matrix $\mathbf{W} \in \mathbb{R}^{d\times c}$ and bias $\mathbf{b} \in \mathbb{R}^c$:

$$\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{y}_i||_2^2. \tag{5}$$

Simply, the objective function can be written as:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{y}_i||_2^2. \tag{6}$$

where the k-th column of $\mathbf{W}$ comprises the $d+1$ dimensional vector $\mathbf{w}_k = (b_k, \mathbf{w}_k^T)^T$ and $\mathbf{x}_i$ is the corresponding augmented vector $(1, \mathbf{x}_i^T)^T$. In the RFS model, the robust objective function is

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{y}_i||_{2,1} \tag{7}$$

where the residual is not squared and in which the outliers have less influence than in the squared residual. Adding a regularization term with a parameter γ, the model (7) develops into a robust feature selection model

$$\min_{\mathbf{W}} \frac{1}{\gamma}||\mathbf{X}^T\mathbf{W} - \mathbf{Y}||_{2,1} + ||\mathbf{W}||_{2,1} \tag{8}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$ is the label matrix. Due to $||\mathbf{X}^T\mathbf{W}-\mathbf{Y}||_{2,1} = ||\mathbf{Y}-\mathbf{X}^T\mathbf{W}||_{2,1}$, and let $\mathbf{E} = \frac{1}{\gamma}(\mathbf{Y}-\mathbf{X}^T\mathbf{W})$, then problem (8) is equivalent to

$$\min_{\mathbf{W},\mathbf{E}} ||\mathbf{E}||_{2,1} + ||\mathbf{W}||_{2,1}. \tag{9}$$

The above problem (9) can be rewritten as

$$\min_{\mathbf{U}} ||\mathbf{U}||_{2,1} \quad s.t. \quad \mathbf{AU} = \mathbf{Y} \tag{10}$$

where $\mathbf{U} = \begin{bmatrix}\mathbf{W}\\ \mathbf{E}\end{bmatrix} \in \mathbb{R}^{m\times c}$ $(m = n + d)$, $\mathbf{A} = \begin{bmatrix}\mathbf{X}^T & \gamma\mathbf{I}\end{bmatrix} \in \mathbb{R}^{n\times m}$ and $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix. The real challenge of this model is to solve the $L_{2,1}$-norm optimization problem in these two terms, and Nie *et al.* [28] propose a simple yet efficient algorithm to solve this problem.

There is another supervised feature selection algorithm, namely Similarity Preserving Feature Selection (SPFS) [30], which has a similar objective function to Eq.(8). Both of them aim at preserving the global similarity structure rather than the local geometric structure of data. The differences between RFS and SPFS are the regression target and regression loss [30]. Subsequently, we will introduce a novel formulation of robust dimensionality reduction method which is based on linear discriminant analysis rather than regression model.

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

3 Proposed Method

3.1 Motivation and Problem Formulation

In this subsection, we formulate the robust LDA based on $L_{2,1}$-norm. From the literature [12], the conventional LDA can be written as following problem:

$$\min_{\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} Tr(\mathbf{W}^T\boldsymbol{S_w}\mathbf{W}) \tag{11}$$

which have following vector form:

$$\min_{\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} \sum_{k=1}^{c}\sum_{\mathbf{x}_i\in\boldsymbol{\pi}_k} \left\|\mathbf{W}^T(\mathbf{x}_i-\boldsymbol{\mu}_k)\right\|_2^2, \tag{12}$$

where $\boldsymbol{\mu}_k$ is the mean of samples in k-th class. Obviously, the objective function of Eq.(12) employs L_2-norm distance criterion to evaluate loss function value, and it is known that squared loss function will enlarge the error if training set contains outliers. In our model, we use $L_{2,1}$-norm instead of L_2-norm which can reduce the influence of outliers in the objective function value. Therefore, our new robust LDA is to solve the following problem:

$$\min_{\mathbf{m}_k,\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} \sum_{k=1}^{c}\sum_{\mathbf{x}_i\in\boldsymbol{\pi}_k} \left|\left|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\right|\right|_2 \tag{13}$$

where the outliers in the residual $\left|\left|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\right|\right|_2$ have less effect than in the square residual $\left|\left|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\right|\right|_2^2$ [28]. In addition, $\mathbf{m}_k$ in Eq.(13) is a weighted class mean which is different from $\boldsymbol{\mu}_k$ in Eq.(12). Specifically, the weights of all samples can be learned adaptively by our model, and the ideal case is that the weights of outliers are very small, which can alleviate the influence of outliers in the calculation of class mean.

An illustration conducted on FERET [31] data set to demonstrate the advantage of weighted class mean is shown in Fig. 1, in which the samples in the first three columns are occluded by the insertion of small black blocks as an attempt to simulate as the outliers in training data. Instead, we keep in the next four columns the normal samples while we place in the eighth and ninth columns the mean of the previous columns, respectively without and with the black blocks. It is evident that the ninth column which represents the role played by $\boldsymbol{\mu}_k$ in Eq.(12) is blurred by outliers and it is difficult to identify which class the sample belongs to. That is why the conventional LDA is sensitive to outliers. On the contrary, let us focus on the last column which represents the weighted centre of the left seven columns, i.e., $\mathbf{m}_k$ in Eq.(13) of our model and the image appears as more distinct and less affected by the presence of outliers which seem as almost eliminated. We should point out that the aforementioned desirable property is attributed to the $L_{2,1}$-norm which is able to make the objective function value more stable and less attracted towards the outliers by imposing that the weights of the outliers are close to zero.

In addition, in Fig. 2, another visualization is presented to show the weights of each sample, i.e., d_{ik} in Eq.(14) that learned by our method adaptively. We conduct this experiment on AR [32] data set which contains 100 classes and every class has 26 samples, therefore each subfigure in Fig. (2) is a 100×26 matrix. Then we modify the data by choosing the first 10%-50% samples and affect them by superimposition of a black area that represents outliers in each class. As shown in Fig. (2), our method assigns the weights to each sample adaptively, and every row of the weight matrix represents the weights of samples in the same class. Obviously, the weights corresponding to the outliers are close to zero, i.e., the black area in each subfigure, which proves the above-mentioned theoretical deduction. In brief, our method is capable of robustness to outliers.

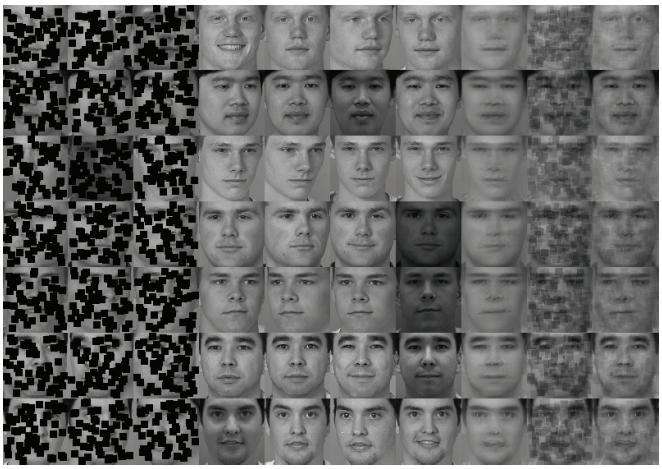

Fig. 1: Class mean sample on occluded FERET face data. The first seven columns are training samples in different classes, including three occluded samples representing the outliers in each class (first three columns). The eighth and ninth columns are the class mean sample without/with outliers respectively. The last column is the weighted class mean sample.

