Linear Ranking Analysis

Weihong Deng, Jiani Hu, Jun Guo
Beijing University of Posts and Telecommunications,
Beijing, 100876, China
{whdeng, jnhu, guojun}@bupt.edu.cn

Abstract

We extend the classical linear discriminant analysis (LDA) technique to linear ranking analysis (LRA), by considering the ranking order of classes centroids on the projected subspace. Under the constrain on the ranking order of the classes, two criteria are proposed: 1) minimization of the classification error with the assumption that each class is homogenous Guassian distributed; 2) maximization of the sum (average) of the k minimum distances of all neighboring-class (centroid) pairs. Both criteria can be efficiently solved by the convex optimization for one-dimensional subspace. Greedy algorithm is applied to extend the results to the multi-dimensional subspace. Experimental results show that 1) LRA with both criteria achieve state-of-the-art performance on the tasks of ranking learning and zero-shot learning; and 2) the maximum margin criterion provides a discriminative subspace selection method, which can significantly remedy the class separation problem in comparing with several representative extensions of LDA.

1. Introduction

Dimension reduction helps pattern classification by selecting a low-dimensional subspace that preserves the class separability, such as the wide applications on face recognition [7][6][4]. Moreover, it provides a low-dimensional, usually 1D or 2D, graphical representations that are useful for preliminary analyses and data visualization in various areas. For example, in genomics, one would like to find the single combinations of gene mutations that cause a set of subclasses of a disease. In psychology, one usually want to visualize a group of samples belonging to multi-classes in a 2D plot from which conclusions can be drawn. Fisher's linear discriminant analysis (LDA) [11] is one of the most important method for dimension reduction. It selects the $(C-1)$-dimensional, wherein C is the class number, subspace by simultaneously maximizing the between-class scatter and minimizing the within-class scatter. However,

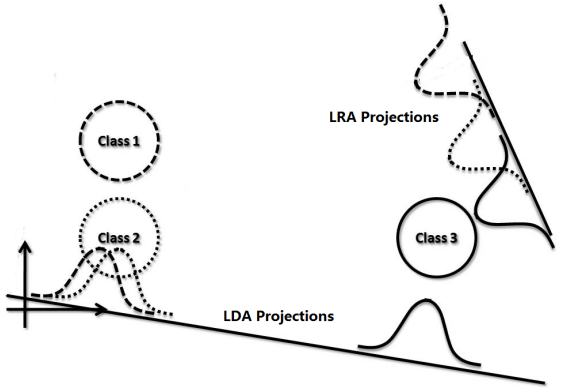


Figure 1. There are three classes (named 1, 2, and 3) of samples, which are drawn from a Gaussian distribution in each class. LDA finds a projection direction that maximize the between-class scatter, while merges class 1 and class 2. The proposed LRA aims to minimize the classification error (or maximize neighboring-class margin) while preserving the ranking order "Class 1$\succ$Class 2$\succ$Class 3"

LDA has two apparent limitations as follows.

- When the dimension is less than $C-1$, LDA is suboptimal and could merge the close classes. This is defined as the *class separation* problem [24] in the literature.
- LDA does not model the relative strength of the high-level semantic attributes, which is recently shown to be useful for zero-shot learning and recognition[20].

To address these two limitations, we extend the classical LDA technique to linear ranking analysis (LRA), by considering the ranking order of classes centroids on the projected subspace. Under the constrain on the ranking order of the projected class centroids, two criteria are proposed: 1) minimization of the classification error with the assumption that each class is homogenous Guassian distributed; 2) maximization of the sum (average) of the k minimum distances of all neighboring-class (centroid) pairs. Both criteria can be efficiently solved by the convex optimization for one-dimensional subspace. Greedy algorithm is applied to extend the results to the multi-dimensional subspace.

Experimental results on synthetic data set, Outdoor Scene Recognition, and Public Figure Face Database show that LRA with both criteria achieves state-of-the-art performance on the tasks of ranking learning and zero-shot learning. The results on two UCI machine learning repository and USPS handwriting digits show that the maximum margin criterion is a potential discriminative subspace selection method, which significantly reduces the class separation problem in comparing with several representative extensions of LDA.

