Pattern Recognition Letters 84 (2016) 149–155

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Orthogonal discriminant analysis revisited

Yong Wang[a,c,*], Jian-Bin Xie[b], Yi Wu[c]

[a] Low Speed Aerodynamics Institute of China Aerodynamics Research and Development Center, Mianyang, 621000, China
[b] Department of Electronic Science and Engineering, National University of Defense Technology, Changsha, 410073, China
[c] Department of Mathematics and Systems Science, National University of Defense Technology, Changsha, 410073, China

ARTICLE INFO

Article history:
Received 27 March 2016
Available online 26 September 2016

Keywords:
Dimensionality reduction
Discriminant analysis
Orthonormality
Face recognition

ABSTRACT

Orthogonal discriminant analysis (ODA) methods extend traditional discriminant analysis (DA) methods under the condition of orthonormality of features. Despite many practical successes of the ODA methods in the literature of face recognition, some basic properties and crucial problems with respect to the ODA methods have not been explored or solved yet. For this sake, we revisit ODA in this paper. First, we introduce a new technique quite different from traditional one to answer one open problem raised by Cai et al. (IEEE Transactions on Image Processing, 2006), i.e., a unified theoretical justification for understanding and explaining the experimental phenomenon that the eigenvalues of the ODA methods are consistently larger than their DA counterparts. Comprehensive comparisons and extensive experiments on twenty real data sets verify our theoretical conclusion. Second, we reveal a fundamental problem concerning the usability of the ODA methods through our experiments, i.e., they are not consistently better than those of the corresponding DA methods in terms of the performance of recognition, especially when they were used onto low-dimensional problems.

1. Introduction

Linear dimensionality reduction methods are computationally simple, analytically tractable, and have the direct ability to deal with the testing data points [2], thus they are widely used in recognition. However, these methods often suffer from some limitations, for example, the basis functions obtained by some of these methods are non-orthogonal or statistically correlated, and some of them can not extract nonlinear structures in the data or make full use of the natural structure and correlation information between nearby pixels encoded in tensor objects. Therefore, orthogonalization [4,5], uncorrelation [16,17,28], kernelization [25], and tensorization [29] extensions of the linear methods have been proposed in the literature.

Among these extensions, orthogonalization technique has attracted substantial interests in the past few decades [4–6,10,14,15,18,30]. This technique was generally performed on discriminant analysis (DA) methods, such as LDA (Linear Discriminant Analysis) [1], LPP (Locality Preserving Projections) [13] and Locality Sensitive Discriminant Analysis (LSDA) [3]. Mathematically, the basis functions of these methods can be equivalently reformulated as the eigenvectors $\mathbf{g}$ of a certain generalization eigenvalue decomposition problem $\mathbf{XAX}^T\mathbf{g} = \lambda\mathbf{XBX}^T\mathbf{g}$ associated with the maximum eigenvalues, where $\mathbf{X}$ is the sample set for model training, while $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices that describe certain desired statistical or geometrical properties. Generally, the basis functions or the eigenvectors of $\mathbf{XAX}^T\mathbf{g} = \lambda\mathbf{XBX}^T\mathbf{g}$ are non-orthogonal. Thus, orthogonalization extensions of the DA methods, named as orthogonal discriminant analysis (ODA) methods, add a step to enforce these bases to be orthogonal to one another. Extensive experimental results in the literature have shown that the ODA methods classify better than their corresponding DA methods in high-dimensional face recognition.

Despite many practical successes of the ODA methods, there are several unresolved issues: (a). Several empirical studies [4,15,18] have shown that the eigenvalues of the ODA method are consistently larger than its DA counterpart. However, it is still unclear whether there is any theoretical guarantee to justify this experimental phenomenon. (b). The effectiveness of the ODA methods is primarily demonstrated on several face data sets, such as the ORL and Yale face databases, a natural question that arises is whether they will success on other data sets or there will have any new insights on other data sets. (c). No statistical testing was performed to compare the recognition results of the ODA methods with their corresponding DA methods, thus it is unclear whether the ODA methods are marginal compared to or statistically better than the DA methods. If we can prove that the ODA methods classify statistically better than the corresponding DA methods on

* Corresponding author. Fax: +86 816 2467566.
E-mail address: nudt604@aliyun.com (Y. Wang).

http://dx.doi.org/10.1016/j.patrec.2016.09.010
0167-8655/

different kinds of data sets, then the usability of the orthogonalization extensions is very significant. On the contrary, if we can not prove this judgement, for example, give some counterexamples through experiments, then orthogonalization extension should be used with caution.

To answer the above mentioned issues and complete the underpinning of the orthogonalization extension, in this paper, we firstly provide rigorous mathematical analysis to show that the eigenvalues of the ODA method are indeed larger than its DA counterpart. This mathematical justification offers a unified theoretical view for understanding and explaining the empirical observations in several existing references. A series of experiments on twenty real data sets verifies our theoretical conclusion. Secondly, comprehensive comparisons and extensive experiments reveal that the ODA methods outperform their corresponding DA methods when they were used onto high-dimensional small sample size problem. However, when they were used onto low-dimensional problem, they do not always classify better than their DA counterparts.

The rest of this paper is organized as follows. Section 2 gives a brief review of the general framework of the DA and ODA methods, as well as a class of DA methods studied in this paper. Section 3 discusses the theoretical comparison between orthogonal discriminant analysis and discriminant analysis. In Section 4, we report on experiments. Finally, we provide some concluding remarks in Section 5.

2. Background and related work

Suppose that there are N D-dimensional samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \Re^D$ which are grouped into a single data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \Re^{D\times N}$. Assume the class label of the sample $\mathbf{x}_i$ is $c_i \in \{1, 2, \cdots, M\}$, where M is the number of classes and each class has N_j samples with $N = \sum_{j=1}^{M} N_j$. In effect, linear dimensionality reduction methods seek to find d unitary basis functions $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_d$, which are assembled in a linear transformation matrix $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_d] \in \Re^{D\times d}$, to map the original data $\mathbf{x}_i, i = 1, 2, \cdots, N$ onto a lower d-dimensional (typically $d \ll D$) space $\mathbf{y}_i = \mathbf{G}^T\mathbf{x}_i \in \Re^d$ such that $\mathbf{y}_i$ "represents" $\mathbf{x}_i$ well in terms of some optimal criterions.