Although the new model has several merits, the problem (13) is difficult to solve because it has two variables need to be optimized. In the next subsection, we propose a simple yet efficient iterative algorithm to solve the problem.

3.2 An Efficient Iterative Algorithm to Solve the Problem (13)

Since the problem (13) is dependent on $\mathbf{W}$ and $\mathbf{m}_k$, it is difficult to solve it directly. In this paper, we use an efficient iterative algorithm to solve it. The detailed algorithm is described in Algorithm 1, and the theoretical analysis of the algorithm is given at the end of this section. In each iteration, we solve the following problem:

$$\min_{\mathbf{m}_k,\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} \sum_{k=1}^{c}\sum_{\mathbf{x}_i\in\boldsymbol{\pi}_k} d_{ik}\left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\right\|_2^2 \tag{14}$$

where $d_{ik} = \frac{1}{2\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\|_2}$ and $\boldsymbol{\pi}_k$ represents the samples in the k-th class.

We alternately update one variable while fixing the other variable each time. First, we fix $\mathbf{W}$ to solve for $\mathbf{m}_k$ and simply optimize problem (14) in k-th class. As a result, the problem (13) becomes the problem (15):

$$\min_{\mathbf{m}_k} \sum_{\mathbf{x}_i\in\boldsymbol{\pi}_k} d_{ik}\left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{m}_k)\right\|_F. \tag{15}$$

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

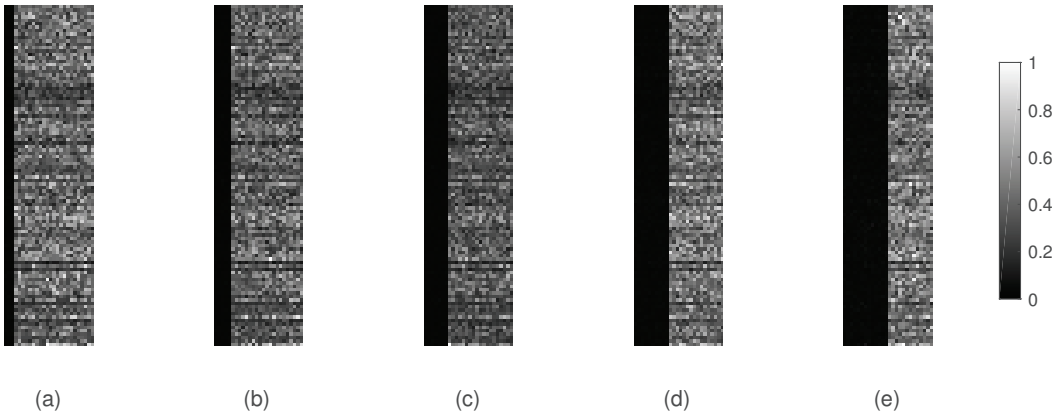

Fig. 2: Weight matrix generated by RLDA on AR data set with 10%(a), 20%(b), 30%(c), 40%(d), 50%(e) outliers respectively.

We assume the weight d_{ik} is a constant which can be initialized. Taking the partial derivative of problem (15) w.r.t. $\mathbf{m}_k$ and setting it to zero, we have:

$$\sum_{\mathbf{w}_i \in \boldsymbol{\pi}_k} d_{ik}(\mathbf{W}\mathbf{W}^T\mathbf{m}_k - \mathbf{W}\mathbf{W}^T\mathbf{x}_i) = 0 \tag{16}$$

$$\Rightarrow \quad \mathbf{W}\mathbf{W}^T\mathbf{r}_k = 0 \tag{17}$$

where $\mathbf{r}_k = \sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}(\mathbf{m}_k - \mathbf{x}_i)$. Then we can let $\mathbf{r}_k = \alpha\mathbf{W} + \beta\overline{\mathbf{W}}$, where $\overline{\mathbf{W}}$ is the orthogonal complement of $\mathbf{W}$. Substituting $\mathbf{r}_k$ into Eq.(17) and we have

$$\alpha\mathbf{W}\mathbf{W}^T\mathbf{W} + \beta\mathbf{W}\mathbf{W}^T\overline{\mathbf{W}} = 0. \tag{18}$$

It is obvious that only setting $\alpha = 0$, the Eq.(18) is possibly set up; then we can get

$$\mathbf{r}_k = \sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}(\mathbf{m}_k - \mathbf{x}_i) = \beta\overline{\mathbf{W}} \tag{19}$$

$$\Rightarrow \quad \mathbf{m}_k = \frac{1}{\sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}}\Big(\sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}\mathbf{x}_i + \beta\overline{\mathbf{W}}\Big). \tag{20}$$

Substituting Eq.(19) into Eq.(17), we can arrive at:

$$\mathbf{W}\mathbf{W}^T\overline{\mathbf{W}}\beta = 0 \tag{21}$$

From Eq.(21), we know that regardless of the numerical value of β, the value of Eq.(21) is always zero. Therefore we can let $\beta = 0$ and arrive at:

$$\mathbf{m}_k = \frac{\sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}\mathbf{x}_i}{\sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}} \tag{22}$$

where $\mathbf{m}_k$ is a weighted class mean which is different from the one usually adopted in the conventional model. Subsequently, we fix $\mathbf{m}_k$ and optimize $\mathbf{W}$, then problem (13) reduces to:

$$\min_{\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} Tr(\mathbf{W}^T\mathbf{M}\mathbf{W}) \tag{23}$$

where $\mathbf{M} = \sum_{k=1}^{c}\sum_{\mathbf{x}_i \in \boldsymbol{\pi}_k} d_{ik}(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T$. Then the Lagrangian function of the problem (23) is:

$$\mathcal{L}(\mathbf{W}, \boldsymbol{\Lambda}) = Tr(\mathbf{W}^T\mathbf{M}\mathbf{W}) - Tr(\boldsymbol{\Lambda}(\mathbf{W}^T\boldsymbol{S_t}\mathbf{W} - \mathbf{I})) \tag{24}$$

where $\boldsymbol{\Lambda}$ is the Lagrangian coefficient matrix. Taking the derivative of Eq.(24) w.r.t. $\mathbf{W}$ to zero, we know that the solution of problem (23) is reduced to solving the following eigen-decomposition problem:

$$\boldsymbol{S_t}^{-1}\mathbf{M}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda}. \tag{25}$$

Finally, the columns in projection matrix $\mathbf{W}$ are formed by the eigenvectors of $\boldsymbol{S_t}^{-1}\mathbf{M}$ corresponding to the first m smallest eigenvalues.

