Neurocomputing 79 (2012) 140–157

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Robust linearly optimized discriminant analysis

Zhao Zhang *, Tommy W.S. Chow

Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 8 May 2011
Received in revised form
20 October 2011
Accepted 29 October 2011
Communicated by X. Gao
Available online 15 November 2011

Keywords:
Linear discriminant analysis
Robustness
Orthogonality
Marginal inter- and intra-class scatters
Dimensionality reduction
Image recognition

ABSTRACT

Supervised *Fisher Linear Discriminant Analysis* (LDA) is a classical dimensionality reduction approach. LDA assumes each class has a Gaussian density and may suffer from the singularity problem when handling high-dimensional data. We in this work consider more general class densities and show that optimizing LDA criterion cannot always achieve maximum class discrimination with the geometrical interpretation. By defining new marginal inter- and intra-class scatters, we elaborate a pairwise criteria based optimized LDA technique called robust *linearly optimized discriminant analysis* (LODA). A multimodal extension of LODA is also presented. In extracting the informative features, two effective solution schemes are proposed. The kernelized extension of our methods is also detailed. Compared with LDA, LODA has four significant advantages. First, LODA needs not the assumption on intra-class distributions. Second, LODA characterizes the inter-class separability with the marginal criterion. Third, LODA avoids the singularity problem and is robust to outliers. Fourth, the delivered projection matrix by LODA is orthogonal. These properties make LODA more general and suitable for discriminant analysis than using LDA. The delivered results of our investigated cases demonstrate that our methods are highly competitive with and even outperform some widely used state-of-the-art techniques.

1. Introduction

Utilizing dimensionality reduction to extract a small number of features from high-dimensional datasets before performing classification is important. The extracted features are supposed to be holding the most representative information from high-dimensional datasets which usually consist of redundant information and noise. In the last decades, many linear or nonlinear, supervised or unsupervised, global or local dimensionality reduction methods, e.g. [2,3,10,18,19,22,23,26,32] have been proposed. Details can be referred to [1]. *Fisher Linear Discriminant Analysis* (LDA) [2,3] and *Principal Component Analysis* (PCA) [3] are the two most representative techniques of global supervised and unsupervised methods, respectively. Linearized PCA and LDA are widely applied in the pattern recognition community and have been proven to be more powerful for performing feature representation and extraction. PCA and LDA is that they are both single-modal methods [34], making them unable to deliver satisfactory results when handling multimodal data distributions. It is also noticed that, when the class labels accompanied with the data samples are available, LDA tends to be more effective and efficient than using the unsupervised PCA for pattern classification and image recognition [9].

Different from structure preservation based PCA, LDA focuses on achieving maximum class discrimination for classification. But it is noted that LDA suffers from several drawbacks. One of the major drawbacks is LDA requires intra-class scatter to be nonsingular. But this requirement poses certain limitation to some emerging applications, which are mostly described by high-dimensional distributions, because most of real data have many attributes, e.g., gene distributions and global climate patterns. Thus this may limit its certain real applications. To address this problem, many approaches, including [4–7,18], have been proposed. Despite some success, performing these types of LDA variants for dimensionality reduction may lose certain important discriminative information [12]. The second shortcoming of LDA is that it requires an implicit assumption that each class has a unimodal Gaussian distribution [2,3], enabling LDA to find the best directions for discrimination [8]. Note that this condition usually cannot be satisfied in many practical applications, because most real datasets deliver more complex distributions. In another word, LDA is usually unable to find the optimal discriminative directions when intra-class distributions are multimodal or when there are some faraway points. Another disadvantage of LDA is the rank limitation on the inter-class scatters. To overcome these problems, some methods, e.g. [8,13–16], have recently been presented. But it is noted that these methods either still surfer from the singularity problem caused by the intra-class scatter or focus on utilizing the nonparametric intra- and inter- class sample neighbor information around each data point for discriminant analysis.

* Corresponding author.
E-mail address: cszzhang@gmail.com (Z. Zhang).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.10.015

The processing of real life applications can be quite demanding because virtually all real datasets are not unimodally distributed. Also, most of the sample points in a class are usually located closely forming a dense region (or, field), creating similar embeddings [10,11]. There are often many faraway points or outliers. For discriminant analysis, we should take this distribution behavior and the intra-class dense regions into account when we aim at defining an effective discrimination criterion. From this perspective, there exists a drawback in the LDA intra-class scatter. Recall that the intra- and inter-class scatters of LDA are formulated based on the intra-class means and global mean. For those complex real distributions, sample means of different classes may be the same or very close due to the effect of faraway points, although most of inter-class points may be distributed in separated clusters. In this case, minimizing the original LDA intra-class scatter is equivalent to congregating the embeddings of all data points. Besides, LDA inter-class scatter is measured as a sum of pairwise distances between each class mean and global mean, but a cumulative large distance cannot guarantee a high separability between pariwise classes. Thus, there are many research works [18,25,27,30], which have been proposed to improve the original LDA inter-class scatter. But it should be noted that these methods all rely on the sample means of total within-class points, implying that they are single-modal and are usually incapable of embedding multimodal datasets respectably. Most importantly, these methods are sensitive to the faraway points and do not consider the neighborhood information of data. To address this issue, *Local Fisher Discriminant Analysis* (LFDA) [20] has recently been introduced to overcome the weakness of LDA against intra-class multimodality and outliers [20]. We in this paper also consider the neighborhood information of data as LFDA, but we mainly focus on utilizing the density region computed from each class and defining new marginal intra- and inter-class scatter criteria for improving the original LDA. To the best of our knowledge, to data no related work on this topic has been studied. The followings highlight the four major contributions of our work:

- By constructing the neighborhood sub-graph of each class via nearest neighbor search, data graph of each class l is divided into two parts, namely l-density-region and region out of the l-density-region according to the degrees of vertices. New marginal intra- and inter-class scatters are then formulated based on the density-regions. As a result, our proposed criterion is robust against the faraway sample points or outliers. Geometrical comparisons of our proposed marginal scatter criteria and the original LDA scatter criteria are also provided.
- Based on the newly defined large margin scatters, the drawbacks of LDA scatters can be effectively overcome. We then propose a general and robust linearly optimized discriminant analysis (LODA) technique. LODA finds the important marginal discriminant directions without assuming the class densities.
- LODA is linear, which makes it fast and suitable for real-world applications. Considering that nonlinear and multimodal structures are common in real data [20], we also focus on defining pairwise nonlinear and multimodal scatters and extending LODA to the nonlinear and multimodal dimensionality reduction scenarios for mining the nonlinear and multimodal structures hidden in the datasets.
- To avoid the singularity problem and compute the transforming axes steadily, the strategies of *trace ratio (TR) optimization* [19,22] and *maximum margin criterion* (MMC) [18] are used to formulate our presented problems. As a result, the obtained transforming basis vectors are guaranteed to be mutually orthogonal, implying that the similarity will not be changed if it is based on the Euclidean distance measure [22].
- The theoretical comparisons between our work and related works are elaborated. We mathematically show that our criterion is more general and exhibit a strong generalization capability. That is, PCA, LDA, MMC and *Locality Preserving Projections* (LPP) [23] can be interpreted with our criteria as special cases.

The outline of the paper is described as follows. In Section 2, we briefly review LDA. In Section 3, we describe our new marginal LODA scatter criteria mathematically. We then detail the multimodal extension of LODA. Two effective solution schemes are also presented. We then discuss the connections between this present work and the previous related works. We in Section 4 describe the simulation settings and evaluate our proposed techniques using benchmark datasets. Finally, we offer the concluding remarks in Section 5.

2. Fisher Linear Discriminant Analysis (LDA)

Let $x_i \in \Re^n (i=1,2,\ldots N)$ be vectors of N n-dimensional data and $y_i (\in \{1,2,\ldots,c\})$ be the associated class labels, where c is the number of classes. Then classical LDA aims to compute an optimal transformation matrix Ξ that maps each pattern x_i of $X=[x_1|x_2|\ldots|x_N]$ from an n-dimensional space to a feature vector in a d-dimensional space ($d \le n$). The embedding of each x_i is then given by $\Xi^{\mathrm{T}} x_i$, where T denotes the transpose of a matrix or a vector. Let $S^{(b)}$ and $S^{(w)}$ denote the LDA inter-class scatter and intra-class scatter given as

$$S^{(b)} = \sum_{l=1}^{c} N_l (\overline{M^{(l)}} - \overline{M})(\overline{M^{(l)}} - \overline{M})^{\mathrm{T}}, \tag{1}$$

$$S^{(w)} = \sum_{l=1}^{c} \sum_{i:y_i=l} (x_i - \overline{M^{(l)}})(x_i - \overline{M^{(l)}})^{\mathrm{T}}, \tag{2}$$

where N_l is the number of data points belonging to class $l \in \{1,2,\ldots,c\}$, $\sum_{i:yi=l}$ denotes the summation over i such that $y_i=l$, $\overline{M^{(l)}} = (1/N_l)\sum_{i:y_i=l} x_i$ is the average vector of points in class l and $\overline{M} = (1/N)\sum_{l=1}^{c}\sum_{i:y_i=l} x_i$ is the global mean of all the data points. Then LDA finds the $n \times d$ transformation matrix Ξ from the following *ratio trace* (or, *determinant ratio*) [19,22] problem for discrimination:

$$\Xi^* = \underset{\Xi \in \Re^{n\times d}}{\operatorname{argmax}} Tr((\Xi^{\mathrm{T}} S^{(w)} \Xi)^{-1}(\Xi^{\mathrm{T}} S^{(b)} \Xi)) = \underset{\Xi \in \Re^{n\times d}}{\operatorname{argmax}} \frac{\left|\Xi^{\mathrm{T}} S^{(b)} \Xi\right|}{\left|\Xi^{\mathrm{T}} S^{(w)} \Xi\right|}, \tag{3}$$

where $tr(H)$ denotes the matrix trace of matrix H. That is, LDA finds a transformation matrix Ξ such that the inter-class scatter is maximized while the intra-class scatter is minimized. Provided that the intra-class scatter $S^{(w)}$ has full rank, then the optimal transforming axes of LDA are given by the eigenvectors $\{\xi_j\}_{j=1}^{d}$ associated with the generalized eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$ of the following generalized eigen-problem:

$$(S^{(w)})^{-1} S^{(b)} \xi_j = \lambda_j \xi_j. \tag{4}$$

Note that $S^{(w)}$ is a positive semi-definite matrix, thus the inverse of $S^{(w)}$ may be singular. In order to ensure the stability of Eq. (4), we need to generalize $S^{(w)}$ by adding a regularization term and compute the transforming axes by solving the following generalized eigen-problem: $(S^{(w)} + \mu I)^{-1} S^{(b)} \xi = \lambda \xi$, where μ is a positive small number and I is an identity matrix. Then a solution Ξ is analytically given as $\Xi = (\xi_1|\xi_2|\ldots|\xi_d)$.

3. Robust linearly optimized discriminant analysis (LODA)

3.1. Formulation of LODA

For a given data matrix $X=[x_1|x_2|\ldots|x_N]$ with class labels $y_i(\in\{1, 2, \ldots, c\})$, we firstly conduct k-nearest neighbor search (NNS) to determine the nearest neighbor set $N_+^{(x_i)}$ of each sample $x_i^{(l)}$ in subset $X^{(l)}=[x_1^{(l)}|x_2^{(l)}|\cdots|x_{N_l}^{(l)}]$ belonging to the class $l(\in\{1,2, \ldots, c\})$, where N_l denotes the number of sample points of the lth class. It is noted that $x_i^{(l)}$ is the closest neighbor of the point $x_i^{(l)}$ itself. We then construct c neighborhood sub-graphs within each given class. Considering that, in real applications, sample points from the same object or class tend to be densely distributed with similar embeddings [10,11,23], thus most data points within an object or a class are distributed closely together. And commonly there is a small number of sample data points that are far away from the dense cluster center and are usually treated as outliers. The simple two-class case problem shown in Fig. 1(a) is a representative example of the real-world datasets.

In order to elaborate the argument, we only show the neighborhood sub-graph G_N^r within the class r in Fig. 1(a) to make the figure clearer. We conduct the $k(=3)$ nearest neighbor search to find the neighbors of each data point and construct the neighborhood sub-graph G_N^r for the class r. Based on the above descriptions, the following definitions are stated prior to formulating our proposed new scatter criteria.

Definition 1. (l-density-graph). The l-density-graph $\tilde{G}_D^l$ of class l consists of vertices $\{\hat{x}_i^l\}_{i=1}^{q_l}$ that are included in class l and have degree $D^l \geq (\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\beta$ in the neighborhood sub-graph G_N^l within class l. Where $\tilde{D}_i^l, i=1,2,\ldots,q_l$ denotes the degree of each vertex x_i^l, q_l is the number of vertices in graph $\tilde{G}_D^l$, $\hat{x}_i^l$ denotes a node or a vertex in the l-density-graph $\tilde{G}_D^l$, the density-data matrix $\widehat{X^{(l)}}=\left[\widehat{x_1^l}|\widehat{x_2^l}|\cdots|\widehat{x_{q_l}^l}\right]$, β is a controlling parameter. $\max(\tilde{D}_i^l)$ and $\min(\tilde{D}_i^l)$ denote the maximal and minimal values of $\tilde{D}_i^l$, respectively.

Definition 2. (l-density-region). The l-density-region Δ^l within class l is defined by the region that is formed by the vertices in $\widehat{X}^{(l)}=\left[\widehat{x_1^l}|\widehat{x_2^l}|\cdots|\widehat{x_{q_l}^l}\right]$ and edges or links included in the l-density-graph $\tilde{G}_D^l$. The sub-graph within the l-density-region is called the l-density-graph of class l.