2.1. The general framework: DA vs ODA

Despite the different motivations of the DA methods, they can be interpreted in the following optimal criterion [1,3,11–13,27]:

$$\mathbf{g}_{opt} = \arg_{\mathbf{g}}\max \ J(\mathbf{g}) = \arg_{\mathbf{g}}\max \ \frac{\mathbf{g}^T(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}})\mathbf{g}}{\mathbf{g}^T(\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})\mathbf{g}}, \tag{1}$$

where $\mathbf{A}, \mathbf{B} \in \Re^{N\times N}$ are symmetric matrices that encode certain desired statistical or geometric properties. It is easily observed that the basis functions $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_d$ of the DA methods are the maximum eigenvalue solutions to the generalized eigen-problem $\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{g} = \lambda\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{g}$, ordered according to their eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and $\mathbf{g}_k, k = 1, 2, \cdots, d$ is the eigenvector corresponding to λ_k. The basis functions can also be regarded as the eigenvectors of the matrix[1] $\left(\mathbf{X}\mathbf{B}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{A}\mathbf{X}^T$ associated with the maximum eigenvalues. Generally, this matrix is not symmetric, thus the basis functions of the DA methods are non-orthogonal.

To obtain orthogonal basis functions, the ODA methods sequentially extract features by maximizing the optimal criterion of the corresponding DA methods subject to the orthogonality of features. More formally, the objective function of the ODA methods can be written as

$$\mathbf{g}_1 = \arg_{\mathbf{g}}\max \ \frac{\mathbf{g}^T(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}})\mathbf{g}}{\mathbf{g}^{\mathbf{T}}(\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})\mathbf{g}}, \tag{2}$$

and for $k = 2, \cdots, d$,

$$\mathbf{g}_k = \arg_{\mathbf{g}}\max \ \frac{\mathbf{g}^T(\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}})\mathbf{g}}{\mathbf{g}^T(\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})\mathbf{g}},$$

$$\text{subject to } \mathbf{g}_k^T\mathbf{g}_1 = \mathbf{g}_k^T\mathbf{g}_2 = \cdots = \mathbf{g}_k^T\mathbf{g}_{k-1} = 0. \tag{3}$$

It is easy to check that the first basis function $\mathbf{g}_1$ of the ODA methods is the same as the first basis vector of the corresponding DA methods. The other basis functions of the ODA methods are generally computed by using Lagrange's method of indeterminate multipliers. The detail derivations can be traced back to [4–6,10,14,15,18,30], with the starting point and initial paper on the subject seems to be [6].

2.2. Examples of the DA methods

Several discriminant analysis methods can be formulated in the general form of Eq. (1), including LDA, LPP, NPE (Neighborhood Preserving Embedding), LSDA and MFA (Marginal Fisher Analysis) [2]. It should be noted that the list above is certainly not complete in this rapidly developing field of research, we just review the most important and popular among them.

2.2.1. LDA

LDA aims to find optimal discriminant vectors on which the distances between different classes in the projection space are maximized while the distances within the same class are minimized. The choice of $\mathbf{A}$ in LDA is the between-class scatter matrix while the choice of $\mathbf{B}$ is the within-class scatter matrix [1,6,30], which are defined as

$$\mathbf{A}^{LDA} = \frac{1}{N}\sum_{j=1}^{M} N_j(\mu_j - \mu)(\mu_j - \mu)^T,$$

$$\mathbf{B}^{LDA} = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j}(\mathbf{x}_i^j - \mu_j)(\mathbf{x}_i^j - \mu_j)^T, \tag{4}$$

where $\mathbf{x}_i^j$ is the ith sample in the jth class, $\mu = 1/N\sum_{j=1}^{M}\sum_{i=1}^{N_j}\mathbf{x}_i^j$ is the overall mean and $\mu_j = 1/N_j\sum_{i=1}^{N_j}\mathbf{x}_i^j$ is the mean of the jth class.

2.2.2. LPP

LPP [12,13] seeks a subspace to ensure that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" in the original space, then $\mathbf{y}_i$ and $\mathbf{y}_j$ are "close" in the projection space. Let $N_K(\mathbf{x}_i)$ denotes the set of K nearest neighbors of $\mathbf{x}_i$ which share the same label with $\mathbf{x}_i$, then the choice of $\mathbf{A}$ in LPP $\mathbf{A}^{LPP}$ is

$$A_{ij}^{LPP} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_K(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_K(\mathbf{x}_i) \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

and $\mathbf{B}^{LPP}$ is a diagonal matrix with $B_{ii}^{LPP} = \sum_{j=1}^{N} A_{ij}^{LPP}$.

2.2.3. NPE

NPE [11] aims at finding a low-dimensional space that optimally preserve the local neighborhood structure of the data manifold. Let $\mathbf{W}$ be an $N \times N$ local reconstruction coefficient matrix, which is defined as $W_{ij} = 0$ if $\mathbf{x}_j \notin N_K(\mathbf{x}_i)$ and the other W_{ij} are computed by

[1] $\mathbf{X}\mathbf{B}\mathbf{X}^T$ may be singular and thus does not have inverse, especially when $N < D$. In this case, we can employ a procedure similar to the PCA + LDA or the Fisherface method proposed by Belhumeur et al. [1], namely projecting the data set onto a suitable intermediate dimension using PCA to make it non-singular. For the sake of simplicity, we still use $\mathbf{X}$ and D to denote the data matrix in the PCA subspace and the dimensionality.

[2] In this paper, the solved generalized eigen-peoblem of LPP, NPE and MFA has been equivalently reformulated to follow the optimal criterion of (1).

minimizing the following reconstruction cost function under the constraint of $\sum_{j=1}^{N} W_{ij} = 1$:

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{j=1}^{N} W_{ij}\mathbf{x}_j \right\|_2^2. \tag{6}$$

Then, the choices of $\mathbf{A}$ and $\mathbf{B}$ for NPE are $\mathbf{A}^{NPE} = \mathbf{W} + \mathbf{W}^T - \mathbf{W}\mathbf{W}^T$ and $\mathbf{B}^{NPE} = \mathbf{I}_N$ which is the $N \times N$ identity matrix.