Algorithm 1 that is a summarization of the previous formulations, is described as follows. In each iteration, $\mathbf{m}_k$ is calculated with the current d_{ik}. Then $\mathbf{M}$ is calculated with the current d_{ik} and $\mathbf{m}_k$, the projection matrix $\mathbf{W}$ is updated by current $\mathbf{M}$. Finally, the update of d_{ik} occurs based on the current $\mathbf{W}$. The iterative algorithm procedure is repeated until the algorithm converges. We prove the convergence of our optimization algorithm in the next subsection.

3.3 Theoretical analysis of Algorithm 1

The Algorithm 1 monotonically decreases the objective function of the problem in Eq.(13) in each iteration. To prove it, we first introduce the following lemma [28]:

Lemma 1. *For any nonzero vector $\boldsymbol{u}, \boldsymbol{u}_t \in \mathbb{R}^c$, the following inequality holds:*

$$||\mathbf{u}||_2 - \frac{||\mathbf{u}||_2^2}{2||\mathbf{u}_t||_2} \leq ||\mathbf{u}_t||_2 - \frac{||\mathbf{u}_t||_2^2}{2||\mathbf{u}_t||_2}. \tag{26}$$

The convergence of the Algorithm 1 is described in the following theorem:

Theorem 1 *Algorithm 1 monotonically decreases the objective function of the problem formulated with Eq.(13) in each iteration.*

Proof: The problem (13) can be written as follows:

$$\min_{\mathbf{W}^T\boldsymbol{S_t}\mathbf{W}=\mathbf{I}} Tr\Big[\sum_{k=1}^{c}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m}_k)\frac{1}{2||\mathbf{W}^T(\mathbf{x}_i - \mathbf{m}_k)||_2} (\mathbf{x}_i - \mathbf{m}_k)^T\mathbf{W}\Big]. \tag{27}$$

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

Algorithm 1 An efficient iterative algorithm to solve the problem (13).

Input:
$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$;
$\mathbf{y} = [y_1, ..., y_n] \in \mathbb{R}^n$;
The final dimension: m
Initialitation:
$t = 0$;
$d_{ik}^{(0)} = 1$;
repeat
Calculate $\mathbf{m}_k^{(t+1)} = \dfrac{\sum\limits_{\mathbf{x}_i\in\boldsymbol{\pi}_k} d_{ik}^{(t)}\mathbf{x}_i}{\sum\limits_{\mathbf{x}_i\in\boldsymbol{\pi}_k} d_{ik}^{(t)}}$
Calculate $\mathbf{M}^{(t+1)} = \sum\limits_{k=1}^{c}\sum\limits_{\mathbf{x}_i\in\boldsymbol{\pi}_k} d_{ik}^{(t)}\big(\mathbf{x}_i - \mathbf{m}_k^{(t+1)}\big)\big(\mathbf{x}_i - \mathbf{m}_k^{(t+1)}\big)^T$
Calculate $\mathbf{W}^{(t+1)}$ which is the eigenvectors matrix of $\boldsymbol{S}_t^{-1}\mathbf{M}^{(t+1)}$ corresponding to its m smallest eigenvalues.
Update $d_{ik}^{(t+1)} = \frac{1}{2||\mathbf{W}^{(t+1)T}(\mathbf{x}_i - \mathbf{m}_k^{(t+1)})||_2}$.
$t = t + 1$.
until *Convergence*
Output: $\mathbf{W} \in \mathbb{R}^{d\times m}$

Let us denote $\mathbf{u}$ by $\mathbf{u} = (\mathbf{x}_i - \mathbf{m}_k)^T\mathbf{W}$ and substitute it into Eq.(27). Then we have:

$$\min_{\mathbf{W}^T\boldsymbol{S}_t\mathbf{W}=\mathbf{I}} Tr\bigg(\sum_{k=1}^{c}\mathbf{u}^T\theta\mathbf{u}\bigg) \tag{28}$$

where $\theta = \frac{1}{2||\mathbf{u}^T||_2}$. Thus, in the t-th iteration,

$$\mathbf{u}_{t+1} = \arg\min_{\mathbf{W}^T\boldsymbol{S}_t\mathbf{W}=\mathbf{I}} Tr(\mathbf{u}^T\theta_t\mathbf{u}) \tag{29}$$

$$\Rightarrow Tr(\mathbf{u}_{t+1}^T\theta_t\mathbf{u}_{t+1}) \le Tr(\mathbf{u}_t^T\theta_t\mathbf{u}_t) \tag{30}$$

This is to say,

$$\frac{||\mathbf{u}_{t+1}^k||_2^2}{2||\mathbf{u}_t^k||_2} \le \frac{||\mathbf{u}_t^k||_2^2}{2||\mathbf{u}_t^k||_2}. \tag{31}$$

According to Lemma 1, we have

$$\sum_{k=1}^{c}\bigg(||\mathbf{u}_{t+1}^k||_2 - \frac{||\mathbf{u}_{t+1}^k||_2^2}{2||\mathbf{u}_t^k||_2}\bigg) \le \sum_{k=1}^{c}\bigg(||\mathbf{u}_t^k||_2 - \frac{||\mathbf{u}_t^k||_2^2}{2||\mathbf{u}_t^k||_2}\bigg). \tag{32}$$

Combining Eq.(31) and Eq.(32), we arrive at

$$\sum_{k=1}^{c}||\mathbf{u}_{t+1}^k||_2 \le \sum_{k=1}^{c}||\mathbf{u}_t^k||_2 \tag{33}$$

Therefore, Algorithm 1 will monotonically decrease in each iteration the objective function of the formulated problem in Eq.(13). □

4 Experimental results

In this section, we evaluate the performance of our method and compare it with other *state-of-the-art* methods on a toy data set, on UCI data sets and on four face data sets[1] including YaleB [33], CUM-PIE (POSE C05, POSE C07, POSE C09, POSE C27), FERET [34] and AR. The experiments are implemented by Matlab R2010b on a computer with Intel Core i7-2600 3.4GHz CPU and Windows 10 operating system. We use the 1-Nearest-Neighbor algorithm as the classifier to calculate the recognition rate in our experiments.