We present the procedures for constructing the l-density-graph $\tilde{G}_D^l$ within each class l in Table 1, where W is a symmetric matrix, F actually reflects the importance of the vertices in the neighborhood graph, that is the bigger the value of F_{ii}, the more important the corresponding vertex is. It is noticed that β is the key parameter for determining the range of the l-density-region Δ^l within class l. According to Definition 1, the bound of degree D^l is $\max(\tilde{D}_i^l)\geq D^l\geq\min(\tilde{D}_i^l)$. Thus the bound of β is

$$\frac{(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))}{\max(\tilde{D}_i^l)}\leq\beta\leq\frac{(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))}{\min(\tilde{D}_i^l)}. \tag{5}$$

It is noted that when k in NNS is set to N_l, all data points have the same degrees, i.e. N_l-1. In this case, the lower bound and upper bound of β will be the same. According to Definitions 1 and 2, we have circled the constructed density-regions of the two classes using the schematic illustration in Fig. 1(a). In this case, $\beta=3$, $\min(\tilde{D}_i^r)=3$ over i and the lower bound of D^r is 5. Based on Definition 2, we can easily obtain the sample mean $\tilde{M}^{(r)}=(1/q_r)\sum_{j=1}^{q_r}\widehat{x_j^r}$ within class r through the points belonging to the r-density-region. Note that we

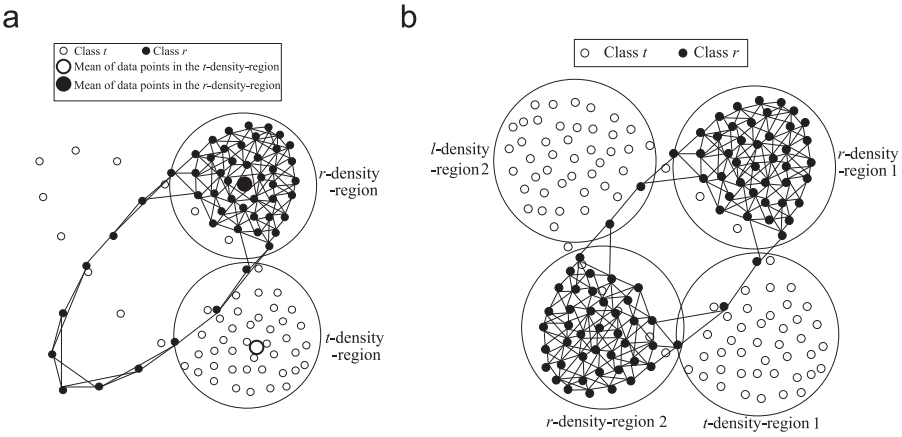

Fig. 1. Neighborhood graph construction based on a two-class case.

Table 1
Procedures for constructing the l-density-graph $\tilde{G}_D^l$ within each class l.

(1) $X^{(l)}\leftarrow\left(x_1^l|x_2^l|\ldots|x_{N_l}^l\right)$; % *Data matrix of class l.*
(2) Determine the neighbors of each point and construct the neighborhood graph G_N^l within class l.
(3) Construct a weight matrix W whose entries $W_{i,j}=1$ if x_i^l and x_i^l are neighbors, and else 0.
(4) Define a row or column vector F with entries $F_i=\sum_j W_{i,j}$. Sort F in descending order.
(5) Determine the vertices $\hat{x}_i^l$ whose degrees $\tilde{D}_i^l$ are bigger than D^l according to Step (4) and Eq. (5) and then form the density-data matrix $\widehat{X^{(l)}}=\left[\widehat{x_1^l}|\widehat{x_2^l}|\ldots|\widehat{x_{q_l}^l}\right]$.
(6) The neighborhood relationships among the samples of $\widehat{X^{(l)}}$ form the final l-density-graph.

have also shown the computed sample means of the two classes in Fig. 1(a).

3.1.1. Newly defined intra-class scatter

We define a marginal intra-class scatter criterion and compare it with the scatter $S^{(w)}$ of LDA. We firstly consider a typical three-class case in Fig. 2(a), in which the class label information of data points and the corresponding class means are also shown. We see that the three means over all the intra-class data points are closely together due to the effect of faraway sample points. This phenomenon is rather common in real data, because most real data are rather noisy and filled with some redundant information.

Considering that the LDA intra-class scatter $S^{(w)}$ aims at minimizing the difference between each within-class point and their class mean by optimizing the criterion in Eq. (2). As a result, data points of different classes in Fig. 2(a) tend to be congregated when minimizing $S^{(w)}$, though maximizing the LDA inter-class scatter $S^{(b)}$ can improve the inter-class separation to some extent. Note that we will detail the disadvantages of optimizing $S^{(b)}$ in next section. According to Definitions 1 and 2 and our computational method of the class means, we can easily achieve the density-graph of each class and the density-region based sample means, which are shown in Fig. 2(b). We see clearly that the class means computed by using our method will not be affected by the faraway points or outliers, implying that our proposed criterion will be robust against the outliers. Actually, this observation is significantly important for subsequent classification process, because the decision boundary of classifier is usually sensitive to the outliers. And more importantly, the differences or margins between the class means are larger than those shown in Fig. 2(a). This will significantly contribute to characterizing the inter-class separability. Then the density-region based marginal intra-class scatter $L^{(w)}$ of LODA is defined as minimizing

$$L^{(w)} = \sum_{l=1}^{c} \frac{q_l}{N_l} \sum_{i=1}^{N_l} (x_i^l - \tilde{M}^{(l)})(x_i^l - \tilde{M}^{(l)})^{\mathrm{T}}. \tag{6}$$

Clearly, $L^{(w)}$ shares the same form as the scatter $S^{(w)}$ of LDA, but it is noted that the intra-class compactness in LODA is measured as the sum of distances between each within-class data point and its class mean computed from the density-region. It is worth noting that minimizing $L^{(w)}$ is also capable of achieving enhanced inter-class separation compared with minimizing scatter $S^{(w)}$ due to the fact that the margins of inter-class means have been significantly enlarged in the $L^{(w)}$ criterion. Then $L^{(w)}$ can be interpreted as the following matrix form:

$$\begin{aligned} L^{(w)} &= \sum_{l=1}^{c} \frac{q_l}{N_l} \sum_{i=1}^{N_l} \left[x_i^l - \frac{1}{q_l} \sum_{j=1}^{q_l} \widehat{x}_j^l \right] \left[x_i^l - \frac{1}{q_l} \sum_{j=1}^{q_l} \widehat{x}_j^l \right]^{\mathrm{T}} \\ &= \sum_{l=1}^{c} \frac{q_l}{N_l} \left[\sum_{i=1}^{N_l} x_i^l x_i^{l\mathrm{T}} - \frac{1}{q_l} \sum_{i=1}^{N_l} \sum_{j=1}^{q_l} x_i^l \widehat{x}_j^{l\mathrm{T}} - \frac{1}{q_l} \sum_{j=1}^{q_l} \sum_{i=1}^{N_l} \widehat{x}_j^l x_i^{l\mathrm{T}} \right. \\ &\quad \left. + \frac{N_l}{q_l^2} \sum_{i=1}^{q_l} \sum_{j=1}^{q_l} \widehat{x}_i^l \widehat{x}_j^{l\mathrm{T}} \right] \\ &= \sum_{l=1}^{c} \left[\frac{q_l}{N_l} X^{(l)} X^{(l)\mathrm{T}} - \frac{1}{N_l} \left(L_+^{(w)} + L_+^{(w)\mathrm{T}} \right) + \frac{1}{q_l} \widehat{X}^{(l)} \hat{e}^{(l)\mathrm{T}} \hat{e}^{(l)} \widehat{X}^{(l)^{\mathrm{T}}} \right], \end{aligned} \tag{7}$$

where $L_+^{(w)} = X^{(l)} e^{(l)^{\mathrm{T}}} \hat{e}^{(l)} \widehat{X}^{(l)^{\mathrm{T}}}$, $e^{(l)}$ is a $1 \times N_l$ vector of all ones, and $\hat{e}^{(l)}$ is a $1 \times q_l$ vector of all ones.

3.1.2. Newly defined inter-class scatter

Next we will detail the inter-class scatter $L^{(b)}$ for LODA. We begin with another three-class case in Fig. 3. We observe from Fig. 3(a) that the mean $\overline{M^{(t)}}$ of class t is far away from the other two means $\overline{M^{(l)}}$ and $\overline{M^{(r)}}$ that are close together. On one hand, by defining the intra-class density-regions in Fig. 3(b), we can characterize the inter-class separability with a marginal criterion. In another word, the sum of distances between the inter-class means in Fig. 3(b) is much larger than that of Fig. 3(a) by applying our definition method.

On the other hand, recall that the inter-class scatter $S^{(b)}$ in LDA is measured as the sum of distances between every class mean $\overline{M^{(j)}}$, $j=1, 2, \ldots, c$ and their global mean $\overline{M}$. It is important to notice that a large sum of the distances between $\overline{M^{(j)}}$ and $\overline{M}$, for instance in Fig. 3(a) is unable to guarantee that $\overline{M^{(l)}}$ and $\overline{M^{(r)}}$ can be mapped far away in the reduced space. As a result, minimizing $S^{(w)}$ simultaneously is equivalent to pushing data points of class l and class r closely together. This will directly result in a low inter-class separation and a high classification error; this is critical for performing discrimination. Note that the pairwise inter-class criteria [18,25,27,30] has been addressed to improve the scatter $S^{(b)}$. But these criteria still rely on class means of all the intra-class data points, making them suffer from the same problem as LDA does. That is to say that for real datasets, the intra-class mean is notably affected by the faraway points and then congregated. This work defines a new density-region based marginal inter-class

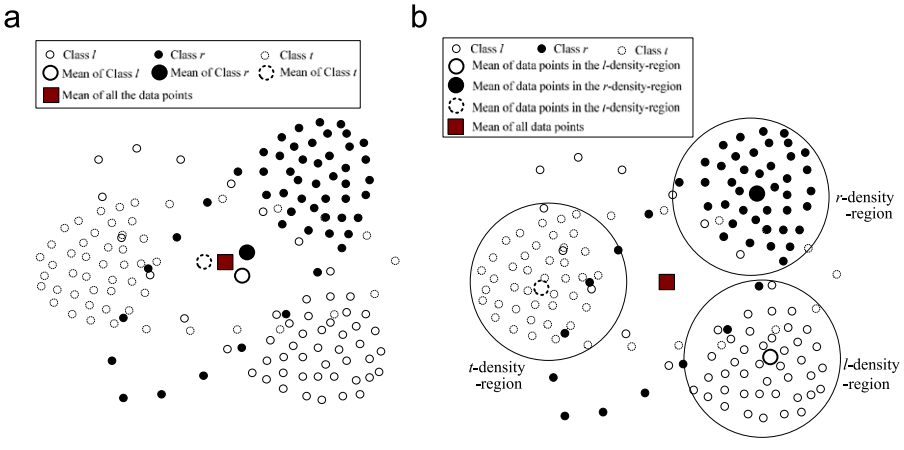

Fig. 2. Geometrical comparison of the intra-class scatter criteria of LDA (left) and LODA (right).

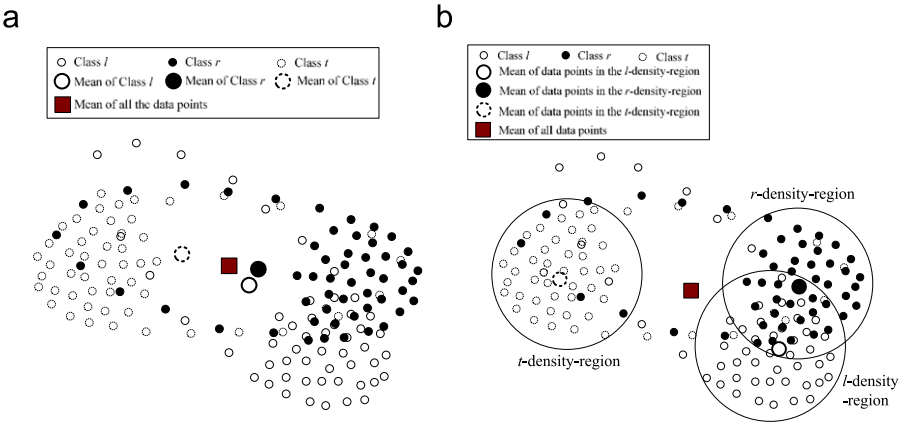

Fig. 3. Geometrical comparison of the inter-class scatter criteria of LDA (left) and LODA (right).

Table 2
Maximum margin criterion based linearly optimized discriminant analysis (LODA).

Input: Data points $\{(x_i,y_i)|x_i\in\Re^n,y_i(\in\{1,2,\ldots,c\})\}_{i=1}^N$
Dimensionality of embedding space $d(1\le d\le n)$
Output: $n\times d$ transformation matrix $\Xi=[\tilde{\varphi}_1|\tilde{\varphi}_2|\ldots|\tilde{\varphi}_d]$
(1) $X\leftarrow(x_1|x_2|\ldots|x_N)$; % data matrix.
(2) For each class $l\in\{1,\ldots,c\}$, using the procedures in Table 1 to construct the l-density-graph.
(3) Compute the sample mean $\tilde{M}^{(l)}$ of class l using the data points in $\widehat{X}^{(l)}$ of the l-density-region.
(4) $L^{(b)}=\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}\left[(q_\eta/q_l)\widehat{X}^{(l)}\hat{e}^{(l)\mathrm{T}}\hat{e}^{(l)}\widehat{X}^{(l)\mathrm{T}}-(L_+^{(b)}+L_+^{(b)\mathrm{T}})+(q_l/q_\eta)\widehat{X}^{(\eta)}\hat{e}^{(\eta)\mathrm{T}}\hat{e}^{(\eta)}\widehat{X}^{(\eta)\mathrm{T}}\right]$;
$L^{(w)}=\sum_{l=1}^{c}\left[(q_l/N_l)X^{(l)}X^{(l)\mathrm{T}}-(1/N_l)(L_+^{(w)}+L_+^{(w)T})+(1/q_l)\widehat{X}^{(l)}\hat{e}^{(l)}\mathrm{T}\hat{e}^{(l)}\widehat{X}^{(l)\mathrm{T}}\right]$.
(5) $\{\varphi_r,\widehat{\lambda_{[r]}}\}_{r=1}^d\leftarrow$ standard eigenvectors and eigenvalues of $(L^{(b)}-L^{(w)})\tilde{\varphi}=\tilde{\lambda}\tilde{\varphi}$, where $\widehat{\lambda}_{[1]}\ge\widehat{\lambda}_{[2]}\ge\cdots\ge\widehat{\lambda}_{[d]}$ and $\tilde{\varphi}_r{}^{\mathrm{T}}\tilde{\varphi}_r=1,\quad\tilde{\varphi}_r{}^{\mathrm{T}}\varphi_{r-1}^{\sim}=0,\quad$ for $\quad\forall r\in\{1,2,\ldots,d\}$. Output $\Xi=[\tilde{\varphi}_1|\tilde{\varphi}_2|\ldots|\tilde{\varphi}_d]$.

scatter $L^{(b)}$ to improve the above pairwise criteria by maximizing

$$L^{(b)}=\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}q_lq_\eta(\tilde{M}^{(l)}-\tilde{M}^{(\eta)})(\tilde{M}^{(l)}-\tilde{M}^{(\eta)})^{\mathrm{T}}. \tag{8}$$