2.2.4. LSDA

LSDA [3] tries to maximize the margin between data points from different classes at each local neighborhood, i.e., the nearby points with the same label are mapped close to each other while the nearby points with different labels are mapped far apart. Two graphs, i.e. *within-class graph* $G_w = \{\mathbf{X}, \mathbf{W}_w\}$ and *between-class graph* $G_b = \{\mathbf{X}, \mathbf{W}_b\}$, are constructed to model both geometrical and discriminant structure of the data in LSDA. The similarity matrices of G_w and G_b are defined as:

$$W_{w,ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_w(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_w(\mathbf{x}_i) \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

$$W_{b,ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_b(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_b(\mathbf{x}_i) \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $N_w(\mathbf{x}_i)$ contains the points with the same label as $\mathbf{x}_i$ among its K nearest neighbors, while $N_b(\mathbf{x}_i)$ contains the neighbors having different labels.

The choices of $\mathbf{A}$ and $\mathbf{B}$ for LSDA are $\mathbf{A}^{LSDA} = \alpha\mathbf{L}_b + (1-\alpha)\mathbf{W}_w$ and $\mathbf{B}^{LSDA} = \mathbf{D}_w$, where both $\mathbf{D}_w$ and $\mathbf{D}_b$ are diagonal matrices whose entries are $D_{w,ii} = \sum_j W_{w,ij}$ and $D_{b,ii} = \sum_j W_{b,ij}$, respectively. $\mathbf{L}_b = \mathbf{D}_b - \mathbf{W}_b$ and $\alpha \in [0, 1]$ is a suitable constant.

2.2.5. MFA

MFA [27] seeks to preserve the intraclass compactness and at the same time suppress the interclass separability. Let $N_{K_1}^{+}(\mathbf{x}_i)$ denotes the index set of the K_1 nearest neighbors of the sample $\mathbf{x}_i$ which share the same label with $\mathbf{x}_i$, while $N_{K_2}^{-}(\mathbf{x}_i)$ denotes the set of the K_2 nearest neighbors of $\mathbf{x}_i$ among the samples whose labels are different to that of $\mathbf{x}_i$. Define

$$W_{ij}^{+} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_{K_1}^{+}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{K_1}^{+}(\mathbf{x}_i) \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

$$W_{ij}^{-} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_{K_2}^{-}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{K_2}^{-}(\mathbf{x}_i) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Then, the choices of $\mathbf{A}$ and $\mathbf{B}$ for MFA are $\mathbf{A}^{MFA} = \mathbf{D}^{-} - \mathbf{W}^{-}$ and $\mathbf{B}^{MFA} = \mathbf{D}^{+} - \mathbf{W}^{+}$, where both $\mathbf{D}^{-}$ and $\mathbf{D}^{+}$ are diagonal matrices whose entries are $D_{ii}^{+} = \sum_j W_{ij}^{+}$ and $D_{ii}^{-} = \sum_j W_{ij}^{-}$, respectively.

3. Orthogonal discriminant analysis versus discriminant analysis

3.1. The basis functions of ODA

From the constraints in (3), $\mathbf{g}_k$ must be found in the $(D-k+1)$-dimensional subspace S^{D-k+1} orthogonal to the $k-1$ features g_i, $i = 1, 2, \cdots, k-1$. Let $\mathbf{C}^{(k-1)} = \mathbf{I}_D - \sum_{i=1}^{k-1} \mathbf{g}_i\mathbf{g}_i^T$, where $\mathbf{I}_D$ is the $D \times D$ identity matrix. Suppose the Singular Value Decomposition (SVD) of $\mathbf{C}^{(k-1)}$ is

$$\begin{aligned} \mathbf{C}^{(k-1)} &= \mathbf{U}^{k-1}\mathbf{\Sigma}^{k-1}(\mathbf{V}^{k-1})^T \\ &= \begin{bmatrix} \mathbf{U}_s^{k-1} & \mathbf{U}_s^{\tilde{k-1}} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_s^{k-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_s^{k-1} & \mathbf{V}_s^{\tilde{k-1}} \end{bmatrix}^T \\ &= \mathbf{U}_s^{k-1}\mathbf{\Sigma}_s^{k-1}(\mathbf{V}_s^{k-1})^T. \end{aligned} \tag{11}$$

where $\mathbf{U}^{k-1}$ and $\mathbf{V}^{k-1}$ are orthogonal matrices, $\mathbf{\Sigma}_s^{k-1} \in \Re^{s\times s}$ is non-singular and $s = rank(\mathbf{C}^{(k-1)})$. Then, we can obtain the following theorem.

Theorem 1. *S^{D-k+1} can be spanned by $\mathbf{U}_s^{k-1}$, i.e., $S^{D-k+1} = range(\mathbf{U}_s^{k-1})$.*

Proof. Since $\mathbf{g}_i$, $i = 1, 2, \cdots, k-1$ are orthogonal and unitary, then we have $\mathbf{C}^{(k-1)}\mathbf{g}_i = (\mathbf{I}_D - \sum_{i=1}^{k-1} \mathbf{g}_i\mathbf{g}_i^T)\mathbf{g}_i = \mathbf{g}_i - \mathbf{g}_i = 0$. Moreover, it is easily observed that the rank of $\mathbf{C}^{(k-1)}$ is $D-k+1$. Thus, $range(\mathbf{C}^{(k-1)})$ is a $(D-k+1)$-dimensional subspace orthogonal to the $k-1$ features $\mathbf{g}_i$, $i = 1, 2, \cdots, k-1$, that is S^{D-k+1}. Finally, the column vectors of $\mathbf{U}_s^{k-1}$ are a set of orthogonal bases of $range(\mathbf{C}^{(k-1)})$ [9]. Therefore, $S^{D-k+1} = range(\mathbf{U}_s^{k-1})$. □

Theorem 1 shows that $\mathbf{g}$ in (3) can be denoted by $\mathbf{g} = \mathbf{U}_s^{k-1}\xi$. Then, the optimal problem of (3) becomes:

$$\xi_k = \arg\max_{\xi} \frac{\xi^T(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi}{\xi^T(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi}. \tag{12}$$

It is easy to check that the optimal feature $\boldsymbol{\xi}_k$ of (12) is the eigenvector corresponding to the largest eigenvalue of $(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi = \lambda(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi$. Then, we obtain $\mathbf{g}_k$ as $\mathbf{g}_k = \mathbf{U}_s^{k-1}\xi_k/\|\xi_k\|$. Since $(\mathbf{U}_s^{k-1})^T\mathbf{U}_s^{k-1} = \mathbf{I}_s$, we see that $\mathbf{g}_k$ is a unitary vector.