4.1 Visualization on toy data set

In this subsection, we present an experiment on a toy data set to illustrate the property of robustness to outliers of our method. In Fig. 3, the toy data set consists of 200 samples which are divided into two classes (the red points and the yellow points), and we put six outlier samples in Fig. 3(b). Three lines are the projections of one-dimensional subspace learnt by RLDA, LDA and MMC. In Fig. 3(a), we apply three methods to the data set without outliers and obtain the projection vectors $\omega_{LDA} = [0.0240, -0.0010](\theta_{LDA} = 2.44°)$, $\omega_{MMC} = [-0.9985, 0.0542](\theta_{MMC} = 3.11°)$, $\omega_{RLDA} = [-0.0024, 0.0013](\theta_{RLDA} = 2.99°)$, which indicates that they obtain similar projection vectors and all work well. However, in Fig. 3(b) we add the outlier samples in the data set and get the projection vectors $\omega_{LDA} = [0.0190, -0.0135](\theta_{LDA} = 35.55°)$, $\omega_{MMC} = [-0.8258, 0.5639](\theta_{MMC} = 34.33°)$, $\omega_{RLDA} = [-0.0167, 0.0012](\theta_{RLDA} = 4.13°)$. It is evident that as the influence of outliers, both LDA and MMC do not work well while our method continues to have a similar result as in Fig. 3(a). This example illustrates that our method can avoid the influence of outliers and find a projection vector which can preserve the information that helps in classification and discrimination between classes even though the training set contains outliers.

4.2 Experiments on UCI data sets

In this subsection, we present experiments on three low-dimensional data sets, Wine data set, Balance Scale data set and Australian data set taken from UCI Machine Learning Repository.[2] In each experiment, we randomly select 30%, 40%, 50%, 60% and 70% of all samples for training and the rest of them for testing respectively. For fair comparison with LDA, the final dimension is set to $c-1$. All experiments are repeated ten times and results are presented in Table 2, Table 3 and Table 4, where the mean and standard deviation in these tables represent the mean of classification accuracy and standard deviation. The baseline results are obtained by applying 1-Nearest-Neighbor classifier on original data sets directly. Table 1 shows the details of the data sets from UCI database and the Letter-BA2[3] data set which will be used in the final discussion. In order to verify the validity of our experimental results, some statistical tests, namely one tail T-test, are conducted on every paired comparison where one method of each pair is the proposed method and the another one is one of competitors.

On the Wine data set, according to the p-value generated by one-tail T-test, our method has similar performances with LDA. However, except for LDA, our method outperforms the other methods significantly and it achieves the highest recognition accuracy on all training rates. From Table 3, we

1. http://www.face-rec.org/databases/

2. https://archive.ics.uci.edu/ml/datasets.html
3. http://www.cs.nyu.edu/ roweis/data.html

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

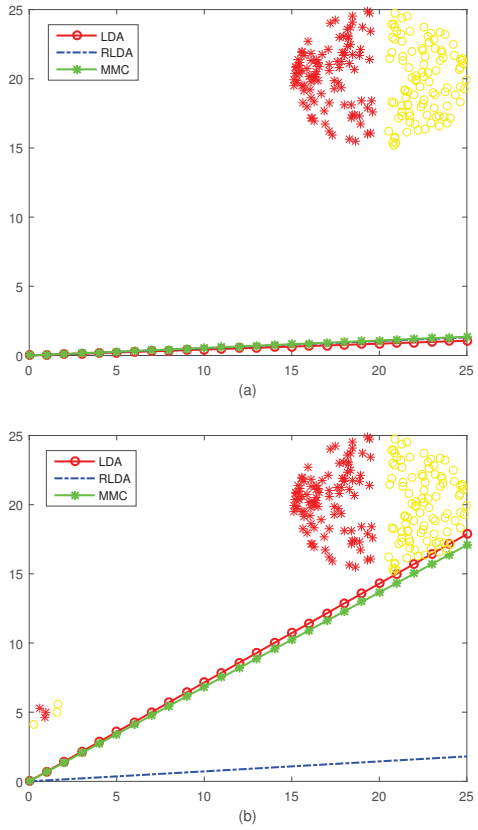

Fig. 3: Projection vectors learnt by RLDA, LDA and MMC on artificial data set without/with outliers.

TABLE 1: Descriptions of the data sets

Data sets	No.of variables	No.of class	No.of instances
Balance	4	3	625
Australian	14	2	690
Wine	13	3	178
Heart	13	5	294
Vehicle	18	4	846
Cancer	9	2	683
Letter-BA2	80	15	585

can conclude that the observed differences of accuracy are not significant between the proposed method and the three compared methods including LDA, NMMP and ALDE, when the training rate is 40%. However, in the cases of other training rates, the superiority of the proposed method seem to be evident. On the Australian data set, our method works well in most of comparisons and achieves the best accuracy in correspondence to training rates. These experimental results show that our methods also have competitive performance on the low-dimensional real data sets.

To further visualize the performance of subspace learning of our method, we project the Wine date set onto a two-dimensional subspace generated by PCA, LDA, MMC, NMMP, ALDE, TR, RLDA. Specifically, we randomly select about 30% of all samples as training set to learn a two-dimensional subspace, then the rest of the samples are projected onto the two-dimensional subspace. Fig. 4 shows the scatter plots of the first two features of Wine date set and the extracted two-dimensional features by different methods. From Fig. 4, we can conclude that the samples are almost separable in the subspace generated by LDA, MMC, Trace Ratio and RLDA. In particular, we note that there are no overlapped samples in the LDA and RLDA subspace as well as no contiguous heterogeneous samples in RLDA subspace. These results confirm that RLDA also can effectively extract information that is able to discriminate between classes, even though the training data do not contain outliers.

4.3 Experiments on face data sets

In this section, classification experiments are conducted on several real-world face data sets by comparing our method with five state-of-the-art dimensionality reduction algorithms to evaluate the ability of outlier suppression in high-dimensional data. First, we randomly select the training samples and randomly add some small black block occlusions into them as outliers, where the number of outliers in training samples is gradually increasing in each experiment, and the maximum number of outliers is half the number of training samples per class. Then, for computational convenience, each image is resized to a suitable size. Last, we apply PCA to reduce the dimensionality that can help to overcome the SSS problem. All the parameters about the preprocessing are displayed in Table 5. All experiments are repeated twenty times in the different dimensions to obtain the mean recognition accuracy in classification as a performance measure. Fig. 5, 6, 7, 8 show the recognition accuracy of our method (RLDA) and five comparison experiments in different dimensions, in which the numerical value of x-coordinate represents the number of outliers in the training set per class. For instance, in Fig. 5, the numerical value of x-coordinate is 3 that denotes training data set contains $3 \times c$ outliers in total, where c is the number of classes.

In what follows, we describe the details of each experiment.

AR [32] data set contains over 4000 color images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions, illumination conditions and occlusions(sun glasses and scarf). We use a subset (50 men and 50 women, each person has 26 images) of AR and the color images are converted into gray images.

In Fig. 5, our approach obtains stable results, even though the number of outliers reaches the maximum in different dimensions. With the increase of outliers, the superiority of our method becomes more obvious. It is worth noting that the recognition rates of LDA and NMMP are always higher than those of other comparative methods.

FERET [31] data set is one of the largest publicly available data sets with two versions, gray and color FERET data. We use the gray FERET face data set which includes 200 distinct individuals in which each individual has seven different images. Some images are taken at different gestures, facial expressions and light conditions. Each image in the data set is 256 gray-levels after having been cropped.

In this experiment, our method is not the best one when the training samples do not contain outliers. However, whenever the training samples contain outliers, the

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

TABLE 2: Experimental results on Wine data set with different training rates.