That is, the inter-class separation in LODA is measured as the sum of the pairwise distances of the class means computed from the density-regions. By utilizing this definition, the rank limitation on the inter-class scatter is relaxed, implying that LODA is able to extract more meaningful features than LDA, especially for a large class number c. And most importantly, maximizing $L^{(b)}$ and minimizing $L^{(w)}$ simultaneously can guarantee enhanced intra-class compactness and inter-class separation because the difference or margin between each class mean and other class means have been significantly enlarged by LODA. Similarly, the symmetric marginal inter-class scatter matrix $L^{(b)}$ can also be described by using the following matrix form:

$$\begin{aligned}L^{(b)}&=\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}q_lq_\eta\left[\frac{1}{q_l}\sum_{i=1}^{q_l}\widehat{x_i^l}-\frac{1}{q_\eta}\sum_{j=1}^{q_\eta}\widehat{x_j^\eta}\right]\left[\frac{1}{q_l}\sum_{i=1}^{q_l}\widehat{x_i^l}-\frac{1}{q_\eta}\sum_{j=1}^{q_\eta}\widehat{x_j^\eta}\right]^{\mathrm{T}}\\&=\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}q_lq_\eta\left[\frac{1}{q_l^2}\sum_{i,j=1}^{q_l}\widehat{x_i^l}\widehat{x_j^l}^{\mathrm{T}}-\frac{1}{q_lq_\eta}\sum_{i=1}^{q_l}\sum_{j=1}^{q_\eta}\widehat{x_i^l}\widehat{x_j^\eta}^{\mathrm{T}}\right.\\&\quad\left.-\frac{1}{q_lq_\eta}\sum_{j=1}^{q_\eta}\sum_{i=1}^{q_l}\widehat{x_j^\eta}\widehat{x_i^l}^{\mathrm{T}}+\frac{1}{q_\eta^2}\sum_{i,j=1}^{q_\eta}\widehat{x_i^\eta}\widehat{x_j^\eta}^{\mathrm{T}}\right]\\&=\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}\left[\frac{q_\eta}{q_l}\widehat{X}^{(l)}\hat{e}^{(l)\mathrm{T}}\hat{e}^{(l)}\widehat{X}^{(l)^{\mathrm{T}}}-(L_+^{(b)}+L_+^{(b)\mathrm{T}})+\frac{q_l}{q_\eta}\widehat{X}^{(\eta)}\hat{e}^{(\eta)\mathrm{T}}\hat{e}^{(\eta)}\widehat{X}^{(\eta)^{\mathrm{T}}}\right]\end{aligned} \tag{9}$$

where matrices $L_+^{(b)}=\widehat{X}^{(l)}\tilde{W}^{(l\eta)}\widehat{X}^{(\eta)^{\mathrm{T}}}$ and $\tilde{W}^{(l\eta)}$ is a $q_l\times q_\eta$ matrix with all the elements equaling to one.

3.1.3. The objective function and solution

LODA solution scheme 1: To compute the transforming basis vectors $\{\zeta_r\}_{r=1}^d$ of LODA and avoid the matrix singularity, a *maximum margin criterion* (MMC) [18] based objective function for LODA is defined as

$$\underset{\Xi\in\Re^{n\times d}}{Max}\ Tr(\Xi^{\mathrm{T}}(L^{(b)}-L^{(w)})\Xi)\quad\text{subject to}\quad\Xi^{\mathrm{T}}\Xi=I. \tag{10}$$

By introducing the Lagrangian function $\hat{J}$ of Eq. (10) with the multiplier λ_i, we can then obtain

$$\hat{J}(\zeta_i,\lambda_i)=(L^{(b)}-L^{(w)})\zeta_i-\lambda_i(\|\zeta_i\|^2-1). \tag{11}$$

By taking the derivatives with respect to ζ_i and λ_i, and zeroing it, we can get the following typical eigenvalue problem: $(L^{(b)}-L^{(w)})\zeta_i{}^*=\lambda_i{}^*\zeta_i{}^*$, where λ_i^* and ζ_i^* are the standard eigenvalues and transforming basis vectors of the matrix $(L^{(b)}-L^{(w)})$, respectively. That is, the optimal projection matrix Ξ^* can be obtained by solving

$$\Xi^*=\underset{\Xi\in\Re^{n\times d},\Xi^{\mathrm{T}}\Xi=I}{\operatorname{argmax}}\ Tr(\Xi^{\mathrm{T}}(L^{(b)}-L^{(w)})\Xi). \tag{12}$$

From Eq. (12), the orthogonal projection matrix $\Xi=[\zeta_1|\zeta_2|\ldots|\zeta_d]$ can be analytically obtained. In this paper, we refer to the MMC based LODA as LODA. The algorithmic procedures of LODA are summarized in Table 2.

LODA solution scheme 2: Based on the newly defined intra-class and inter-class scatters $L^{(w)}$ and $L^{(b)}$, LODA can also be simply

defined as the following LDA-style trace ratio (TR) form problem:

$$\underset{\Xi \in \Re^{n\times d}}{Max}\ Tr(\Xi^{\mathrm{T}}L^{(b)}\Xi)/Tr(\Xi^{\mathrm{T}}L^{(w)}\Xi) \quad \text{subject to} \quad \Xi^{\mathrm{T}}\Xi = I. \tag{13}$$

Note that this TR problem is usually converted into the simplified ratio trace (RT) problem to find the optimal projection matrix by $\Xi^* = \mathrm{argmax}_{\Xi \in \Re^{n\times d}} Tr((\Xi^{\mathrm{T}}L^{(w)}\Xi)^{-1}(\Xi^{\mathrm{T}}L^{(b)}\Xi))$ as LDA. Then this problem can be easily solved by using the generalized eigen-decomposition approach, but the obtained solution may not optimal and does not necessarily best optimize the corresponding TR optimization problem [21]. In this paper, the iterative ITR [19,22] is employed to solve the TR problem of LODA. ITR tackles the TR problem in Eq. (13) by directly optimizing the objective $Tr(\Xi^{\mathrm{T}}L^{(b)}\Xi)/Tr(\Xi^{\mathrm{T}}L^{(w)}\Xi)$ under the assumption, i.e., column vectors of Ξ are orthogonal together. For given λ^v at each iteration v, Ξ^v can be obtained by the following trace difference (TD) problem:

$$\Xi^v = \mathrm{argmax}_{\Xi^{\mathrm{T}}\Xi = I} Tr(\Xi^{\mathrm{T}}(L^{(b)} - \lambda^v L^{(w)})\Xi), \tag{14}$$

and then ITR renews λ^{v+1} as TR value given by Ξ^v: $\lambda^{v+1} = Tr((\Xi^v)^{\mathrm{T}}L^{(b)}\Xi^v)/Tr((\Xi^v)^{\mathrm{T}}L^{(w)}\Xi^v)$ until convergence. Mathematical proof show that ITR can converge to the global optimum [19]. See the theoretical proofs of ITR in [19]. It is also noted that, under TR criterion, the orthogonal constraint $\Psi^{\mathrm{T}}\Psi = I$ is always imposed. In another word, the obtained projection matrix is also guaranteed to be orthogonal and the similarity between the data points can be efficiently preserved if it is based on Euclidean distance measure according to [19,22].

It is worth noting that, in ITR, Ξ^0 needs to be initialized as an arbitrary orthogonal matrix, which implies that ITR maybe unstable due to the randomness. And most importantly, the orthogonal initialized Ξ^0 is difficult to be constructed and a bad initialization may greatly increase the number of iterations in the optimization. In this paper, we initialize λ^0 instead of initializing Ξ^0 to be the orthogonal matrix as [22]. The solutions of LODA can then be effectively solved by this revised ITR. In summary, for the positive semi-definite matrices $L^{(b)}$ and $L^{(w)}$, the algorithmic procedures are described in Table 3. The construction steps of the scatters $L^{(w)}$ and $L^{(b)}$ are the same as Table 2. Similarly, this work refers to this TR criterion based LODA as TR-LODA.

3.1.4. Comparison between LODA, LDA and MMC

We mainly compare our proposed LODA algorithm with LDA and MMC from the following four aspects:

(1) LDA and MMC are based on utilizing all the data points to construct the intra-class scatter $S^{(w)}$ and inter-class scatter $S^{(b)}$, whilst the scatters $L^{(w)}$ and $L^{(b)}$ of LODA are based on the density-graphs and density-regions. As a result, our proposed criteria are more robust to the outliers in real applications than LDA and MMC criteria. It is noted that, for an evenly distributed dataset, degrees $\tilde{D}_i^l, i = 1,2,\ldots,N_l$ of all the vertices may be the same. Or when k value in NNS is set to N_l, the l-density-graph and l-density-region within class l will be consisted of all data points belonging to the class l. Or when $\beta = (\max(\tilde{D}_i^l) + \min(\tilde{D}_i^l))/\min(\tilde{D}_i^l)$, that is $X^{(l)} = \tilde{X}^{(l)}$. In these cases, the sample mean of class l can be calculated as $\tilde{M}^{(l)} = (1/N_l)\sum_{j=1}^{N_l} x_j^l$ which is equivalent to $\overline{M^{(l)}}$ in LDA, thus the intra-class scatter $L^{(w)}$ of LODA can then be transformed into

$$\begin{aligned} L^{\tilde{(w)}} &= \sum_{l=1}^{c}\sum_{i=1}^{N_l}\left[x_i^l - \frac{1}{N_l}\sum_{j=1}^{N_l}x_j^l\right]\left[x_i^l - \frac{1}{N_l}\sum_{j=1}^{N_l}x_j^l\right]^{\mathrm{T}} \\ &= \sum_{l=1}^{c}\sum_{i=1}^{N_l}(x_i^l - \overline{M^{(l)}})(x_i^l - \overline{M^{(l)}})^{\mathrm{T}} \end{aligned} \tag{15}$$

which is just the intra-class scatter $S^{(w)}$ of LDA. Similarly, the marginal scatter $L^{(b)}$ can be reformulated as

$$L^{\tilde{(b)}} = \sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c} N_l N_\eta (\overline{M^{(l)}} - \overline{M^{(\eta)}})(\overline{M^{(l)}} - \overline{M^{(\eta)}})^{\mathrm{T}}. \tag{16}$$

Clearly, this is just the pairwise inter-class scatter criterion introduced in [27,30]. Thus, the scatter criterion in Eq. (16) can be considered as a generalized version of our $L^{(b)}$ criterion. Especially, when data points of each class are evenly distributed (i.e. $\tilde{M}^{(l)} = \overline{M^{(l)}}$ for each class l) and the sample means of all classes are equal, namely

$$\tilde{M}^{(i)} = \tilde{M}^{(j)} = \overline{M^{(i)}} = \overline{M^{(j)}} = \overline{M}, \quad i = 1,2,\ldots,c-1, \quad j = i+1,2,\ldots,c. \tag{17}$$

Obviously, scatter $L^{\tilde{(b)}}$ will be equivalent to scatter $S^{(b)}$, i.e. $L^{\tilde{(b)}} = S^{(b)} = 0_{N\times N}$, where $0_{N\times N}$ is a $N\times N$ matrix with all zeros, implying that the LDA method is considered as a special case of our LODA. When the above two conditions are satisfied simultaneously, LODA can be reduced to MMC. Also, when k in NNS is set to N_l, that is the lower bound and upper bound of β are the same, thus MMC is also regarded as a special case of our LODA. It is also important to note that when condition in Eq. (17) is satisfied, LDA, MMC and our techniques will become inefficient in achieving inter-class separation, because there does not exists any inter-class difference. This is a drawback of our LODA, but it is noted that this condition is difficult to be satisfied in real applications. (2) There is no assumption on intra-class distribution in LODA. Without prior information on the data distributions, the inter-class margin can better characterize inter-class separability than the inter-class variance of LDA. These make LODA more general than applying LDA for discriminant analysis. (3) LODA is built on the TR criterion and MM criterion, thus the singularity problems can be effectively avoided by LODA because there is no need to compute the matrix inverse of any matrix. Moreover, the obtained projection matrix is always orthogonal under the two solution schemes. It should be noted that the LDA projection matrix is not orthogonal, thus the similarity may be changed if it is based on the Euclidean distance [19,22]. (4) Similar to LDA and MMC, the density- regions based LODA is also single-modal, thus they tend to map intra-class clusters into a single cluster in the real applications. We will in next section detail a natural multimodal extension of LODA, which we call multimodal LODA (MLODA), by defining new multimodal pairwise inter- and intra-class scatter criteria.

Table 3
TR criterion based robust linearly optimized discriminant analysis (TR-LODA).

(1) Initialize $\lambda^v = 0$, step $v = 0$.
(2) Solve the standard eigenvalue problem $(L^{(b)} - \lambda^v L^{(w)})\pi = \tau^v \pi$ and calculate the vectors $\{\pi_\delta^v\}_{\delta=1}^d$ of $(L^{(b)} - \lambda^v L^{(w)})$ by conducting the eigen-decomposition.
(3) Projection matrix $\Xi^v = \{\pi_\delta^v\}_{\delta=1}^d$ is formed by the eigenvectors according to the first d leading eigenvalues $\{\tau_\delta^v\}_{\delta=1}^d$ of the matrix $(L^{(b)} - \lambda^v L^{(w)})$.
(4) Renew λ^{v+1} by computing $Tr((\Xi^v)^{\mathrm{T}}L^{(b)}\Xi^v)/Tr((\Xi^v)^{\mathrm{T}}L^{(w)}\Xi^v)$.
(5) If $|\lambda^{v+1} - \lambda^v| < \varepsilon$, go to Step 6; else $v = v+1$, Steps 2–4 repeat.
(6) Output $\lambda^* = \lambda^v$, and $\Xi^* = \mathrm{argmax}_{\Xi^{\mathrm{T}}\Xi = I} Tr(\Xi^{\mathrm{T}}(L^{(b)} - \lambda^v L^{(w)})\Xi)$.

3.2. Multimodal extension of LODA

3.2.1. Multimodal intra-class and inter-class scatters

Intrinsic multimodal structure is common in real applications [11]. Thus, preserving the multimodal structures for dimensionality reduction and discriminant analysis is also an important issue that needs to be addressed. We first take the two-class case problem in Fig. 1(b) as an example, in which each class has two isolated clusters, i.e. *multimodal*. Accordingly the density-region within each class is consisted of two density sub-regions, because the density-graph within the same class may be disconnected due to multimodal distributions. Next we will extend LODA to multimodal scenarios for endowing LODA the capability to handle multimodal datasets.