3.2. The theoretical comparison between ODA and DA

In this subsection, we prove theoretically that the eigenvalues of the ODA method are consistently larger than its DA counterpart which had been observed experimentally in previous studies [4,15,18].

From Section 2 and Section 3.1, we have known that the kth basis function $\mathbf{g}_k^{DA}$ of the DA method is given by the eigenvector corresponding to the kth largest eigenvalue λ_k^{DA} of $\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{g} = \lambda\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{g}$. While, the kth transformation vector $\mathbf{g}_k^{ODA}$ of its ODA counterpart is given by $\mathbf{g}_k^{ODA} = \mathbf{U}_s^{k-1}\xi_{k,\max}/\|\xi_{k,\max}\|$, where $\boldsymbol{\xi}_{k,\max}$ is the eigenvector corresponding to the largest eigenvalue $\lambda_{k,\max}$ of $(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi = \lambda(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U}_s^{k-1}\xi$. Then, the kth value of the DA method is

$$J(\mathbf{g}_k^{DA}) = \frac{(\mathbf{g}_k^{DA})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{g}_k^{DA}}{(\mathbf{g}_k^{DA})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{g}_k^{DA}} = \lambda_k^{DA}, \tag{13}$$

while the kth value of the ODA method is

$$\begin{aligned} J(\mathbf{g}_k^{ODA}) &= \frac{(\mathbf{g}_k^{ODA})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{g}_k^{ODA}}{(\mathbf{g}_k^{ODA})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{g}_k^{ODA}} \\ &= \frac{(\xi_{k,\max})^T[(\mathbf{U}_s^{i-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_s^{i-1}]\xi_{k,\max}}{(\xi_{k,\max})^T[(\mathbf{U}_s^{i-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{U}_s^{i-1}]\xi_{k,\max}} \\ &= \lambda_{k,\max}. \end{aligned} \tag{14}$$

That is, each of the eigenvalues is associated with the value of the corresponding optimal criterion. In the following, we clarify that $\lambda_{k,\max} \geq \lambda_k^{DA}$, which implies that the eigenvalues of the ODA method are consistently larger than its corresponding DA method. This observation is based on the following theorem known as the Poincare Separation Theorem [23].

Theorem 2. *Let $\mathbf{R}$ be an $n \times n$ Hermitian matrix and $\mathbf{S}$ is an $n \times k$ matrix such that $\mathbf{S}^T\mathbf{S} = \mathbf{I}_k$, then*

$$\lambda_i(\mathbf{S}^T\mathbf{R}\mathbf{S}) \geq \lambda_{n-k+i}(\mathbf{R}), i = 1, \cdots, k \tag{15}$$

where $\lambda_i(\cdot)$ denotes the ith eigenvalue of a matrix.

To obtain our result, we also need the following Corollary of Theorem 2 [24] and a Lemma [8].

Corollary 1. *Let* $\mathbf{P}$ *be an* $n \times k$ *matrix such that* $\mathbf{P}^T\mathbf{Q}\mathbf{P} = \mathbf{I}_k$, *where* $\mathbf{Q}$ *is a positive definite matrix. Then*

$$\lambda_i(\mathbf{P}^T\mathbf{R}\mathbf{P}) \geq \lambda_{n-k+i}(\mathbf{Q}^{-1}\mathbf{R}), i = 1, \cdots, k. \tag{16}$$

Lemma 1. *The eigenvalues* $\lambda_i, i = 1, \cdots, n$ *of* $\mathbf{B}^{-1}\mathbf{A} \in \Re^{n\times n}$ *are invariant under any non-singular transformation matrix* $\mathbf{\Psi}$, *i.e.,*

$$\lambda_i\{(\mathbf{\Psi}^T\mathbf{B}\mathbf{\Psi})^{-1}\mathbf{\Psi}^T\mathbf{A}\mathbf{\Psi}\} = \lambda_i(\mathbf{B}^{-1}\mathbf{A}). \tag{17}$$

From the above facts, we can obtain the following theorem which gives the main result of this section. It should be noted that the derivation presented here is partly motivated by [10].

Theorem 3. *The eigenvalues of the ODA method are consistently larger than its DA counterpart, that is*

$$\lambda_{k,\max} \geq \lambda_k^{DA}, k = 1, \cdots, d. \tag{18}$$

Proof. It is easy to see that $\lambda_{1,\max} = \lambda_1^{DA}$ since both of them are the first largest eigenvalue of $\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{g} = \lambda\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{g}$. Now, we show that $\lambda_{k,\max} \geq \lambda_k^{DA}, k = 2, \cdots, d$.

Since $\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}$ is a positive definite matrix, it is easy to check that $\mathbf{B}^{k-1} \triangleq (\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}$ is also a positive definite matrix. Then, the SVD of $\mathbf{B}^{k-1}$ has the form of $\mathbf{B}^{k-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{\Sigma}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{B}^{k-1}$ and the column vectors of $\mathbf{U}$ are the corresponding eigenvectors. Denote $\mathbf{E}^{k-1} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$ and $\mathbf{A}^{k-1} = (\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^T\mathbf{U}_s^{k-1}$, it is easily observed that

$$\mathbf{I}_s = (\mathbf{E}^{k-1})^T\mathbf{B}^{k-1}\mathbf{E}^{k-1} = (\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\,\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}. \tag{19}$$

Then, by using Corollary 1, we have

$$\begin{aligned}
&\lambda_1\{[(\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}]^{-1}\\
&\quad(\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}\}\\
&= \lambda_1\{(\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}\}\\
&\geq \lambda_{D-s+1}((\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}).
\end{aligned} \tag{20}$$

Moreover, utilizing Lemma 1, we obtain

$$\begin{aligned}
&\lambda_1\{[(\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}]^{-1}\\
&\quad(\mathbf{U}_s^{k-1}\mathbf{E}^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}\mathbf{E}^{k-1}\}\\
&= \lambda_1(\{(\mathbf{E}^{k-1})^T[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]\mathbf{E}^{k-1}\}^{-1}\\
&(\mathbf{E}^{k-1})^T[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]\mathbf{E}^{k-1})\\
&= \lambda_1\{[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]^{-1}[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]\}.
\end{aligned} \tag{21}$$