Method	30%		40%		50%		60%		70%	
	mean	std	mean	std	mean	std	mean	std	mean	std
Baseline	0.6824	0.0302	0.6776	0.0400	0.6898	0.0379	0.6775	0.0462	0.6679	0.0591
LDA	0.8696†	0.0646	0.8869†	0.0493	0.9193†	0.0276	0.8944†	0.0394	0.8396†	0.1046
MMC	0.6848	0.0259	0.7037	0.0473	0.7273	0.0467	0.7056	0.0545	0.7057	0.0557
NMMP	0.6688	0.0930	0.6925	0.0746	0.6864	0.0846	0.7239	0.0830	0.7283	0.0542
TR	0.8264	0.0388	0.8065	0.0741	0.8591	0.0322	0.8732	0.0398	0.8472†	0.0730
ALDE	0.6456	0.0256	0.6505	0.0493	0.6557	0.0432	0.6183	0.0354	0.6264	0.0425
RLDA	**0.8744**	0.0672	**0.8879**	0.0481	**0.9216**	0.0365	**0.9028**	0.0412	**0.8509**	0.1120

TABLE 3: Experimental results on Balance data set with different training rates.

Method	30%		40%		50%		60%		70%	
	mean	std	mean	std	mean	std	mean	std	mean	std
Baseline	0.7276	0.0153	0.7061	0.0162	0.6811	0.0191	0.6652	0.0238	0.6487	0.0258
LDA	0.8612	0.0207	0.8720†	0.0229	0.8712	0.0156	0.8728	0.0183	0.8642	0.0092
MMC	0.8628	0.0218	0.8715	0.0159	0.8708	0.0297	0.8748	0.0240	0.8690†	0.0199
NMMP	0.8674	0.0179	0.8819†	0.0174	0.8696	0.0181	0.8696	0.0368	0.8599	0.0359
TR	0.8635	0.0195	0.8661	0.0175	0.8689	0.0319	0.8736	0.0145	0.8652	0.0242
ALDE	0.8610	0.0180	0.8757†	0.0173	0.8721	0.0209	0.8768	0.0194	0.8631	0.0169
RLDA	**0.8874**	0.0197	**0.8837**	0.0217	**0.8929**	0.0284	**0.8940**	0.0143	**0.8813**	0.0300

TABLE 4: Experimental results on Australian data set with different training rates.

Method	30%		40%		50%		60%		70%	
	mean	std	mean	std	mean	std	mean	std	mean	std
Baseline	0.5934	0.0316	0.5957	0.0612	0.5799	0.0304	0.6134	0.0295	0.6010	0.0376
LDA	0.6178†	0.0419	0.5599	0.0705	0.5767†	0.1371	0.5938	0.0794	0.5812	0.1247
MMC	0.6035	0.0466	0.5570	0.0618	0.5735	0.0602	0.5873	0.0775	0.5633	0.0984
NMMP	0.5400	0.0615	0.5560	0.0484	0.5430	0.0467	0.5786	0.0588	0.5536	0.0511
TR	0.6083	0.0301	0.6024	0.0367	0.5942	0.0292	0.6011	0.0610	0.5778	0.0660
ALDE	0.5988	0.0327	0.6014	0.0286	0.5701	0.0396	0.6062	0.0420	0.6111	0.0343
RLDA	**0.6340**	0.0454	**0.7121**	0.0685	**0.6192**	0.0181	**0.6580**	0.0527	**0.6754**	0.0684

† The observed differences of accuracy are not significant (p-value > 0.05) between competitors and the proposed method according to one tail T-test.

TABLE 5: Some parameters in preprocessing.

	AR	FERET	YaleB	PIE
Training number	1300	800	1207	5777
Testing number	1300	600	1207	5777
Class	100	200	38	68
Original size	165×120	80×80	32×32	32×32
Occlusion number	50	50	20	20
Occlusion size	8×8	8×8	6×6	6×6
Resize	16×16	16×16	16×16	16×16

performance of the comparison methods decreases significantly while the performance of our method keeps stable and has not changed dramatically, which demonstrates the robustness to outliers of proposed method.

YaleB [35] data set is an extension of Yale data set which contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. In our experiment, we simply use the cropped images and resize them to 32×32 pixels. This data set now has 38 individuals and around 64 near frontal images under different illuminations per individual.

On this data set, our method has a similar performance with LDA on the subspace of 15 and 20 dimensions, but it outperforms the other methods on the subspace of 5 and 10 dimensions. Thus, with a reduced number of dimensions it permits to reach superior performances than competitors.

CMU-PIE [36] data set contains 41368 images of 68 people, in which each person is presented in 13 different poses, under 43 different illumination conditions, and with 4 different expressions. In our experiment, we only use a subset of PIE which contains 11554 images; they are divided into 4 classes (POSE C05, POSE C07, POSE C09, POSE C27) based on the different poses. Each image contains 256 gray-levels and is resized to 32×32 pixels after having been cropped.

Similarly to the AR face recognition experiment, in this data set our method makes significant improvements in comparison to other methods. It is worth noting that the final selected dimension is not beyond the number of classes.

Besides, for fair comparison to LDA, we also compare the performance between our method and other methods when the final dimension is set to $c - 1$. The experimental

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

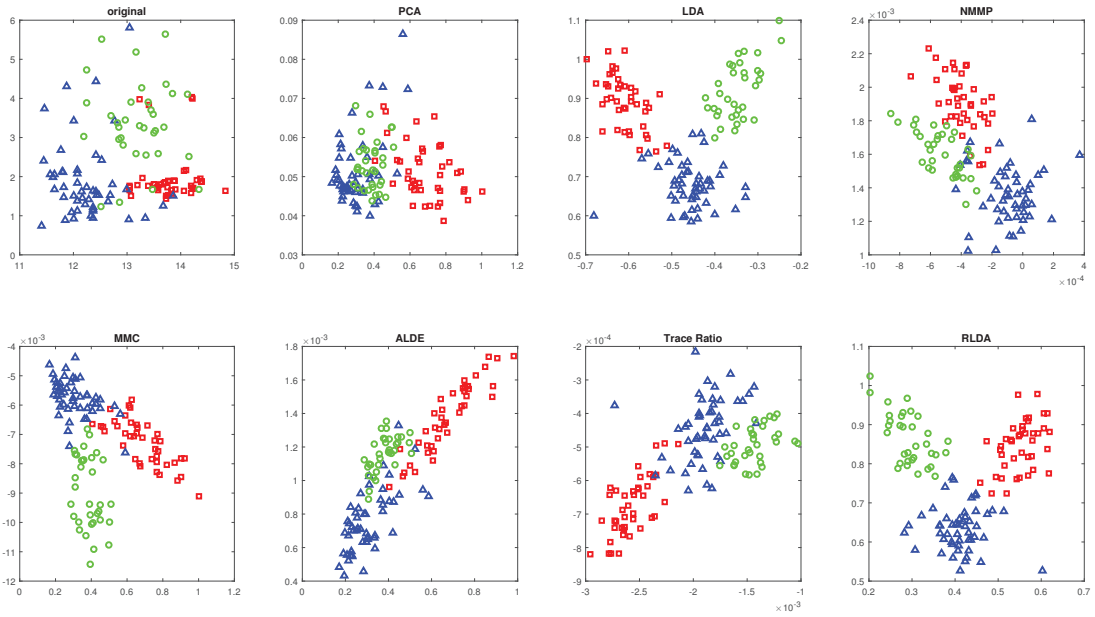

Fig. 4: Scatter plot of Wine data set projected onto a two-dimensional subspace generated by PCA, LDA, MMC, NMMP, ALDE, TR, RLDA