Similarly, the density-graphs based multimodality preserving intra-class scatter criterion $ML^{(w)}$ of MLODA can be defined as the following pairwise description from $L^{(w)}$:

$$ML^{(w)} = \sum_{l=1}^{c} \frac{q_l}{N_l} \sum_{i=1}^{N_l} \sum_{j=1}^{q_l} \left(x_i^l - \widehat{x_j^l}\right)\left(x_i^l - \widehat{x_j^l}\right)^{\mathrm{T}} A_{i,j}^{(l)}, \tag{18}$$

which is minimized to measure the intra-class compactness. To preserve the multimodal structures of the datasets, weight matrix $A^{(l)}$ is added to represent the proximity around each data point $\widehat{x_i^l}$ in the l-density-region of class l. Note that $A^{(l)}$ can be similarly defined as the similarity matrix of *Locality Preserving Projections* (LPP) [23]. $A^{(l)}$ also aims at keeping the projections of all similarity neighboring pairs from the same class in close vicinity of the original space still close in the reduced space. As a result, the intrinsic multimodality can be effectively preserved. For the simple-minded method [10], $A_{i,j}^{(l)} = 1$ if x_i^l (or, $\widehat{x_i^l}$) and $\widehat{x_j^l}$ are mutually neighbors of a class, and $A_{i,j}^{(l)} = 0$ if x_i^l (or, $\widehat{x_i^l}$) is not a neighbor of $\widehat{x_j^l}$. Thus minimizing $ML^{(w)}$ is equivalent to improving the tightness of all the data points in the same class without losing multimodal distributions. Based on similar algebra computation, we can interpret the multimodal intra-class scatter $ML^{(w)}$ with the following matrix form:

$$\begin{aligned} ML^{(w)} &= \sum_{l=1}^{c} \frac{q_l}{N_l} \sum_{i=1}^{N_l} \sum_{j=1}^{q_l} (x_i^l - \widehat{x_j^l})(x_i^l - \widehat{x_j^l})^{\mathrm{T}} A_{i,j}^{(l)} \\ &= \sum_{l=1}^{c} \frac{q_l}{N_l} \left(\sum_{i=1}^{N_l} \left[\sum_j A_{i,j}^{(l)} \right] x_i^l x_i^{l\mathrm{T}} - \sum_{i=1}^{N_l} \sum_{j=1}^{q_l} x_i^l A_{i,j}^{(l)} \widehat{x_j^l}^{\mathrm{T}} \right. \\ &\quad \left. - \sum_{j=1}^{q_l} \sum_{i=1}^{N_l} \widehat{x_j^l} A_{j,i}^{(l)} x_i^{l\mathrm{T}} + \sum_{j=1}^{q_l} \left[\sum_i A_{i,j}^{(l)} \right] \widehat{x_j^l} \widehat{x_j^l}^{\mathrm{T}} \right) \\ &= \sum_{l=1}^{c} \left(\frac{q_l}{N_l} X^{(l)} \tilde{V}^l X^{(l)\mathrm{T}} - \left(L_-^{(w)} + L_-^{(w)\mathrm{T}} \right) + \frac{q_l}{N_l} \widehat{X}^{(l)} \tilde{U}^l \widehat{X}^{(l)^{\mathrm{T}}} \right), \end{aligned} \tag{19}$$

where $L_-^{(w)} = X^{(l)} Q^{(l)} \widehat{X^{(l)}}^{\mathrm{T}}$, $Q^{(l)}$ is a $N_l \times q_l$ matrix with entries $Q_{i,j}^{(l)} = (q_l/N_l) A_{i,j}^{(l)}$, $\tilde{U}^l$ and $\tilde{V}^l$ are diagonal matrices with entries being the column and row sums of $A^{(l)}$, respectively.

Analogous to the definition of pairwise inter-class criterion $L^{(b)}$, the density-region based marginal multimodal inter-class scatter criterion $ML^{(b)}$ for our MLODA can be formulated as the following pairwise form:

$$ML^{(b)} = \sum_{l=1}^{c-1} \sum_{\eta=l+1}^{c} \sum_{i=1}^{q_l} \sum_{j=1}^{q_\eta} (\widehat{x_i^l} - \widehat{x_j^\eta})(\widehat{x_i^l} - \widehat{x_j^\eta})^{\mathrm{T}} \tag{20}$$

which will be maximized to measure the inter-class separation. Clearly, $ML^{(b)}$ can be considered as a multimodal interpretation of the pairwise inter-class criteria [27,30]. But note that our scatter $ML^{(b)}$ is defined based on the density-regions rather than all the data points of each class. Clearly, the scatter matrix $ML^{(b)}$ is defined as the sum of pairwise distances between points $\widehat{x_i^l}$ and $\widehat{x_j^\eta}$ from l-density-region and η-density-region, respectively. Thus maximizing $ML^{(b)}$ can push points of different density-regions far apart without losing the multimodal structures in the dimension-reduced space. By using matrix interpretation, scatter $ML^{(b)}$ can be reformulated as

$$\begin{aligned} ML^{(b)} &= \sum_{l=1}^{c-1} \sum_{\eta=l+1}^{c} \sum_{i=1}^{q_l} \sum_{j=1}^{q_\eta} (\widehat{x_i^l} - \widehat{x_j^\eta})(\widehat{x_i^l} - \widehat{x_j^\eta})^{\mathrm{T}} \\ &= \sum_{l=1}^{c-1} \sum_{\eta=l+1}^{c} \left[q_\eta \sum_{i=1}^{q_l} \widehat{x_i^l} \widehat{x_i^l}^{\mathrm{T}} - \sum_{i=1}^{q_l} \sum_{j=1}^{q_\eta} \widehat{x_i^l} \widehat{x_j^\eta}^{\mathrm{T}} \right. \\ &\quad \left. - \sum_{j=1}^{q_\eta} \sum_{i=1}^{q_l} \widehat{x_j^\eta} \widehat{x_i^l}^{\mathrm{T}} + q_l \sum_{j=1}^{q_\eta} \widehat{x_j^\eta} \widehat{x_j^\eta}^{\mathrm{T}} \right] \\ &= \sum_{l=1}^{c-1} \sum_{\eta=l+1}^{c} \left[q_\eta \widehat{X}^{(l)} \widehat{X}^{(l)^{\mathrm{T}}} - (L_-^{(b)} + L_-^{(b)\mathrm{T}}) + q_l \widehat{X}^{(\eta)} \widehat{X}^{(\eta)^{\mathrm{T}}} \right], \end{aligned} \tag{21}$$

where $L_-^{(b)} = \widehat{X}^{(l)} \tilde{Q}^{(l\eta)} \widehat{X}^{(\eta)^{\mathrm{T}}}$ and $\tilde{Q}^{(l\eta)}$ is a $q_l \times q_\eta$ matrix with all entries $\tilde{Q}_{i,j}^{(l\eta)} = 1$. Clearly, $ML^{(b)}$ is symmetric.

3.2.2. The objective function and solution

Similar to LODA, maximizing marginal $ML^{(b)}$ and minimizing marginal $ML^{(w)}$ can exhibit enhanced intra-class compactness and inter-class separation. Based on the scatters $ML^{(w)}$ and $ML^{(b)}$, the objective function of MLODA can be similarly formulated as the following TR criterion and MM criterion based problems:

$$J_{TR-MLODA}(\Xi) = \max_{\Xi \in \Re^{n \times d}} Tr(\Xi^{\mathrm{T}} ML^{(b)} \Xi) / Tr(\Xi^{\mathrm{T}} ML^{(w)} \Xi) \quad \text{subject to} \quad \Xi^{\mathrm{T}} \Xi = I, \tag{22}$$

$$J_{MLODA}(\Xi) = \max_{\Xi \in \Re^{n \times d}} Tr(\Xi^{\mathrm{T}} (ML^{(b)} - ML^{(w)}) \Xi) \quad \text{subject to} \quad \Xi^{\mathrm{T}} \Xi = I. \tag{23}$$

Similarly, we refer to TR criterion based robust MLODA as TR-MLODA and MM criterion based MLODA as MLODA. Notice that the TR optimization process of TR-MLODA is similar in spirit to that of optimizing the TR-LODA criterion. Detailed computational issue will not provided due to the page limitation.

The transforming basis vectors of MLODA can be similarly obtained as LODA by solving a standard eigen- problem. Let $\{\zeta_r\}_{r=1}^{d}$ be the eigenvectors, ordered according to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, we can then take as transforming axes of MLODA the eigenvectors corresponding to maximum eigenvalues of the following eigen-value problem: $(ML^{(b)} - ML^{(w)})\zeta_j = \lambda_j \zeta_j$. Then a solution of MLODA is analytically given as $\Xi = (\zeta_1 | \zeta_2 | \ldots | \zeta_d)$.

It is important to note that the processes of computing the projection axes of MLODA and TR-MLODA are also stable as the computation processes can avoid the matrix inverse operation. Most importantly, the obtained projection matrices are orthogonal. The algorithmic procedures of obtaining the transforming basis vectors from MLODA and TR-MLODA can be similarly implemented as Tables 1 and 2.

3.2.3. Comparison with LODA, LDA, MMC, PCA, LPP and LFDA

Comparison with LDA, MMC and LODA: MLODA is a natural extension of LODA, so they share some common advantages. The margins of inter-class clusters and intra-class clusters can also be enlarged by MLODA and the compactness of intra-class points can be shrunk at the same time. Also, there is no assumption on intra-class data distributions in MLODA. Different from LDA, MMC and LODA, MLODA can avoid computing the intra-class means and global mean. Thus, for a multimodally distributed dataset, using the MLODA criterion will not project the intra-cluster data

points into a single cluster. Also, MLODA is robust against the outliers as LODA.

Connection with PCA, LPP and LFDA: For an evenly distributed dataset, degrees $\tilde{D}_i^l, i=1,2,\ldots,N_l$ of all the vertices may be the same, or when $\beta=(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/(\min(\tilde{D}_i^l))$. Thus the density-region within each class contains all the data points of a class. As a result, the sample set $X^{(l)}$ out of the l-density-region is empty and set X^l will be generalized to X^l, then the intra-scatter matrix $ML^{(w)}$ can be transformed into

$$\tilde{M}L^{(w)}=\sum_{l=1}^{c}\sum_{i,j=1}^{N_l}(x_i^l-x_j^l)(x_i^l-x_j^l)^{\mathrm{T}}A_{i,j}^{(l)}, \tag{24}$$

which can be considered as a supervised weighted metric for LPP. It is also noted that when $A_{i,j}^{(l)}$ is similarly defined as [20], $\tilde{M}L^{(w)}$ is equivalent to the local scatter matrix of the LFDA algorithm.

It is noted that

$$\sum_{l>\eta}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}+\sum_{l<\eta}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}$$
$$=\sum_{l=1}^{c}\sum_{\eta=1}^{c}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}.$$

Also because

$$\sum_{l=\eta}^{c}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}=\sum_{l=\eta}^{c}\sum_{i,j=1}^{N_l}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}=0_{N\times N},$$
$$\sum_{l>\eta}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}=\sum_{l<\eta}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}$$

the multimodal inter-class scatter $ML^{(b)}$ can be converted into the following pairwise form:

$$\begin{aligned}\tilde{M}L^{(b)}&=\frac{1}{2}\sum_{l=1}^{c}\sum_{\eta=1}^{c}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}(x_i^l-x_j^\eta)(x_i^l-x_j^\eta)^{\mathrm{T}}\\&=\frac{1}{2}\sum_{l=1}^{c}\sum_{\eta=1}^{c}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}\left[x_i^lx_i^{l\mathrm{T}}-x_i^lx_j^{\eta\mathrm{T}}-x_j^\eta x_i^{l\mathrm{T}}+x_j^\eta x_j^{\eta\mathrm{T}}\right]\\&=\sum_{l=1}^{c}\sum_{\eta=1}^{c}\left[\frac{N_\eta}{2}\sum_{i=1}^{N_l}x_i^lx_i^{l\mathrm{T}}-\frac{1}{2}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}x_i^lx_j^{\eta\mathrm{T}}\right.\\&\quad\left.-\frac{1}{2}\sum_{j=1}^{N_\eta}\sum_{i=1}^{N_l}x_j^\eta x_i^{l\mathrm{T}}+\frac{N_l}{2}\sum_{j=1}^{N_\eta}x_j^\eta x_j^{\eta\mathrm{T}}\right]\\&=\left[\sum_{\eta=1}^{c}\frac{N_\eta}{2}\right]\sum_{l=1}^{c}\sum_{i=1}^{N_l}x_i^lx_i^{l\mathrm{T}}-\frac{1}{2}\sum_{l=1}^{c}\sum_{\eta=1}^{c}\sum_{i=1}^{N_l}\sum_{j=1}^{N_\eta}x_i^lx_j^{\eta\mathrm{T}}\\&\quad-\frac{1}{2}\sum_{l=1}^{c}\sum_{\eta=1}^{c}\sum_{j=1}^{N_\eta}\sum_{i=1}^{N_l}x_j^\eta x_i^{l\mathrm{T}}+\left[\sum_{l=1}^{c}\frac{N_l}{2}\right]\sum_{\eta=1}^{c}\sum_{j=1}^{N_\eta}x_j^\eta x_j^{\eta\mathrm{T}}\\&=\frac{N}{2}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}-\frac{1}{2}\sum_{l=1}^{c}\sum_{i=1}^{N_l}x_i^l\sum_{\eta=1}^{c}\sum_{j=1}^{N_\eta}x_j^{\eta\mathrm{T}}\\&\quad-\frac{1}{2}\sum_{\eta=1}^{c}\sum_{j=1}^{N_\eta}x_j^\eta\sum_{l=1}^{c}\sum_{i=1}^{N_l}x_i^{l\mathrm{T}}+\frac{N}{2}\sum_{j=1}^{N}x_jx_j^{\mathrm{T}}\\&=N\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}-\frac{1}{2}\sum_{l=1}^{c}\sum_{i=1}^{N_l}x_i^l\sum_{j=1}^{N}x_j^{\mathrm{T}}-\frac{1}{2}\sum_{\eta=1}^{c}\sum_{j=1}^{N_\eta}x_j^\eta\sum_{i=1}^{N}x_i^{\mathrm{T}}\\&=N\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}-\sum_{i,j=1}^{N}x_ix_j^{\mathrm{T}}=\sum_{i,j=1}^{N}x_ix_i^{\mathrm{T}}-\sum_{i,j=1}^{N}x_ix_j^{\mathrm{T}}=NS^{(t)}.\end{aligned}\tag{25}$$

Clearly, Eq. (25) is just the total scatter matrix of the PCA criterion, thus maximizing $\Xi^{\mathrm{T}}ML^{(b)}\Xi$ is equivalent to optimizing the PCA criterion.