Since $s = rank(\mathbf{C}^{(k-1)}) = D - k + 1$, from (20) and (21), we have

$$\begin{aligned}
&\lambda_1\{[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]^{-1}[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]\}\\
&\geq \lambda_k((\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}).
\end{aligned} \tag{22}$$

Then, it is easy to check that

$$\begin{aligned}
\lambda_{k,\max} &= \lambda_1\{[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]^{-1}[(\mathbf{U}_s^{k-1})^T\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{U}_s^{k-1}]\}\\
&\geq \lambda_k((\mathbf{X}\mathbf{B}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}) = \lambda_k^{DA},
\end{aligned} \tag{23}$$

which proves this theorem. □

4. Experimental results

In this section, we conduct a series of experiments on twenty real data sets to verify our theoretical conclusion and to empirically investigate the performance of the ODA methods compared to their corresponding DA methods. The methods involved are LDA, LPP, NPE, LSDA and MFA, as well as their corresponding orthogonalization extensions, i.e., OLDA [5,6,10,30], OLPP [4], ONPE [18], OLSDA [15] and OMFA, respectively.

4.1. Databases and experimental setting

Twenty data sets from various data sources were used in our experimental studies, including multi-feature digit data sets, face images, palmprint data, and so on. Glass, Ionosphere, Iris, Liver_disorders, Mfeat_fou, Mfeat_kar, Mfeat_mor, Mfeat_pix, Mfeat_zer, Musk_clean1, Sonar, Vowel and Wine are thirteen data sets from the UCI machine learning repository[3] with dimensionality lower than 300. ORL [21], Yale [26], AR [19] and FERET [22] are four benchmark face databases used widely in face recognition community. PolyU[4] and Fingdb[5] are palmprint database and fingerprint database, respectively. The processed Coil-20 database [20] is a database with 1440 normalized gray-scale images of 20 different objects. These last seven data sets all contain high-dimensional image data. Without any loss of generality, all images are normalized and cropped to a resolution of 32×32 pixels in our experiments.

To ensure our results do not depend on any special choice of the training data and enable us to analyze classification performance from the viewpoint of statistics, we carry out random experiments on all the twenty data sets. Each data set was randomly partitioned 30 times into a training set where each class consists of two-thirds of the whole class and the rest are considered to be the testing set, similar setting as in [30]. In all experiments, classification was performed in the projected space using nearest neighbor classifier, which is used in all the previous references [4–6,10,14,15,18,30].

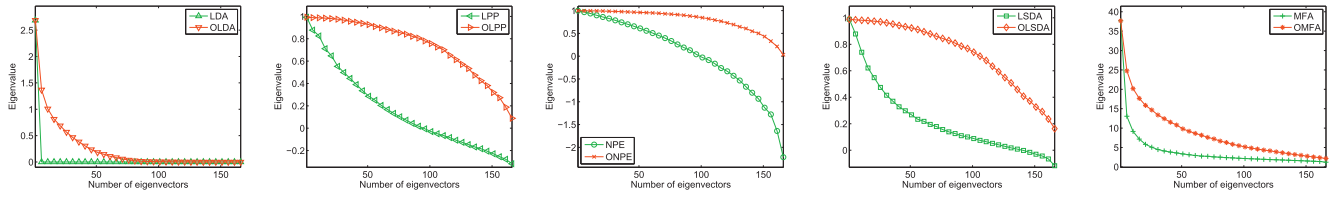
According to theoretical analysis, an upper bound of the dimensionality of the reduced space for LDA is $M - 1$ when there are M different classes. Thus, the upper bound of the reduced dimensionality for LDA is set to $M - 1$. For fair comparison, we use the same matrices $\mathbf{A}$ and $\mathbf{B}$ in the ODA method and its DA counterpart.

4.2. Eigenvalues: ODA vs. DA

In this subsection, we validate our theoretical analysis that the ODA methods are more powerful than their corresponding DA methods in terms of the optimal criterion. As can be seen in Theorem 3, they can be measured by the eigenvalues of ODA compared to DA. All the twenty data sets have been used for this study. Due to page limitation, only the experimental results on the Musk_clean1 data set are presented in Fig. 1. The results confirm our theoretical analysis in Section 3, that is, the eigenvalues of the ODA method are consistently larger than its DA counterpart.

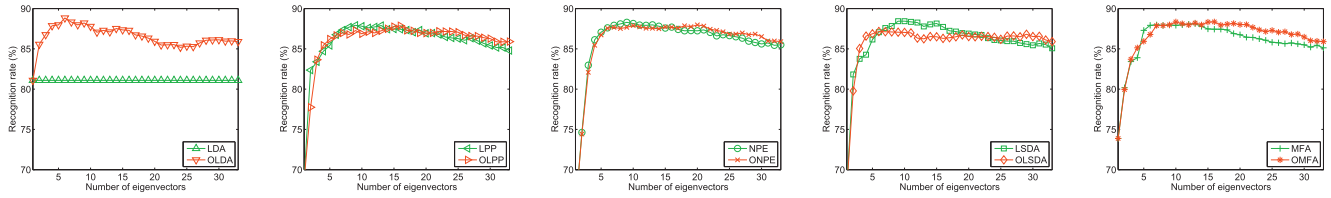
Since the optimal criterions of the different DA methods are designed to provide more discriminative ability and it has been shown that they are potentially related to the recognition power [1,3,11,13,27], we may expect that the ODA methods will achieve better recognition performance than the DA methods, which will be examined in the following experiments.

[3] http://archive.ics.uci.edu/ml
[4] http://www.comp.polyu.edu.hk/~biometrics
[5] http://archive.ics.uci.edu/ml



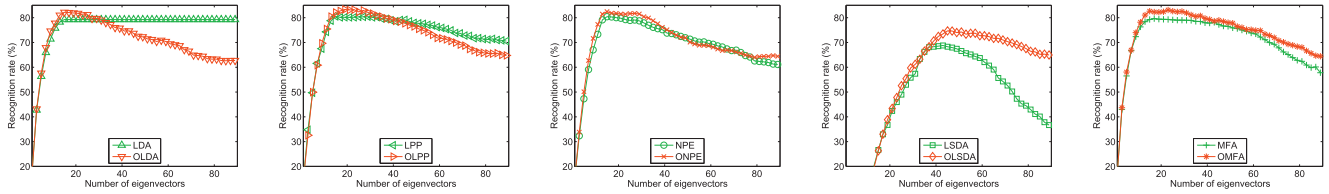
(a) OLDA vs. LDA (b) OLPP vs. LPP (c) ONPE vs. NPE (d) OLSDA vs. LSDA (e) OMFA vs. MFA

Fig. 1. Eigenvalues of the ODA methods and their corresponding DA methods on the Musk_clean1 data set.