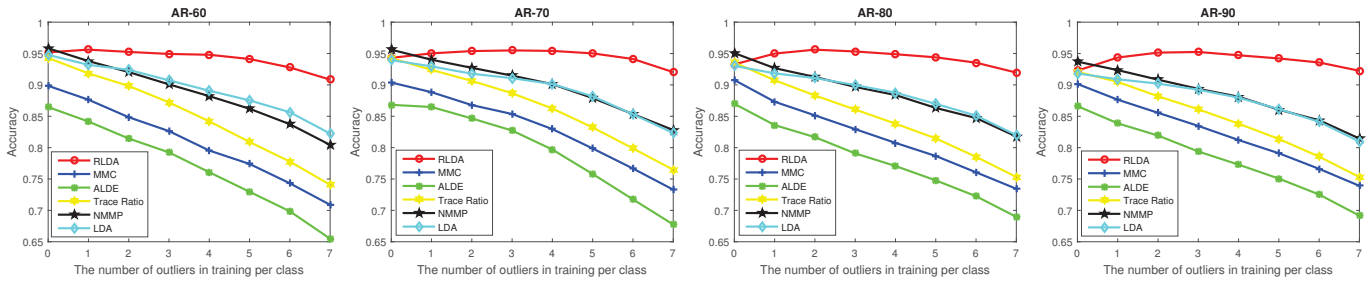

Fig. 5: Recognition accuracy of compared methods in 60,70,80,90 dimensions on the AR database.

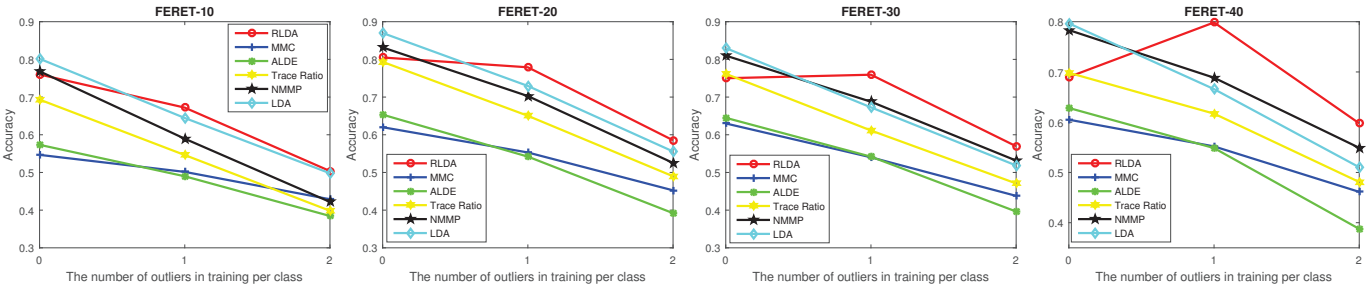

Fig. 6: Recognition accuracy of compared methods in 10,20,30,40 dimensions on the FERET database.

results are presented in Fig. 9. Obviously, the performance of LDA has some improvements in YaleB data set and it works well in the FERET data set when the training samples do not contain outliers. In addition, our method still has good performance in most data sets even in the case of the final dimensions are $c-1$. However, since the class number of FERET database is 200 and the final dimensions are 199, our method does not work well, which might indicate that the limitation of our method is not suitable for high-dimensional subspaces. Recall the results, it observes that our method does not work well when there are no outliers in the training samples, and its performances have a little promotion when the training samples contain few outliers. It possibly indicates that our method is designed for the data with outliers specially, which is both an advantage and a limitation of our method. Nonetheless, the recognition rate of our method has remained stable with the increase of outliers, which can verify the robustness to outliers of our

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

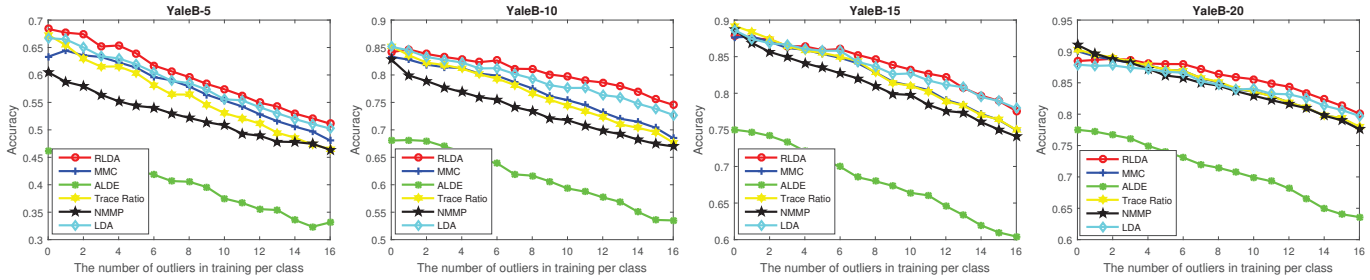

Fig. 7: Recognition accuracy of compared methods in 5,10,15,20 dimensions on the YaleB database.

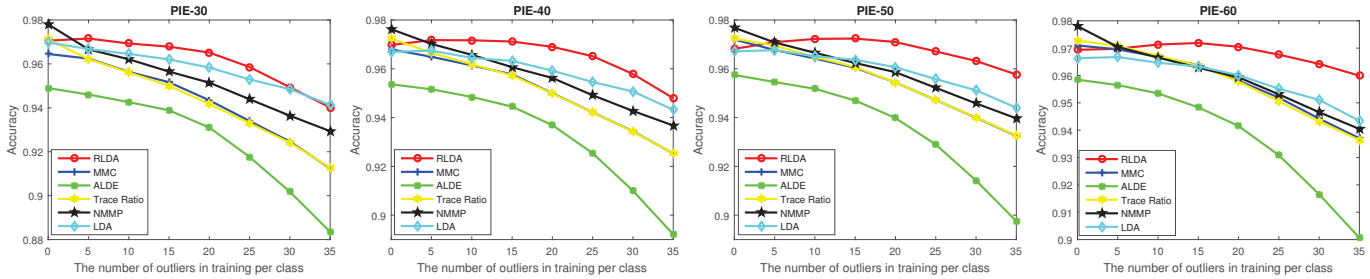

Fig. 8: Recognition accuracy of compared methods in 30,40,50,60 dimensions on the PIE database.