3.3. Kernelized LODA and MLODA for nonlinear dimensionality reduction

This section considers extending LODA and MLODA to the nonlinear scenarios by employing the standard kernel trick [31]. Kernelized LODA and MLODA find matrix $\Upsilon=[\gamma_1,\gamma_2,\ldots,\gamma_d]$ for projections. Let ϕ be the mapping from $\Re^n$ to a higher-dimensional space $\mathbb{Z}^p(p\gg n)$. This mapping can be implicitly defined by using a kernel function. More specifically, the (i,j)th entry of a kernel matrix K is given by $K_{ij}=K(x_i,x_j)=\phi(x_i)^{\mathrm{T}}\phi(x_j)$. The Gaussian RBF kernel [31] with parameter σ, a typical choice of the kernel function, is defined as

$$K(x_i,x_j)=\exp(-||x_i-x_j||^2/2\sigma^2). \tag{26}$$

We take the computational formulation of kernelizing LODA as an example. Rewriting every ζ in $\mathbb{Z}^p$ as an expansion in terms of the mapped training data points, that is $\zeta=\sum_{i=1}^{N}\gamma_i\phi(x_i)=\phi(X)\gamma$, then scatters $(\Xi^{\mathrm{T}}L^{(w)}\Xi)^\phi$ and $(\Xi^{\mathrm{T}}L^{(b)}\Xi)^\phi$ of LODA in the kernel feature space can be written as

$$(\Xi^{\mathrm{T}}L^{(w)}\Xi)^\phi=\Xi^{\phi\mathrm{T}}\sum_{l=1}^{c}\left[\frac{q_l}{N_l}\phi(X^{(l)})\phi(X^{(l)})^{\mathrm{T}}-(L_+^{(w)}+L_+^{(w)\mathrm{T}})^\phi\right.$$
$$\left.+\frac{1}{q_l}\phi(\widehat{X}^{(l)})\widehat{E}^{(l)}\phi(\widehat{X}^{(l)})^{\mathrm{T}}\right]\Xi^\phi, \tag{27}$$

$$(\Xi^{\mathrm{T}}L^{(b)}\Xi)^\phi=\Xi^{\phi\mathrm{T}}\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}\left[\frac{q_\eta}{q_l}\phi\left(\widehat{X}^{(l)}\right)\widehat{E}^{(l)}\phi\left(\widehat{X}^{(l)}\right)^{\mathrm{T}}\right.$$
$$\left.-(L_+^{(b)}+L_+^{(b)\mathrm{T}})^\phi+\frac{q_l}{q_\eta}\phi\left(\widehat{X}^{(\eta)}\right)\widehat{E}^{(\eta)}\phi\left(\widehat{X}^{(\eta)}\right)^{\mathrm{T}}\right]\Xi^\phi, \tag{28}$$

where $(L_+^{(w)})^\phi=\phi(X^{(l)})(W^{(l)})^\phi\phi(\widehat{X}^{(l)})^{\mathrm{T}}$, $(L_+^{(b)})^\phi=\phi(\widehat{X}^{(l)})\widehat{e}^{(l)\mathrm{T}}\widehat{e}^{(\eta)}\phi\left(\widehat{X}^{(\eta)}\right)^{\mathrm{T}}$, $\widehat{E}^{(l)}=\widehat{e}^{(l)\mathrm{T}}\widehat{e}^{(l)}$ for each class l and $(W^{(l)})^\phi$ is a $N_l\times q_l$ matrix with input elements $(W_{i,j}^{(l)})^\phi=(1/N_l)$. By substituting the matrix inner product into Eqs. (27) and (28), we have the following equivalent form for Eq. (27):

$$(\Xi^{\mathrm{T}}L^{(w)}\Xi)^\phi=\Upsilon^{\mathrm{T}}\sum_{l=1}^{c}\left[\frac{q_l}{N_l}K^{(l)}K^{(l)\mathrm{T}}-(K_+^{(w)}+K_+^{(w)\mathrm{T}})+\frac{1}{q_l}\widehat{K_\alpha^{(l)}}\widehat{E}^{(l)}\widehat{K_\alpha^{(l)}}^{\mathrm{T}}\right]\Upsilon, \tag{29}$$

where $K_+^{(w)}=K^{(l)}(W^{(l)})^\phi\widehat{K_\alpha^{(l)}}^{\mathrm{T}}$, $K^{(l)}=\phi(X)^{\mathrm{T}}\phi(X^{(l)})$ and $\widehat{K}_\alpha^{(l)}=\phi((X)^{\mathrm{T}}\phi(\widehat{X}^{(l)})$ of class l is computed as the kernel Gram matrix over data matrix consisting of data points in the l-density-region. Similarly

Fig. 4. Typical sample images of '0'–'9' from the real USPS handwritten digits database.

we can obtain

$$(\Xi^{\mathrm{T}}L^{(b)}\Xi)^{\phi} = \Upsilon^{\mathrm{T}}\sum_{l=1}^{c-1}\sum_{\eta=l+1}^{c}\left[\frac{q_\eta}{q_l}\widehat{K_\alpha^{(l)}}\widehat{E}^{(l)}\widehat{K_\alpha^{(l)}}^{\mathrm{T}} - (K_+^{(b)} + K_+^{(b)\mathrm{T}}) + \frac{q_l}{q_\eta}\widehat{K_\alpha^{(\eta)}}\widehat{E}^{(\eta)}\widehat{K_\alpha^{(\eta)}}^{\mathrm{T}}\right]\Upsilon, \tag{30}$$

where $K_+^{(b)} = \widehat{K}_\alpha^{(l)}\widehat{e}^{(l)\mathrm{T}}\widehat{e}^{(\eta)}\widehat{K}_\alpha^{(\eta)\mathrm{T}}$. Note that the kernel matrices $\widehat{K}_\alpha^{(l)} = \phi(X)^{\mathrm{T}}\phi(\widehat{X}^{(l)})$ and $\widehat{K}_\alpha^{(\eta)} = \phi(X)^{\mathrm{T}}\phi(\widehat{X}^{(\eta)})$ are all defined based on the density regions. By substituting Eqs. (29) and (30) into the problems of Eqs. (10) and (13), the orthogonal transforming axes of the kernelized LODA can be similarly obtained. It is important to notice that MLODA can be similarly extended into the nonlinear

Table 4
Parameter analysis results on the real USPS handwritten digits database.

Value	Digital class l									
	1	2	3	4	5	6	7	8	9	10
(a) $k=p/6+1$										
$\max(\tilde{D}_i^l)$	41	48	45	50	38	36	40	44	37	43
$\min(\tilde{D}_i^l)$	20	20	20	20	20	20	20	20	20	20
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\max(\tilde{D}_i^l)$	1.49	1.42	1.44	1.40	1.53	1.56	1.50	1.45	1.54	1.47
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\min(\tilde{D}_i^l)$	3.05	3.40	3.25	3.50	2.90	2.80	3.00	3.20	2.85	3.15
Mean	2.27	2.41	2.35	2.45	2.22	2.18	2.25	2.33	2.20	2.31
(b) $k=p/4+1$										
$\max(\tilde{D}_i^l)$	59	74	63	88	54	57	55	63	54	60
$\min(\tilde{D}_i^l)$	30	30	30	30	30	30	30	30	30	30
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\max(\tilde{D}_i^l)$	1.51	1.41	1.48	1.35	1.56	1.53	1.56	1.48	1.56	1.50
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\min(\tilde{D}_i^l)$	2.97	3.47	3.10	3.93	2.80	2.90	2.83	3.10	2.80	3.00
Mean	2.24	2.44	2.29	2.64	2.18	2.22	2.20	2.29	2.18	2.25
(c) $k=p/2+1$										
$\max(\tilde{D}_i^l)$	104	110	114	119	114	107	113	109	113	112
$\min(\tilde{D}_i^l)$	60	60	60	60	60	60	60	60	60	60
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\max(\tilde{D}_i^l)$	1.58	1.55	1.53	1.51	1.53	1.56	1.53	1.55	1.54	1.54
$(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\min(\tilde{D}_i^l)$	2.73	2.83	2.90	2.98	2.90	2.78	2.88	2.82	2.88	2.87
Mean	2.15	2.19	2.22	2.25	2.22	2.17	2.21	2.19	2.21	2.21

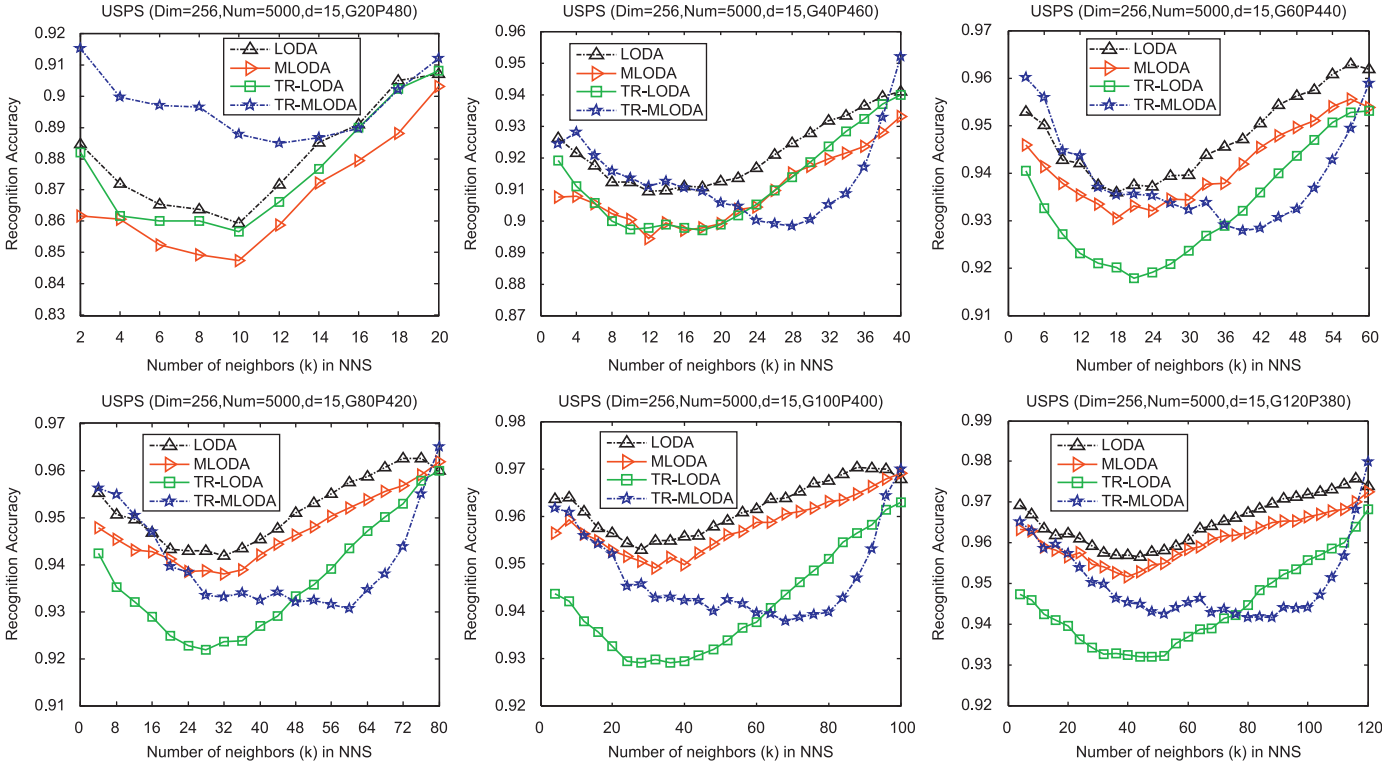

Fig. 5. Recognition accuracies vs. number of neighbors (k) on the USPS handwritten digits database.

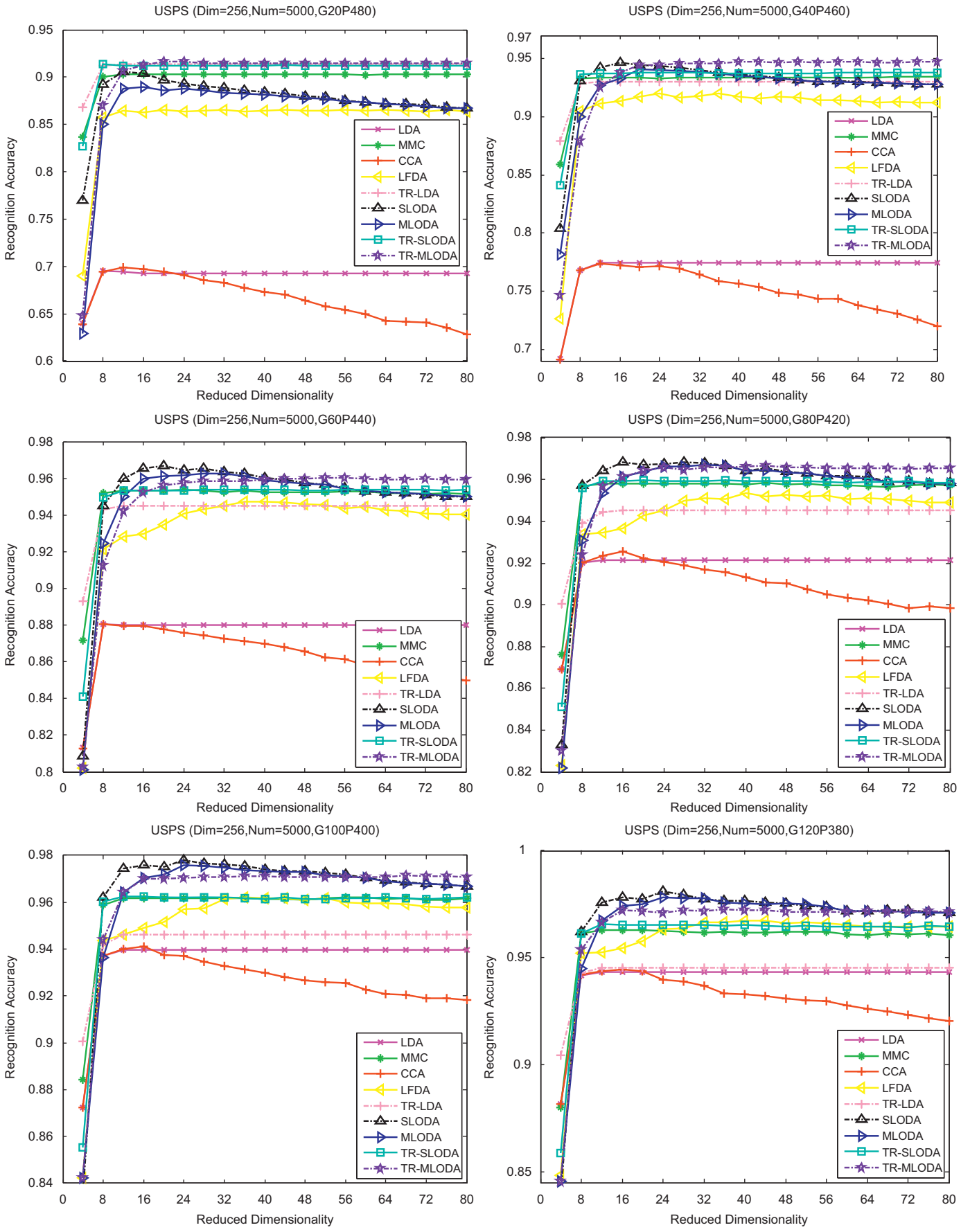

Fig. 6. Recognition accuracies vs. number of reduced dimensions on the USPS handwritten digits database.