(a) OLDA vs. LDA (b) OLPP vs. LPP (c) ONPE vs. NPE (d) OLSDA vs. LSDA (e) OMFA vs. MFA

Fig. 2. Average recognition accuracy versus dimensionality reduction of the different methods on the Ionosphere data set.



(a) OLDA vs. LDA (b) OLPP vs. LPP (c) ONPE vs. NPE (d) OLSDA vs. LSDA (e) OMFA vs. MFA

Fig. 3. Average recognition accuracy versus dimensionality reduction of the different methods on the Yale data set.

4.3. Recognition results: ODA vs. DA

In this subsection, we compare the performance of the ODA methods with their corresponding DA methods for recognition on all the twenty data sets. The recognition process has three steps: First, the projection matrix of each method is computed using the training set; then, all the data points including the training set and the testing set are projected onto the feature space; finally, recognition is performed by classifier in the feature space.

Figs. 2 and 3 show the plots of average recognition rate obtained by the different methods over 30 independent runs versus the variation of dimensions on the Ionosphere and Yale data sets, respectively. As can be seen, the recognition rates vary with the dimension of the feature space. The best average recognition accuracy and the corresponding standard deviations of the different methods are tabulated in Table 1. The best performances among the ODA methods and their corresponding DA methods are highlighted by a star and the performances without significant difference with the best performance (paired t-test at 95% significance level) are boldfaced.

From these results, we can make several interesting observations: (a). From Table 1, we can see that the loss of the OLDA, OLPP, ONPE, OLSDA and OMFA to their corresponding original methods on the twenty data sets are 10, 40, 30, 30 and 25%. While, from the viewpoint of statistic paired t-test, the loss are 5, 20, 15, 30 and 5%, respectively. All the ODA methods loss to their corresponding DA methods on the Mfeat_mor data set. (b). From Figs. 2 and 3, we can see that no method can consistently outperform the others on all the variation of dimensions. Even if one method outperform the other from the statistical viewpoint, it outperform the other only on some of the projected dimensions.

4.4. Statistical testing

In this subsection, we conduct paired t-test and paired Wilcoxon rank sum test at 95% significance level to further compare the recognition results from the view of statistics. The win/tie/loss of the different ODA methods with those of their corresponding DA methods are given in Table 2. It can be observed that the most effective among the ODA methods is the OLDA method, since it statistically outperforms the original LDA method on more than 80% of the twenty data sets. The reason may be that LDA has a serious shortcoming in that the maximum number of features to be extracted is $M-1$ when there are M classes, and thus the discriminatory capability is limited. On the contrary, the OLDA method can extract up to D features, which result in more discriminatory capability and recognition power.

No matter paired t-test or paired Wilcoxon rank sum test, all the ODA methods generally outperform their corresponding DA methods on the seven high-dimensional image data sets, especially on the face databases which are consistent with several existing results in the literature [4,15,18]. On the other hand, the loss of the ODA methods to their corresponding DA methods are almost occurred on the thirteen low-dimensional data sets from the UCI machine learning repository where the dimensionality of the data is smaller than the number of samples. This observation gives us a valuable insight into the introduction of the orthogonal extension, i.e., the ODA methods may be more effective on the high-dimensional small sample size problem [7] where the number of samples is smaller than their dimensionality.

Table 1
Comparison of the recognition accuracy (Mean ± Std%) on twenty real data sets.