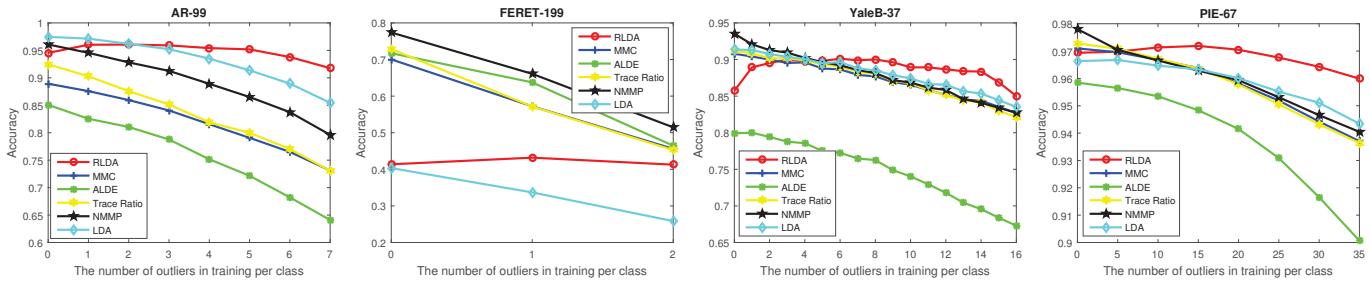

Fig. 9: Recognition accuracy of compared methods in $c-1$ dimensions on the four face databases.

method as well.

Some statistical tests (namely, one-tail T-test) are conducted to show the observed differences of accuracy between our method and each competitor. The statistical test results of all compared pairs are shown in Table 6, although the observed differences of accuracy are not significant on FERET data set, the best recognition rates are always obtained by proposed method when training data contain outliers. Furthermore, the significant differences on other data sets guarantee the evidence of improvements of our method on face classification. In addition, we test the convergence speed of RLDA on all data sets in Fig. 10. It is obvious that our iterative optimization algorithm converges to the optimum within less than 10 iterations on most of data sets. Moreover, on the AR and PIE data sets, proposed method converges only within less than 5 iterations. As a consequence, our method can efficiently achieve to convergence in practice.

4.4 Computational Complexity Analysis

Given a data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, the computational complexity of RLDA is divided into five parts in each iteration.

1. We need $O(nd)$ to calculate variable $\mathbf{m}_k$ according to Eq.(22).
2. We need $O(nd^2)$ to obtain the matrix $\mathbf{M}$.
3. We need $O(max(nd^2, d^3))$ to get the matrix $\boldsymbol{S_t}^{-1}\mathbf{M}$.
4. We need $O(d^3)$ to acquire transformation matrix $\mathbf{W}$ based on eigen-decomposition of $\boldsymbol{S_t}^{-1}\mathbf{M}$.
5. We need $O(2cmd)$ to update d_{ik}, where c is the number of classes and m is the final number of dimensions.

Overall, considering that $n \gg m, n \gg c$, the computational complexity of RLDA is $O(max(nd^2, d^3))$ which is linear with respect to n and efficient for handling large-scale dimensionality reduction problem.

In addition, we also report the average running times (per run) of each method in Table 7 which contains the time of dimensionality reduction and projection computation (training time) on real data sets. As shown in Table 8, the running time of NMMP is the longest one in the case of large

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

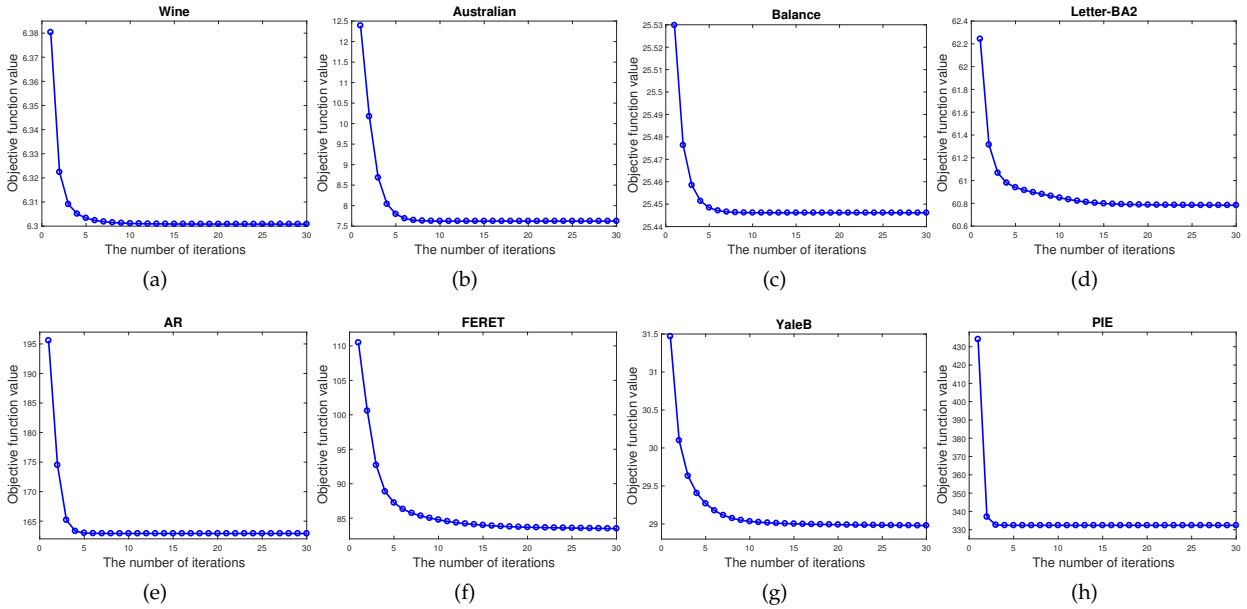

Fig. 10: The convergence curve of RLDA on Wine, Australian, Balance, Letter-BA2, AR, FERET, YaleB and PIE data sets, respectively. (a) Wine. (b) Australian. (c) Balance. (d) Letter-BA2. (e) AR. (f) FERET. (g) YaleB. (h) PIE.

TABLE 6: One tail T-test with the results of Fig. (6)-(9).

Data	AR				FERET				YaleB				PIE			
Dimension	60	70	80	90	10	20	30	40	5	10	15	20	30	40	50	60
LDA					†	†	†	†								
MMC					†			†								
Trace Ratio						†	†	†								
NMMP					†	†	†	†								
ALDE								†								

† The observed differences of accuracy are not significant (p-value > 0.05) between competitors and proposed method according to one tail T-test.