scenarios as kernelizing LODA, but detailed computational formulations will not be provided because of page limitation. It is also noted that LODA and MLODA can be kernelized by applying the KPCA-trick framework [33,35]. See details in [33,35]. By utilizing the KPCA-trick, linear LODA (or, MLODA) can be kernelized directly using a two-stage procedure, i.e. KPCA plus LODA (or, MLODA). Because the size of matrices to be eigen-decomposed in those kernelized methods depends on the number of data points, kernelized formulations can improve the computational efficiency when the sample size is smaller than its input dimensionality of the original space. But it is noted that kernelized methods heavily depend on the kernel family, including kernel function and kernel parameter, due to the fact that different kernel functions produce different mappings and properties [31]. To data there is still no theoretical guarantee of optimal selection of kernels, thus in this paper we mainly focus on evaluating the proposed linear methods.

4. Simulation results and analysis

We in this section conduct extensive simulations, including handwritten digits recognition and object recognition, to verify the validity of our presented methods. We compare the recognition accuracy rates of our LODA family method (i.e., LODA and TR-LODA) and MLODA family (i.e., MLODA and TR-MLODA) with six widely used techniques, including LDA, MMC, CCA [28], LFDA, ITR algorithm based LDA (TR-LDA) [19,22] and PCA. It is noted that the CCA method with the c label coding [28] used in our simulations is the multi-class extension of the binary-class CCA [29]. For classification, the one-nearest-neighbor (1NN) classifier with Euclidean metric is used to avoid the bias caused by the choice of the learning methods. All the used algorithms are implemented in MATLAB 7.1. For all the semi-definite matrix inverse operation and generalized eigen-decomposition involved methods, the regularization factor μ is determined by the 10-fold cross validation methodology. We perform all the simulations on a PC with Intel (R) Core (TM) i5 CPU 650 at 3.20 GHz 3.19 GHz 4 G. In this study, two real databases are evaluated. The first one is the real USPS handwritten digits database [17] and the second one is the real ETH80 object recognition database [24]. The training set will be preliminarily processed by PCA operator to eliminate the null space before dimensionality reduction. For image recognition, after a nearest neighbor classifier (or, learner) is obtained from the training set, the test image data will be projected in the feature space by using the dimensionality reduction matrix learned from training data. The obtained learner is then used for evaluating the recognition accuracies of the test set in the dimension-reduced embedding space.

4.1. Handwritten digits recognition on USPS database

In this section, the USPS handwritten digits database [17] is applied to test our proposed methods. In this study, the publically available handwritten digit set from *http://cs.nyu.edu/~roweis/*

Table 5
Performance comparisons on the USPS handwritten digits database.

Method	Result Simulation setting							
	USPS digit database (Dim=256,G20/P480)				*USPS digit database (Dim=256,G40/P460)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.6938	0.6987	8	0.1686	0.7714	0.7749	8	0.2577
MMC	0.8996	0.9035	72	0.1125	0.9301	0.9345	72	0.1795
CCA	0.6659	0.6992	12	0.1439	0.7489	0.7734	12	0.2282
LFDA	0.8555	0.8657	44	0.2087	0.9053	0.9201	24	0.2456
TR-LDA	0.9087	0.9140	12	0.1402	0.9242	0.9302	12	0.2203
LODA	0.8776	0.9059	12	0.1741	0.9288	0.9470	16	0.2306
MODA	0.8646	0.8894	16	0.1742	0.9210	0.9407	20	0.2322
TR-LODA	0.9082	0.9139	8	0.1897	0.9328	0.9383	60	0.2545
TR-MLODA	0.8989	0.9167	24	0.1887	0.9281	0.9474	80	0.2529
	USPS digit database (Dim=256,G60/P440)				*USPS digit database (Dim=256,G80/P420)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.8778	0.8805	8	0.3349	0.9179	0.9216	8	0.4042
MMC	0.9485	0.9536	24	0.2536	0.9536	0.9585	44	0.3180
CCA	0.8640	0.8804	8	0.3063	0.9091	0.9255	16	0.3784
LFDA	0.9336	0.9473	40	0.5354	0.9416	0.9534	40	0.6036
TR-LDA	0.9390	0.9457	32	0.2956	0.9400	0.9472	36	0.3630
LODA	0.9499	0.9666	20	0.6354	0.9569	0.9686	16	0.8306
MODA	0.9470	0.9628	28	0.6526	0.9535	0.9670	32	0.8474
TR-LODA	0.9476	0.9539	68	0.7190	0.9536	0.9597	36	0.9054
TR-MLODA	0.9477	0.9602	53	0.7207	0.9561	0.9666	44	0.9214
	USPS database (Dim=256,G100/P400)				*USPS database (Dim=256,G120/P380)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.9367	0.9398	8	0.4225	0.9393	0.9433	8	0.4061
MMC	0.9576	0.9623	56	0.3435	0.9577	0.9629	12	0.3406
CCA	0.9260	0.9410	16	0.3989	0.9302	0.9446	16	0.3885
LFDA	0.9514	0.9623	36	0.7224	0.9574	0.9673	40	0.3365
TR-LDA	0.9423	0.9493	32	0.3878	0.9425	0.9502	32	0.3798
LODA	0.9653	0.9782	24	0.9535	0.9677	0.9811	24	0.3071
MODA	0.9628	0.9758	24	0.9806	0.9660	0.9780	24	0.3164
TR-LODA	0.9565	0.9625	12	1.0314	0.9595	0.9656	12	0.3275
TR-MLODA	0.9627	0.9713	68	1.0491	0.9642	0.9725	36	0.3315

data.html is used in our simulations. This sample set includes 16×16 pixels in 8-bit grayscale images of '0' through '9'. Each digit has 1100 images. We show some sample typical images of '0' to '9'in Fig. 4. In this simulation, we select 500 samples from each digit character (totally 5000 examples) for the experiments. Four simulation settings under different numbers of training data samples are tested. The sampled image set is partitioned into different galleries and probe sets, where G_P/P_q means p images per individual accompanied with the underlying class labels are randomly selected for training the learner and the remaining q sample images are used for testing the accuracies. In this simulation, six experimental configurations based on different G_P/P_q, $p=20$, 40, 60, 80, 100 and 120 are evaluated.

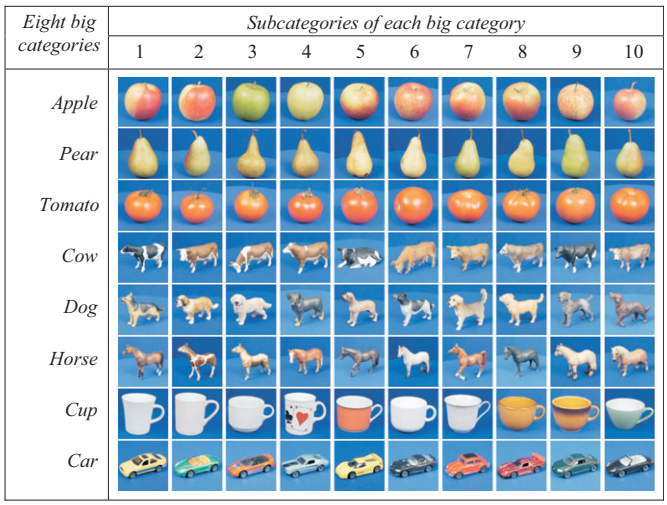

Fig. 7. The categories of the ETH-80 database and each row describes one big category of this database.

4.1.1. Parameter selection of β in density-graph construction

We first investigate the selection of the parameter β. We take the case of G_P/P_q, $p=120$, $q=380$, as an example. Three cases, including $k=p/6+1$, $k=p/4+1$ and $k=p/2+1$ are tested. According to the lower bound and upper bound of β in Eq. (5), for each class l, we report the maximal and minimal values of $\tilde{D}_i^l$, lower and upper bounds of parameter β in Table 4 after the neighborhood graph each class l is constructed. We also record the values of $(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\max(\tilde{D}_i^l)$ and $(\max(\tilde{D}_i^l)+\min(\tilde{D}_i^l))/\min(\tilde{D}_i^l)$ in the bottom of Table 4. Observing from Table 4, we find that the lower bound of β is around 1.5 and the upper bound is around 3. It is also interesting to find that the means are all close to the constant 2 for all the tested cases. For other investigated cases over different G_P/P_q and databases, we can find similar observations. Thus in this present work, the selection of the controlling parameter β is referred to the mean value and is set to 2 in all the simulations.

4.1.2. Performance analysis of k in NNS

We investigate the impact of the number of neighbors, k, involved in NNS on the recognition performance of our proposed

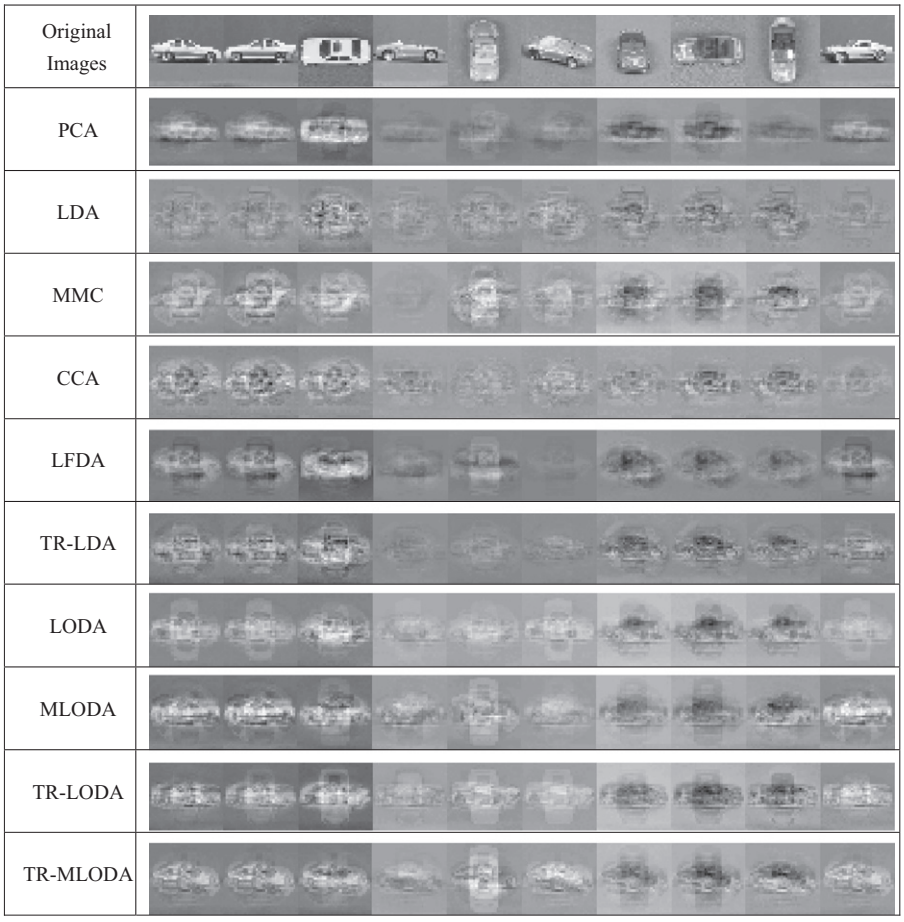

Fig. 8. Visualization of the transforming matrices of PCA, LDA, MMC, CCA, LFDA, TR-LDA and our proposed methods on the car object category from the ETH80 database.

methods. For each setting over G_P/P_q, we first fix the number of training images of each digit and then vary the number of k. For each k, we average the recognition accuracies over 20 random splits of training and test samples. The results are reported in Fig. 5. We see from Fig. 5 that the recognition performance of our methods varies with the increasing number of k. In particular, our methods exhibit an unusual behavior, that is the result initially decreases with the k value and then starts going up after some point till reaching the highest record. According to the observed results, a relatively large k tend to improve the accuracy, therefore the number of k is always set to p-2 for the following handwritten digits recognition simulations.

4.1.3. Handwritten digits recognition

This study aims at testing the LDA, TR-LDA, MMC, LFDA, CCA and our proposed algorithms for recognizing the handwritten digits. In order to ensure that our observed recognition results are not biased from a specific random realization of the training/test set, for each G_P/P_q, we compute the averaged accuracy over 20 random splits of training/test sets. The recognition results are illustrated in Fig. 6, where *Dim* is the dimensionality of the original digital image space and *Num* is the total number of data examples.

Observing from Fig. 6, we can obtain the following conclusions. First, the performance of each method varies with the number change of reduced dimensions. Also, the recognition accuracy of each method increases with the increasing training sample size. Second, our family methods, including LODA, MLODA, TR-LODA and TR-MLODA, deliver the comparable or even better results to TR-LDA and MMC in most cases. Specially, our methods outperform the remaining methods in the latter three settings. The results of LDA and CCA are very close in each setting due to intrinsic relationships between them [28] and both are inferior to the other methods. LFDA also works well in most cases and outperforms the LDA, TR-LDA and CCA methods for the latter three cases. Third, due to the reasonably defined marginal scatter criteria, our investigated cases indicate that our proposed family methods tend to outperform the fully supervised LDA, CCA, LFDA, MMC and TR-LDA.