Data set	LDA	OLDA	LPP	OLPP	NPE	ONPE	LSDA	OLSDA	MFA	OMFA
Glass	58.5 ± 7.1	**71.8 ± 4.3**∗	64.0 ± 5.9	**71.8 ± 4.3**∗	65.5 ± 4.9	**71.8 ± 4.3**∗	63.9 ± 6.3	**71.8 ± 4.3**∗	63.5 ± 6.4	**71.8 ± 4.3**∗
Ionosphere	81.1 ± 3.0	**88.8 ± 2.5**∗	**87.9 ± 2.6**	**87.9 ± 2.2**∗	**88.3 ± 2.3**∗	**88.0 ± 2.2**	**88.4 ± 3.2**∗	87.2 ± 2.5	**88.1 ± 2.3**	**88.3 ± 2.1**∗
Iris	**96.7 ± 1.7**∗	**96.7 ± 1.7**∗	**95.8 ± 1.9**∗	**95.4 ± 2.0**	**95.6 ± 2.1**	**95.8 ± 2.2**∗	**97.6 ± 2.1**∗	95.7 ± 2.1	**96.3 ± 3.0**∗	**95.4 ± 2.0**
Liver_disorders	59.2 ± 4.9	**62.2 ± 4.5**∗	59.3 ± 4.1	**61.9 ± 4.4**∗	60.5 ± 4.6	**62.1 ± 4.4**∗	59.5 ± 4.0	**61.9 ± 4.3**∗	**60.6 ± 4.4**	**62.0 ± 4.3**∗
Mfeat_fou	80.7 ± 1.0	**82.6 ± 0.9**∗	**83.5 ± 0.8**∗	83.1 ± 0.7	**83.1 ± 0.9**∗	**82.9 ± 1.1**	**83.5 ± 1.0**∗	82.9 ± 1.0	**83.7 ± 1.1**∗	**83.6 ± 0.9**
Mfeat_kar	95.3 ± 0.6	**97.1 ± 0.6**∗	**97.7 ± 0.6**∗	97.2 ± 0.6	**97.5 ± 0.5**∗	97.2 ± 0.6	**97.7 ± 0.4**∗	97.2 ± 0.5	**97.8 ± 0.6**∗	**97.6 ± 0.7**
Mfeat_mor	**67.0 ± 1.6**∗	62.2 ± 2.1	**66.5 ± 1.2**∗	55.1 ± 2.2	**59.4 ± 5.2**∗	47.8 ± 5.4	**59.5 ± 3.5**∗	54.2 ± 1.7	**66.3 ± 1.2**∗	65.0 ± 1.6
Mfeat_pix	94.7 ± 0.6	**97.5 ± 0.4**∗	**97.6 ± 0.5**	**97.6 ± 0.4**∗	**97.5 ± 0.4**	**97.6 ± 0.3**∗	**97.5 ± 0.5**	**97.6 ± 0.4**∗	97.4 ± 0.5	**97.9 ± 0.4**∗
Mfeat_zer	77.6 ± 1.6	**80.6 ± 1.0**∗	**80.2 ± 1.3**	**80.6 ± 1.0**∗	78.9 ± 1.7	**80.6 ± 1.0**∗	79.5 ± 1.2	**80.6 ± 1.0**∗	**80.7 ± 1.4**∗	**80.6 ± 1.0**
Musk_clean1	76.5 ± 3.4	**86.8 ± 3.0**∗	**89.2 ± 2.5**∗	**88.4 ± 2.5**	87.0 ± 2.8	**88.4 ± 2.0**∗	87.1 ± 3.1	**90.9 ± 2.8**∗	87.9 ± 2.1	**91.4 ± 2.2**∗
Sonar	71.0 ± 5.0	**84.0 ± 4.5**∗	**83.0 ± 4.2**∗	**82.5 ± 4.2**	**82.7 ± 4.6**∗	**82.5 ± 4.1**	**80.2 ± 6.3**	**82.0 ± 4.0**∗	80.9 ± 5.6	**83.0 ± 4.1**∗
Vowel	97.3 ± 1.1	**97.8 ± 0.9**∗	**97.8 ± 0.9**∗	**97.7 ± 1.0**	96.8 ± 1.1	**97.7 ± 1.0**∗	**97.7 ± 1.0**∗	**97.7 ± 1.0**∗	97.2 ± 1.5	**97.7 ± 0.9**∗
Wine	**96.0 ± 2.3**∗	**95.3 ± 2.4**	**94.5 ± 2.6**∗	89.4 ± 4.0	**84.6 ± 5.5**∗	83.4 ± 5.4	**92.8 ± 3.3**∗	88.6 ± 4.7	**95.8 ± 2.3**	**95.8 ± 2.0**∗
ORL	95.4 ± 1.6	**97.7 ± 1.1**∗	95.4 ± 1.9	**97.7 ± 1.2**∗	95.4 ± 1.6	**97.8 ± 1.2**∗	84.7 ± 3.1	**95.6 ± 1.6**∗	**96.4 ± 1.5**	**98.0 ± 1.1**∗
Yale	79.2 ± 4.4	**82.2 ± 3.5**∗	80.6 ± 4.4	**83.3 ± 3.9**∗	80.8 ± 4.2	**82.4 ± 4.1**∗	68.8 ± 5.4	**74.7 ± 4.8**∗	79.7 ± 4.1	**83.1 ± 4.0**∗
AR	88.3 ± 1.1	**96.9 ± 0.7**∗	76.6 ± 1.7	**96.4 ± 0.8**∗	74.0 ± 1.9	**95.9 ± 0.9**∗	29.8 ± 2.0	**51.1 ± 2.0**∗	89.8 ± 1.1	**97.5 ± 0.5**∗
FERET	75.4 ± 1.3	**93.6 ± 0.9**∗	53.2 ± 1.5	**93.4 ± 0.9**∗	53.2 ± 1.5	**92.6 ± 1.0**∗	9.5 ± 1.5	**19.5 ± 1.9**∗	78.6 ± 1.5	**95.5 ± 1.0**∗
PolyU	97.0 ± 0.8	**99.6 ± 0.2**∗	97.5 ± 0.6	**99.7 ± 0.2**∗	97.8 ± 0.5	**99.5 ± 0.3**∗	96.0 ± 0.8	**99.8 ± 0.2**∗	97.6 ± 0.7	**99.7 ± 0.2**∗
Fingdb	79.6 ± 4.5	**89.6 ± 3.1**∗	82.3 ± 4.5	**89.8 ± 2.8**∗	82.0 ± 4.4	**89.0 ± 3.6**∗	77.3 ± 4.6	**86.2 ± 3.8**∗	82.3 ± 4.1	**89.7 ± 3.2**∗
Coil-20	86.3 ± 1.5	**99.7 ± 0.3**∗	85.7 ± 1.6	**99.5 ± 0.5**∗	85.1 ± 1.4	**99.8 ± 0.4**∗	78.6 ± 2.0	**99.5 ± 0.4**∗	89.5 ± 1.5	**99.9 ± 0.2**∗

Table 2
The win/tie/loss of the different ODA methods versus their corresponding DA methods, after paired *t*-test and paired Wilcoxon Rank Sum Test (WRS test) at 95% significance level (w = win, t = tie, l = loss).

Data set	OLDA vs. LDA		OLPP vs. LPP		ONPE vs. NPE		OLSDA vs. LSDA		OMFA vs. MFA	
	t-test	WRS test	*t*-test	WRS test	*t*-test	WRS test	*t*-test	WRS test	*t*-test	WRS test
Glass	w	w	w	w	w	w	w	w	w	w
Ionosphere	w	w	t	t	t	t	l	t	t	t
Iris	t	t	t	t	t	t	l	l	t	t
Liver_disorders	w	w	w	w	w	t	w	t	t	t
Mfeat_fou	w	w	l	t	t	t	l	l	t	t
Mfeat_kar	w	w	l	l	l	t	l	l	t	t
Mfeat_mor	l	l	l	l	l	l	l	l	l	l
Mfeat_pix	w	w	t	t	t	t	t	t	w	w
Mfeat_zer	w	w	t	t	w	w	w	w	t	t
Musk_clean1	w	w	t	t	w	t	w	w	w	w
Sonar	w	w	t	t	t	t	t	t	w	t
Vowel	w	t	t	t	w	w	t	t	w	t
Wine	t	t	l	l	l	t	l	l	t	t
ORL	w	w	w	w	w	w	w	w	w	w
Yale	w	w	w	w	w	t	w	w	w	w
AR	w	w	w	w	w	w	w	w	w	w
FERET	w	w	w	w	w	w	w	w	w	w
PolyU	w	w	w	w	w	w	w	w	w	w
Fingdb	w	w	w	w	w	w	w	w	w	w
Coil-20	w	w	w	w	w	w	w	w	w	w

4.5. Discussion

In the above subsections, several experiments on a large number of real data sets have been systematically performed. These experiments reveal several interesting points:

(a) The experimental results on all the data sets showed that the eigenvalues of the ODA method are consistently larger than its DA counterpart. These results validate our theoretical analysis in Section 3.2 and offer a potential view for understanding and explaining the observed experimental phenomenon of the ODA methods in several existing references, such as [4,15,18].
(b) Though the optimal criterions of the ODA methods on the training data sets are consistently larger than their corresponding DA methods, the recognition accuracy on the testing data sets in Table 1, and statistical testing in Table 2 showed that the ODA methods do not consistently outperform their corresponding DA methods. Jin et al. [17] and Yang et al. [28] also observed that the discriminatory power of OLDA is much weaker than that of the uncorrelated extension of LDA (ULDA) though the corresponding Fisher criterion ratios of OLDA are much larger than those of ULDA. The reason of these phenomenons is that the transformed feature vectors of the ODA methods are generally statistically correlated which may lead to much information redundancy within the resulting features and make the effective discriminatory information insufficient though the optimal criterions of the ODA methods are much larger [16,17,28].
(c) Experimental results revealed that the ODA methods are more effective on the high-dimensional small sample size problem, no matter paired *t*-test or paired Wilcoxon rank sum test are used, which gives us some helpful instructions for the further use of the orthogolization extension.

5. Conclusion and future work

In this paper, we have introduced a new technique quite different from traditional Largrange's method to sequentially extract features which maximize the optimal criterions of the ODA methods subject to the orthonormality of features. Then, we provide rigorous mathematical analysis to show that the eigenvalues of the

ODA methods are consistently larger than those of the corresponding DA methods, which offers a theoretical guarantee to the empirical observations in several existing references. Comprehensive comparisons and extensive experiments on twenty real data sets not only verify our theoretical conclusion, but also demonstrate that the ODA methods do not consistently outperform their corresponding DA methods in terms of the performance of recognition, especially when they were used onto low-dimensional problems.

Acknowledgments

The authors want to thank the associate editor and anonymous reviewers for helpful comments and suggestions, and thank Prof. Zhi-Hua Zhou and De-Chuan Zhan in Department of Computer Science and Technology of Nanjing University for their helpful discussions. This research was supported by the National Science Foundation of China (61303188), the National Standards Project of China (201406), the Hunan Industry and Information Technology Innovation Project ([2015]-145), and the Innovation Program of China Aerodynamic Research and Development Center.

References

[1] P. Belhumeur, J. Hespanha, D. Kriengman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[2] Y. Bengio, J.F.O. Paiement, P. Vincent, O. Delalleau, N. Le Roux, M. Ouimet, Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering, in: Advances in Neural Information Processing Systems 16, 2004, pp. 177–184.
[3] D. Cai, X. He, K. Zhou, J. Han, H. Bao, Locality sensitive discriminant analysis, in: Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), Jan. 2007.
[4] D. Cai, X.F. He, J.W. Han, H.J. Zhang, Orthogonal Laplacianfaces for face recognition, IEEE Trans. Ima. Proc. 15 (11) (2006) 3608–3614.
[5] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 978–984.
[6] D.H. Foley, J. Sammon John W., An optimal set of discriminant vectors, IEEE Trans. Comput. C-24 (3) (1975) 281–289.
[7] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.
[8] K. Fukunaga, W.L.G. Koontz, A criterion and an algorithm for grouping data, IEEE Trans. Comput. 19 (1970) 917–923.
[9] G. Golub, C. Van Loan, Matrix Computations, third, Johns Hopkins Univ. Press, New York, 1996.
[10] Y. Hamamoto, T. Kanaoka, S. Tomita, On a theoretical comparison between the orthonormal discrimination vector method and discriminant analysis, Pattern Recognit. 26 (12) (1993) 1863–1867.
[11] X.F. He, D. Cai, S.C. Yan, H.J. Zhang, Neighborhood preserving embedding, in: Tenth IEEE International Conference on Computer Vision, 2005, pp. 1208–1213.
[12] X.F. He, P. Niyogi, Locality preserving projections, 2004, pp. 153–160.
[13] X.F. He, S.C. Yan, Y.X. Hu, P. Niyogi, H.J. Zhang, Face recognition using Laplacian faces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[14] H.F. Hu, Orthogonal neighborhood preserving discriminant analysis for face recognition, Pattern Recognit. 41 (6) (2008) 2045–2054.
[15] Y. Jin, Q.Q. Ruan, Orthogonal locality sensitive discriminant analysis for face recognition, J. Inf. Sci. Eng. 25 (2) (2009) 419–433.
[16] Z. Jin, J.Y. Yang, Z.S. Hu, L. Z., Face recognition based on the uncorrelated discriminant transformation, Pattern Recognit. 34 (7) (2001a) 1405–1416.
[17] Z. Jin, J.Y. Yang, Z.M. Tang, Z.S. Hu, A theorem on the uncorrelated optimal discriminant vectors, Pattern Recognit. 34 (10) (2001) 2041–2047.
[18] X.M. Liu, J.W. Yin, Z.L. Feng, J.X. Dong, L. Wang, Orthogonal neighborhood preserving embedding for face recognition, in: IEEE International Conference on Image Processing, 2007, pp. 133–136.
[19] A.M Martinez, R. Benavente, The AR Face Database, CVC Technical Report 24, 1998.
[20] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-20), Technical Report, CUCS-005-96, February 1996.
[21] Olivetti & Oracle Research Laboratory, The Olivetti & Oracle Research Laboratory Face Database of Faces, http://www.camorl.co.uk/facedatabase.html, 1994.
[22] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[23] C.R. Rao, Linear Statistical Inference and Its Applications, Wiley, New York, third edition, 1973.
[24] C.R. Rao, Matrix approximations and reduction of dimensionality in multivariate statistical analysis, in: P.R. Krishnaiah (Ed.), Multivariate Analysis V, North-Holland, Amsterdam, 1980, pp. 3–22.
[25] B. Scholkopf, A. Smola, K. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[26] Yale Univ. Face Database, http://cvc.yale.edu/projects/yalefaces/yalefaces.html, 2002.
[27] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[28] J. Yang, J.Y. Yang, D. Zhang, What's wrong with fisher criterion, Pattern Recognit. 35 (11) (2002) 2665–2668.
[29] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (1) (2004) 131–137.
[30] J.P. Ye, T. Xiong, Computational and theoretical analysis of null space and orthogonal linear discriminant analysis, J. Mach. Learn. Res. 7 (2006) 1183–1204.