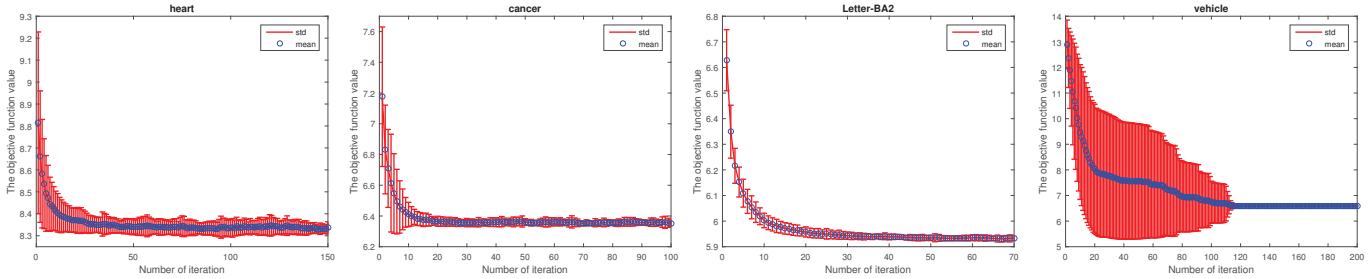

Fig. 11: The standard deviation and mean of objective function value in the 50 rounds of the experiments.

TABLE 7: The mean and std of the converged objective function value of our method using 50 random initialization.

Data	Heart	Letter-BA2	Vehicle	Cancer
Mean±std	8.34±0.02	5.39±0.01	6.72±0.91	6.36±0.03

scale data sets, because it is inevitable to assign k nearest neighbors to each point in the training process. On the contrary, thanks to the fact that MMC obtains projection vectors through matrix $\boldsymbol{S_b} - \boldsymbol{S_w}$ rather than $\boldsymbol{S_w^{-1}}$, it is faster than LDA. We should point out that our method has comparable performance with LDA in most cases. Moreover, as the scale of data increases, the time consumption of our method does not increase significantly.

4.5 Discussion

Since the problem (13) is a non-convex optimization problem, we can only obtain the local optimum solution theoretically. In this subsection, a discussion about the influence of different initializations on the objective function

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

TABLE 8: Running time (per run) on 7 real data sets

Data set	LDA	MMC	TR	NMMP	ALDE	RLDA
Wine	0.01s	0s	0.01s	0.02s	0s	0.02s
Balance	0.02s	0s	0.02s	0.07s	0.01s	0.01s
Australian	0.02s	0s	0.97s	0.05s	0s	0.05s
AR	14.04s	1.34s	2.56s	23.27s	1.34s	26.00s
FERET	6.88s	0.93s	1.53s	11.10s	0.95	14.40s
YaleB	7.71s	1.25s	1.94s	15.09s	1.26s	4.99s
PIE	54.55s	5.23s	22.69s	352.68s	5.28s	21.67s

value is presented. In detail, we use 50 times different random initializations to calculate objective function value in each iteration through the Algorithm 1 on UCI data set. We list the mean and standard deviation of the objective function value after convergence in Table 7. Furthermore, Fig. 11 plots the mean marked as a blue circle as well as the standard deviation markets as red error bar of the objective function value for each iteration in different initializations. Obviously, the objective function value of our method will converge to a stable value with different random initializations, which indicates that the solution obtained by our iterative algorithm is a global optimum on these data sets in practice.

5 Conclusions

In this paper, we propose a new formulation of robust linear discriminant analysis for dimensionality reduction which joints $L_{2,1}$-norm on objective function to alleviate the influence of outliers. An efficient iterative algorithm is proposed to solve the optimization problem and the proof of convergence of the optimization algorithm is also provided. Extensive experiments conducted on Toy and several real-world data sets have achieved excellent performance with comparison to some dimensionality reduction methods, which demonstrates the effectiveness of proposed method. Specifically, the recognition rate allowed by our method does not show any dramatic decline with the increasing presence of outliers in training samples. It demonstrates that our method can efficiently and effectively overcome the problem of outliers.

We will apply the $L_{2,1}$-norm on other algorithms such as multivariate linear regression, PCA and sparse coding and exploit more efficient algorithms to solve the optimization problem in future works.

Acknowledgment

This research is supported in part by the National Natural Science Foundation of China(61402002, 61502002, 61300057), the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry(NO.48,2014-1685), the Natural Science Foundation of Anhui Province(1408085QF120, 1408085MKL94), the Key Natural Science Project of Anhui Provincial Education Department (KJ2016A040), and Open Project of IAT Collaborative Innovation Center of Anhui University (ADXXBZ201511).

References

[1] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen, "Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 320–333, 2006.

[2] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[3] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–7.

[4] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[6] H. Murase, F. Kimura, M. Yoshimura, and Y. Miyake, "An improvement of the auto-correlation matrix in pattern matching method and its application to handprinted hiragana," *Trans. IECE*, vol. 64, no. 3, pp. 276–283, 1981.

[7] H. Kim, P. Howland, and H. Park, "Dimension reduction in text classification with support vector machines," *Journal of Machine Learning Research*, vol. 6, no. Jan, pp. 37–53, 2005.

[8] A. M. Martínez and A. C. Kak, "Pca versus lda," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[9] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, pp. 159–203, 1948.

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

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

[12] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.

[13] M. Skurichina and R. P. Duin, "Stabilizing classifiers for very small sample sizes," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 2. IEEE, 1996, pp. 891–896.

[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[15] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.

[16] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 483–502, 2005.

[17] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[18] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional datawith application to face recognition," *Pattern recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[19] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," in *Advances in neural information processing systems*, 2004, pp. 97–104.

[20] S. Liu, L. Feng, and H. Qiao, "Scatter balance: An angle-based supervised dimensionality reduction," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 2, pp. 277–289, 2015.

[21] H. Zhao, Z. Wang, and F. Nie, "Orthogonal least squares regression for feature extraction," *Neurocomputing*, vol. 216, pp. 200–207, 2016.

[22] Q. Ke and T. Kanade, "Robust subspace computation using l1 norm," 2003.

[23] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-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.

[24] 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.

[25] 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.

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

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/TKDE.2018.2842023, IEEE Transactions on Knowledge and Data Engineering

[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 3, pp. 619–632, 2013.
[31] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.
[32] A. M. Martinez, "The ar face database," *CVC technical report*, 1998.
[33] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 684–698, 2005.
[34] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.
[35] 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 transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.
[36] 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.

Haifeng Zhao received the B.Eng. in the electrical engineering in 1995, and the Ph.D. degree in computer science in 2006 from Anhui University. He is a professor at the School of Computer Science and Technology at Anhui University, Hefei, China. His current research interests include medical image processing, pattern recognition and computer vision.

Zheng Wang received the masters degree from Anhui University, Anhui, China, in 2017. He is currently pursuing the Ph.D. degree at Northwestern Polytechnical University. His research interests include machine learning, data mining and its applications, such as dimensionality reduction, clustering and so on.

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 top journals and conferences: TPAMI, IJCV, TIP, TNNLS/TNN, TKDE, Bioinformatics, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, ACM MM. His papers have been cited more than 6500 times. He is now serving as Associate Editor or PC member for several prestigious journals and conferences in the related fields.

1041-4347 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

View publication stats