The mean and the highest accuracies according to the results of Fig. 6 are recorded in Table 5. The averaged running time

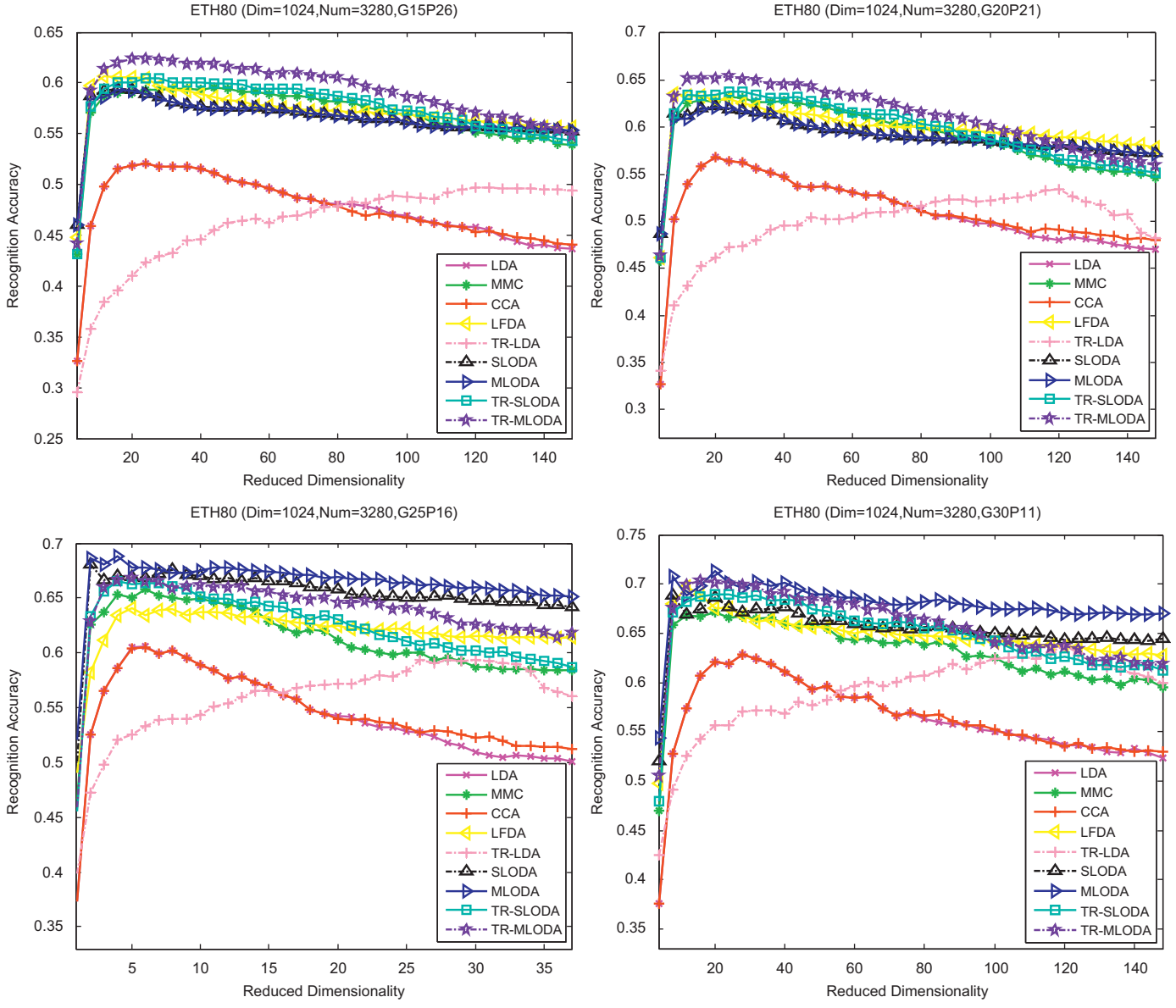

Fig. 9. Recognition accuracies vs. number of reduced dimensions on the ETH80 object database (80 classes).

(including the training and testing time) computed in seconds and the standard deviations over the repetitions are also reported. The best subspaces corresponding to the best accuracies are called the optimal digital image subspaces. Observing the test results, we can find that: (1) in most cases, our proposed methods exhibit the highest accuracies, compared with the other methods. The mean accuracies delivered by our methods are better than those obtained by other methods in most cases. (2) For each evaluated method, the standard deviations computed over repetitions are comparable with each other. (3) Considering the running time performance, our proposed methods are comparable to the remaining methods in most cases.

4.2. Object recognition on ETH80 database

This study addresses an object categorization task using the benchmark ETH80 object recognition database [24]. This database contains images of 8 big categories: *apple*, *car*, *cow*, *cup*, *dog*, *horse*, *pear* and *tomato*. In each big category, there are 10 subcategories, each of which contains 41 images from different viewpoints. Overall, the database contains 3280 images of 80 objects. Each image has 128×128 pixels. In our simulations, we resize each of the images into a size of 32×32 pixels. Each pixel will be considered as an input variable and thus each image corresponds to a data point in a 1024-dimensional space. We show some typical sample images from the ETH80 database in Fig. 7, where we also describe the 8 big categories and 10 subcategories within each big category.

4.2.1. Visualization of transforming matrix

We first examine the visual properties of the transforming matrices obtained by our proposed methods. The performance of our methods is compared with PCA, LDA, MMC, CCA, TR-LDA and LFDA. In this simulation, the car category with 10 objects is tested and each object corresponds to a single class. Then a ten-class case is created. For each method, we randomly select 4 images from each object for learning the optimal object image subspaces. We illustrate the first 10 eigenvectors (or called eigen-pictures) of the transforming matrix obtained by each method. The eigen-pictures are then reshaped into a matrix according to the original object image size, i.e. 32×32. The computed eigen-pictures are exhibited in Fig. 8. We observe from Fig. 4 that the eigen-pictures obtained by LDA, MMC, CCA, TR-LDA, LFDA and our methods are more noisy compared with PCA, which demonstrates that they are capable of capturing more disciminant information about object image details.

4.2.2. Object recognition of 80 categories

In this subsection, we focus on representing and recognizing the object images from the ETH80 database. In this study, each subcategory of the 8 big categories is regarded as a single class. As a result, an 80-class classification problem is then created. In our simulation, four settings over different training sample sizes are evaluated. From the study, similar observation trend of the number of k is found and k value is also set to p-2 for the simulations.

The recognition results under different numbers of training samples in each class are illustrated in Fig. 9. To achieve more accurate accuracy, we compute the recognition accuracies by averaging the results over 10 random splits of training and test samples. The performance of our proposed methods is compared with LDA, MMC, CCA, TR-LDA and LFDA. Observing from the Fig. 9, we can conclude that: (1) for each simulation setting, the recognition performance of all the methods varies with the increasing number of reduced dimensions. Specially, the accuracy of each method initially increases as the number of reduced dimensionalities increases and then starts to decrease after some point. In another word, too higher reduced dimensions may cause the embedding result of each method to deteriorate. (2) Our proposed TR-LODA and TR-MLODA algorithms outperform the remaining methods for settings of G_{15}/P_{26} and G_{20}/P_{21}. In particular, trace ratio criterion based TR-LODA and TR-MLODA are better than LODA and MLODA in these two cases. On the contrary, for the latter two settings, that is G_{25}/P_{16} and G_{30}/P_{11}, the recognition accuracies are significantly improved by using our LODA and MLODA, compared with the other methods,

Table 6
Performance comparisons on the ETH80 object recognition database (80 classes).

Method	Result Simulation setting							
	ETH80 database (Dim=1024,G15/P26)				*ETH80 database (Dim=1024,G20/P21)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.4764	0.5202	24	0.2227	0.5089	0.5683	20	0.2387
MMC	0.5702	0.5954	44	0.2057	0.5916	0.6321	24	0.2216
CCA	0.4762	0.5202	22	0.2223	0.5115	0.5683	20	0.2446
LFDA	0.5730	0.6067	12	0.2845	0.5996	0.6355	8	0.3687
TR-LDA	0.4373	0.4859	148	0.2333	0.4966	0.5344	120	0.2464
LODA	0.5666	0.5941	16	2.6961	0.5905	0.6222	20	3.4383
MODA	0.5659	0.5933	16	2.7268	0.5900	0.6208	20	3.4508
TR-LODA	0.5763	0.6046	24	2.7293	0.5970	0.6379	28	3.4442
TR-MLODA	0.5912	0.6239	20	2.7431	0.6099	0.6544	24	3.4556
	ETH80 database (Dim=1024,G25/P16)				*ETH80 database (Dim=1024,G30/P11)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.5426	0.6055	24	0.2743	0.5630	0.6284	28	0.2712
MMC	0.6122	0.6578	24	0.2541	0.6312	0.6708	20	0.2496
CCA	0.5461	0.6052	24	0.2901	0.5638	0.6284	28	0.2991
LFDA	0.6200	0.6411	20	0.4359	0.6467	0.6969	12	0.5453
TR-LDA	0.5581	0.5940	120	0.28580.2858	0.5902	0.6378	120	0.2876
LODA	0.6545	0.6807	8	2.3938	0.6546	0.6888	8	5.7790
MODA	0.6643	0.6881	16	2.4147	0.6797	0.7125	20	5.8056
TR-LODA	0.6230	0.6652	24	2.4307	0.6491	0.6899	20	5.7732
TR-MLODA	0.6391	0.6679	20	2.4628	0.6576	0.7032	16	5.8566

including TR-LODA and TR-MLODA. (2) MMC and LFDA also work well and are able to deliver the comparable results to our methods in most cases. The accuracies of CCA and LDA are still very close in each case. The performance of CCA, LDA and TR-LDA are worse than the other methods in virtually all the cases. This may be due to that the criteria of CCA, LDA and TR-LDA methods are unable to represent the complex intrinsic distribution of this dataset respectably.

Table 6 details the mean and the highest accuracies based on the results of Fig. 9. The averaged running time that is computed in seconds and the standard deviations over the repetitions are also described. Note that the best subspaces corresponding to the best accuracies are similarly named as the optimal object image subspaces. From the experimental results, we can observe that: (1) the numerical mean and best results shown in the tables are consistent to the results of the above figures in terms of performance superiority. (2) The standard deviation over repetitions produced by each method is comparable, implying that the stability of these methods on this dataset is similar. (3) Considering the running time performance, our methods need lightly more time than the other methods due to fact that the class number is larger, however our proposed methods are capable of exhibiting the better accuracies, including means and best records, compared with other methods in most cases. In particular, our methods need relatively smaller number of reduced dimensions to produce the optimal object image subspace.

4.2.3. Object recognition of 8 categories

We also address another object recognition task using the ETH80 database. In this simulation, each of the 8 big categories is regarded as a single class. Thus an 8-class classification problem is created and tested. Similarly, four experimental configurations over different G_P/P_q are evaluated. We first report the recognition accuracy over different numbers of k in NNS in Fig. 10. For each tested case, we average the accuracy over 20 random splits. The following observations are found. (1) The overall performance of our methods monotonically increases with respect to the k value when the number of k increases. (2) Compared with TR-LODA and TR-MLODA, LODA and MLODA are more robust to the number of k in each case. In this study, we also set the k value to p-2.

In Fig. 11, the recognition results under different numbers of reduced dimensions are illustrated. Observing from the results, we have: (1) the performance of all tested methods vary with the increasing number of reduced dimensions. In particular, the CCA, MMC, LFDA, LODA, MLODA methods exhibit an unusual behavior.

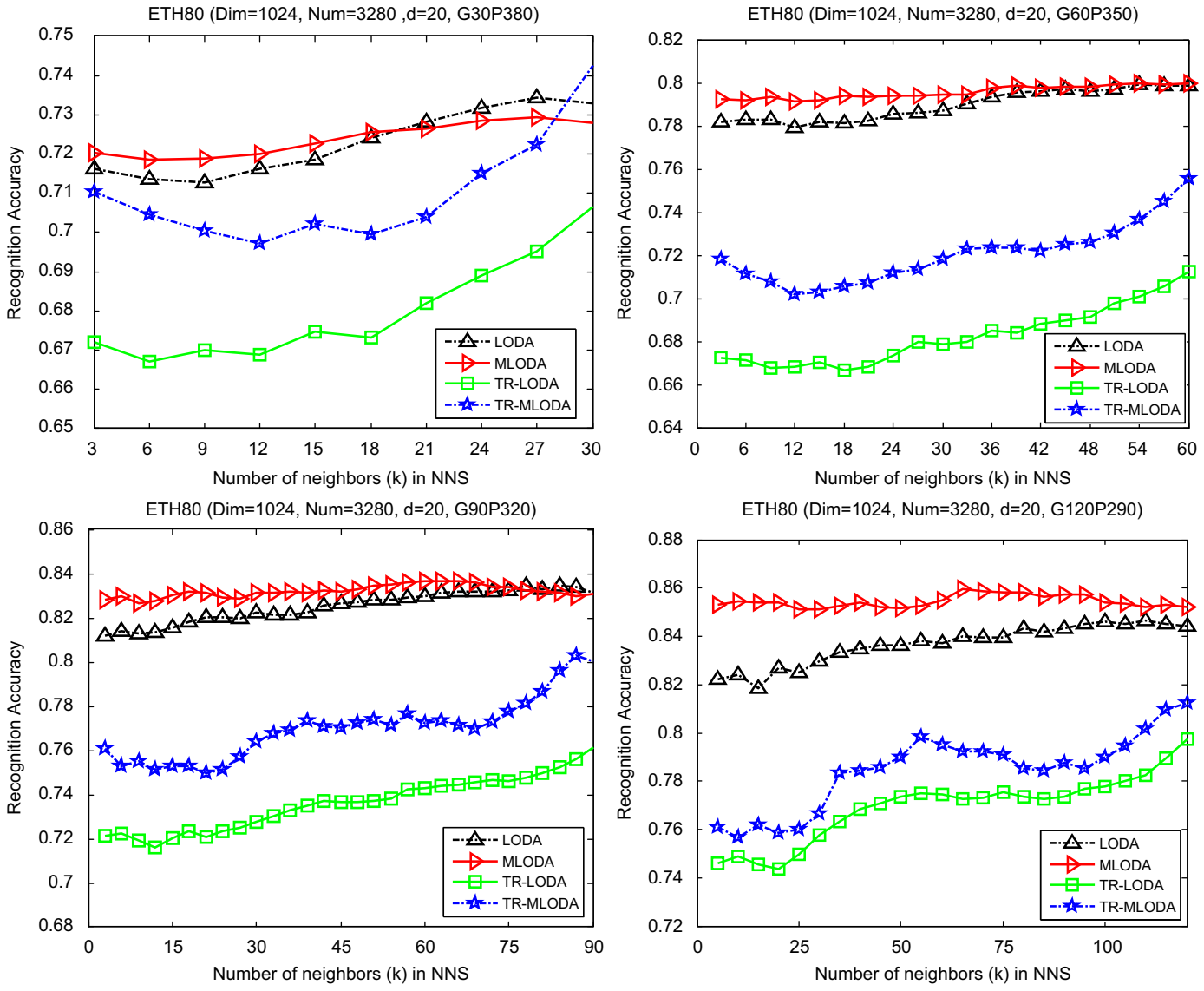

Fig. 10. Recognition accuracies vs. number of neighbors (k) on the ETH80 object database (8 classes).

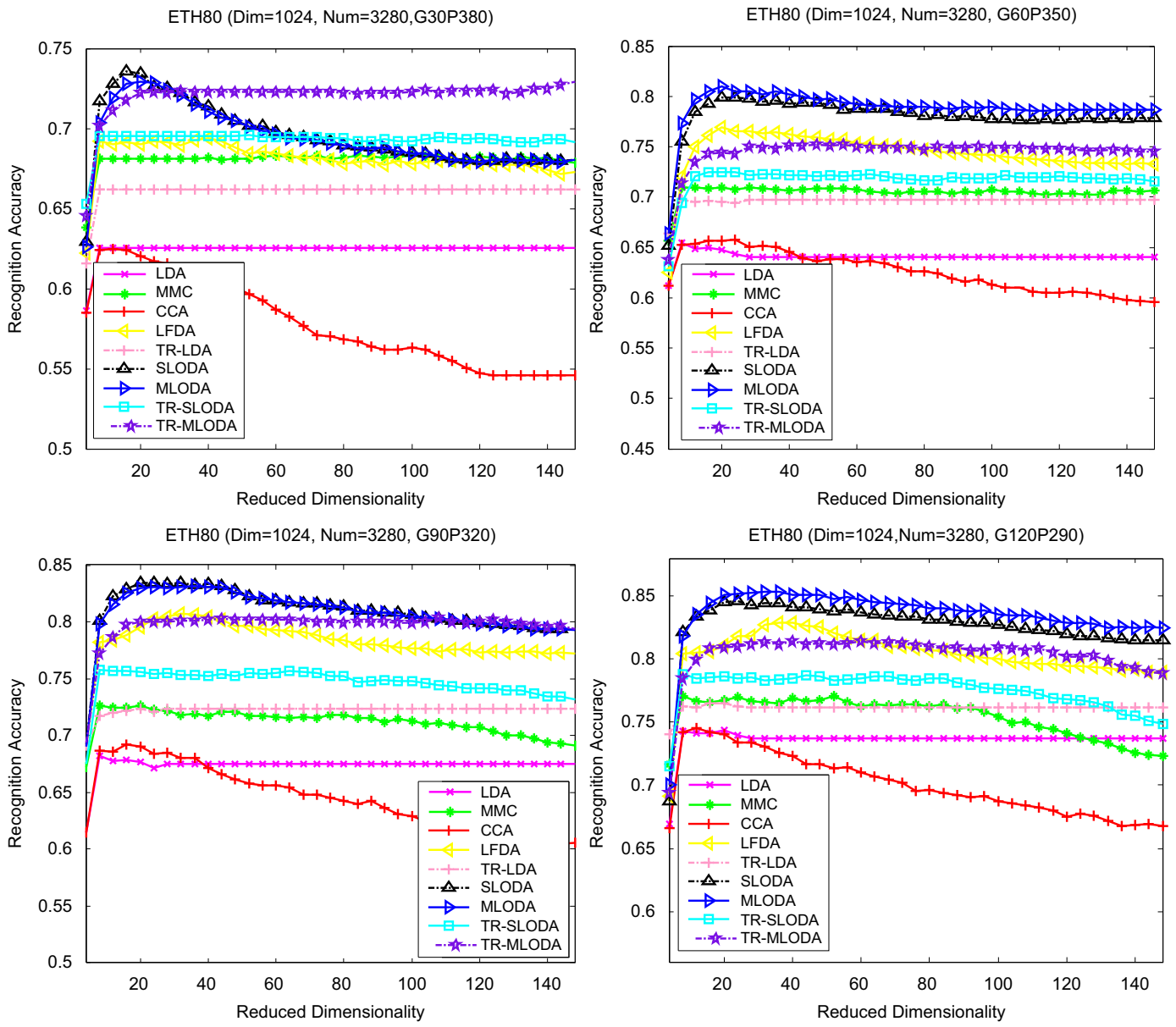

Fig. 11. Recognition accuracies vs. number of reduced dimensions on the ETH80 object database (8 classes).

That is their overall performance initially increases with the increasing reduced dimensions and then starts going down after some point to some extent till reaching the lowest record at around. (2) In all cases, LDA and CCA exhibit comparable results in the beginnings, but LDA outperform CCA as the number of reduced dimensions increases. Their performance is worse than the remaining methods in most cases. (3) Our presented LODA, MLODA and TR-MLODA methods can obtain the highest accuracies in most cases, compared with other methods. LFDA performs well in most cases and delivers the comparable results to our LODA and MLODA family methods. The results MMC and TR-LDA are comparative in each case and they tend to outperform LDA in all the cases. Also, MMC and TR-LDA exhibits slightly worse results to our TR-LODA methods in most cases.

Table 7 records the mean and the highest accuracies as shown in Fig. 11. We also record the averaged running time and standard deviations over the repetitions. The optimal object image subspaces corresponding to the best accuracies are also shown. From Table 7, we find that: (1) the performance superiority, including mean and best results, of each method keeps consistent with the observed results in Fig. 11. Similarly all the methods deliver comparable standard deviations over repetitions. (2) For runtime performance, the computational time of each method is comparative with each other.

5. Concluding remarks

This paper discusses the supervised dimensionality reduction problem. By taking more general class densities into account, we geometrically show that optimizing the LDA scatter criteria is not necessarily to achieve enhanced inter-class discrimination due to the existence of faraway points or outliers. In this paper, we focus on defining new robust criteria to improve the LDA criteria. By taking the distribution behavior of real datasets into account, we construct the density-region within each class. The data points in the density-regions are then used to compute the means and define the marginal scatters. As a result, faraway points or outliers cannot affect the construction of the scatters. We then propose a robust linearly optimized discriminant analysis (LODA) technique.

Table 7
Performance comparisons on the ETH80 object database (8 classes).

Method	Result Simulation Setting							
	ETH80 database (Dim=1024,G30/P380)				*ETH80 database (Dim=1024,G60/P350)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.6241	0.6252	8	0.1971	0.6410	0.6548	8	0.2887
MMC	0.6805	0.6832	60	0.1194	0.7047	0.7099	12	0.1857
CCA	0.5785	0.6245	12	0.1613	0.6258	0.6580	24	0.2473
LFDA	0.6812	0.6939	40	0.2688	0.7439	0.7693	20	0.3316
TR-LDA	0.6607	0.6620	8	0.1615	0.6959	0.6976	8	0.2487
LODA	0.6955	0.7355	16	0.2211	0.7802	0.7986	20	0.2777
MODA	0.6939	0.7293	20	0.2234	0.7887	0.8096	20	0.2726 0.2726
TR-LODA	0.6929	0.6957	52	0.2954	0.7170	0.7243	24	0.3711
TR-MLODA	0.7201	0.7292	148	0.2890	0.7438	0.7524	56	0.3606
	ETH80 database (Dim=1024,G90/P320)				*ETH80 database (Dim=1024,G120/P290)*			
	Mean	Best	Dim	Time(s)	Mean	Best	Dim	Time(s)
LDA	0.6740	0.6820	8	0.3582	0.7359	0.7435	20	0.3983
MMC	0.7120	0.7266	8	0.2505	0.7546	0.7711	52	0.2891
CCA	0.6446	0.6925	26	0.3284	0.7005	0.7455	12	0.3682
LFDA	0.7830	0.8074	32	0.4134	0.8036	0.8290	36	0.5906
TR-LDA	0.7221	0.7241	20	0.3235	0.7617	0.7653	16	0.3625
LODA	0.8098	0.8346	24	0.6653	0.8266	0.8458	24	0.9174
MODA	0.8086	0.8316	36	0.6816	0.8355	0.8535	32	0.9524
TR-LODA	0.7468	0.7576	8	0.7737	0.7757	0.7878	8	1.0188
TR-MLODA	0.7961	0.8035	108	0.7839	0.8032	0.8139	60	1.0440

LODA characterizes the inter-class separability and intra-class compactness by using a large margin criterion. We have shown that LODA is able to tackle the difficulty problems encountered by LDA effectively. We also present a natural multimodal extension of LODA for enabling it to deal with multimodal datasets directly. Mathematical comparisons and analyses between our work and the related work indicate that our LODA is more general with a strong generalization capability for discriminant analysis and can offer some attractive advantages.

The kernelized extension of our method is also addressed. It is found that the performance of kernelized methods heavily depends on the choice of kernel family. But there is still lack of theoretic criteria for selecting the kernels, thus this work only evaluates the presented linear methods. The performance of the proposed family methods is thoroughly evaluated by extensive simulations over benchmark real databases. For handwritten digits recognition and object recognition, the delivered overall performance of our methods is comparable or even better than some widely used state-of-the-art linear discriminant techniques. But we must say determining the optimal reduced dimensions for dimensionality reduction still remains an open issue. Through investigating the selection of the k value in NNS on real databases, we find that our methods almost always tend to deliver better results when the k value is relatively large compared with the training sample number.

Acknowledgments

The authors would like to express our sincere thanks to the anonymous reviewers' comments and suggestions which have made the paper a higher standard.

References

[1] L. Maaten, E. Postma, J. Herik, Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009.
[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, Boston, 1990.
[3] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.
[4] P.N. Belhumeour, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[5] L.F. Chen, H.Y.M. Liao, J.C. Lin, M.D. Kao, G.J. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recog. 33 (10) (2000) 1713–1726.
[6] J.P. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, J. Mach. Learning Res. 6 (2005) 483–502.
[7] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recog. 34 (2001) 2067–2070.
[8] X.P. Qiu, D. Wu, Face recognition by stepwise nonparametric margin maximum criterion, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005, pp. 1567–1572.
[9] S.W. Ji, J.P. Ye, Generalized linear discriminant analysis: a unified framework and efficient model selection, IEEE Trans. Neural Networks 19 (10) (2008) 1768–1782.
[10] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[11] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[12] X.G. Wang, X.O. Tang, Dual-space linear discriminant analysis for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004. pp. 564–569.
[13] K. Fukunaga, J. Mantock, Nonparametric discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 5 (1983) 671–678.
[14] M. Bressan, J. Vitrià, Nonparametric discriminant analysis and nearest neighbor classification, Pattern Recog. Lett. 24 (2003) 2743–2749.
[15] X.P. Qiu, D. Wu, Stepwise nearest neighbor discriminant analysis, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2005, pp. 829–835.
[16] Y.J. Zheng, J.Y. Yang, J. Yang, X.J. Wu, Nearest neighbour line nonparametric discriminant analysis for feature extraction, Electron. Lett. 42 (12) (2006) 679–680.
[17] J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.
[18] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Network 17 (1) (2006) 157–165.
[19] Y. Jia, F. Nie, C.S. Zhang, Trace ratio problem revisited, IEEE Trans. Neural Network 20 (4) (2009) 729–735.
[20] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, J. Mach. Learn. Res. 8 (2007) 1027–1061.
[21] Y. Guo, S. Li, J. Yang, T. Shu, L. Wu, A generalized Foley-Sammon transform based on generalized fisher discriminant criterion and its application to face recognition, Pattern Recog. Lett. 24 (1–3) (2003) 147–158.

[22] H. Wang, S. Yan, D. Xu, X. Tang, T. Huang, Trace ratio vs. ratio trace for dimensionality reduction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, pp. 1–8.
[23] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, IEEE Trans. Patten Anal. Mach. Intell. 27 (3) (2005) 228–340.
[24] B. Leibe, B. Schiele, Analyzing appearance and contour based methods for object categorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2003, pp. 409–415.
[25] M. Loog, Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion, Delft Univ. Press, 1999.
[26] C.P. Hou, J. Wang, Y. Wu, D.Y. Yi, Local linear transformation embedding, Neurocomputing 72 (10–12) (2009) 2368–2378.
[27] M. Loog, R.P.W. Duin, R. Haeb-Umbach, Multiclass linear dimension reduction by weighted pairwise fisher criteria, IEEE Trans. Patten Anal. Mach. Intell. 23 (7) (2001) 762–766.
[28] T. Sun, S. Chen, Class label versus sample label-based CCA, Appl. Math. Comput. 185 (1) (2007) 272–283.
[29] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[30] E.K. Tanga, P.N. Suganthana, X. Yaob, A.K. Qina, Linear dimensionality reduction using relevance weighted LDA, Pattern Recog. 38 (4) (2005) 485–493.
[31] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002 (pp. 25–55).
[32] S. Fiori, Visualization of Riemannian-manifold-valued elements by multidimensional scaling, Neurocomputing 74 (6) (2011) 983–992.
[33] J. Li, X.L. Li, D.C. Tao, KPCA for semantic object extraction in images, Pattern Recog. 41 (10) (2008) 3244–3250.
[34] F.Y. Yao, X.Y. Jing, H.S. Wong, Face and palmprint feature level fusion for single sample biometrics recognition, Neurocomputing 70 (2007) 1582–1586.
[35] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 234–244.

Zhao Zhang is currently working toward the Ph.D. degree at the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong. His current interests include machine learning, pattern recognition and applications.

Tommy W.S. Chow (IEEE M'93–SM'03) received the B.Sc. (First Hons.) and Ph.D. degrees from the University of Sunderland, Sunderland, UK. He is currently a Professor in the Electronic Engineering Department. His research interests are in the area of Machine learning including Supervised and unsupervised learning, Data mining, Pattern recognition and fault diagnostic. In professional activities, he has been an active Committee Member of the HKIE (Hong Kong Institution of Engineers) Control Automation and Instrumentation (CAI) division since 1992, and was the Division Chairman (1997–1998) for the HKIE CAI division. He has authored or coauthored of over 120 technical papers in international journals, 5 book chapters, and over 60 technical papers in international conference proceedings. He is serving as the Associate Editor of Pattern Analysis and Applications, and International Journal of Information Technology.