2. Related Works

2.1. Linear Discriminant Analysis and Extensions

The aim of LDA is to find a low dimensional subspace in which the ratio between the within-class scatter and between-class scatter is minimized. Assuming there are C classes to be analyzed, the subspace is spanned by a set of vectors, w_i, $1 \leq i \leq C-1$, which forms the columns of the matrix $W = [w_1, \ldots, w_{C-1}]$. The ith class contains n_i training samples x_{ij}, $1 \leq j \leq n_i$, and has a centroid of $\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined by

$$S_w = \frac{1}{n}\sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{1}$$

$$S_b = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n_i} n_i(x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \tag{2}$$

where $n = \sum_{i=1}^{C} n_i$ is the sample size of training set and $\mu = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n_i} x_{ij}$ is the global centroid of the training set. The optimal projection matrix W of LDA is computed from the eigenvectors of $S_w^{-1}S_b$, under the assumption that S_w is invertible or the eigenvectors of $(S_w+\sigma I)^{-1}S_b$, when S_w is singular.

Several algorithms [17][18][24][1][5], which emphasizes the close class pairs by adaptively weighting between-class components, have been proposed to address the class separation problem. However, they cannot guarantee the separation of all class pairs, and thus the far apart class pairs may always have influence on the close class pairs. To address this limitation, Bayes optimal LDA (BLDA) does not assign weights to class pairs, but directly minimizes the classification error in the one-dimensional projected space [13]. The minimum error criterion of our work learns the ranking function by a similar optimization procedure to [13], but we have extended it to the applications of the ranking learning and zero-shot learning, rather than only for the classification purpose. Moreover, we propose a maximum margin formulation which is shown to outperform BLDA on the classification of real-world data.

The maximization of the minimum distance between the neighboring class centroids has been explored by Bian and Tao [2]. Our proposed max-k-min criterion is more general and arguably more robust. The max-min distance analysis in [2] focuses on the closest class pair and ignore the global extension. In contrast, our max-k-min criterion considers the global extension of the k closest class pairs, and would be more adaptive to the various class distributions by properly selecting k. Second, the objective function of our criterion is convex and can be solved efficiently by linear programming, but that of [2] is not convex and can only be solved approximately and costly by the semidefinite-programming.

2.2. Binary and Relative Attributes

Most existing works treat attributes as binary predictors indicating the presence or absence of a certain property of an sample. Learning attribute categories have been used to predict texture or color types [10], and provide a middle-cue for object or face recognition. Moreover, the high-level semantics of attributes also enable zero-shot transfer [16][22], or description and localization [9][25]. This may be sufficient for the binary properties, such as 'is Asian' and 'wearing eyeglasses'. These methods model the attributes as binary, but a large number of attributes are not binary, and described naturally in a relative way.

The pioneering work of Parikh and Grauman [20] learned a ranking function on images based on constrains specifying the relative strength of attributes. Given a set of training samples represented by $x_i \in \mathbb{R}^n$. For each attribute, we are given a set of ordered pairs of samples $O = \{(i,j)\}$ and a set of unordered pairs $S = \{(i,j)\}$ such that $(i,j) \in O \Rightarrow i \succ j$, i.e. sample i has a stronger attribute than j, and $(i,j) \in S \Rightarrow i \sim j$, i.e. the attribute of sample i is similar to that of sample j. The goal is learning a ranking function

$$r(x_i) = w^T x_i \tag{3}$$

in order to satisfy the maximum number of following constrains $\forall (i,j) \in O : \; w^T x_i > w^T x_j$ and $\forall (i,j) \in S : w^T x_i = w^T x_j$. The problem can be approximated solved by the modified ranking SVM formulation. Our work also learns the ranking function based on relative strength of attributes, but treat the class as a whole rather than treating the samples individually. Thus, unlike the learning-to-rank formulation, we can impose the distributional assumption and perform marginal analysis to the classes , which we show achieve state-of-the-art performance on the applications of zero-shot learning and dimension reduction.

3. Linear Ranking Analysis

Linear ranking analysis aims to learn the ranking function on the C classes of interest with predefined ordering.

This technique is then extended to solve the zero-shot learning and dimension reduction problems.

3.1. Minimum Error Criterion

Theorem 1 ([13]). *Define a constrained region $\mathcal{A}$ where all vectors w sampled from it generate the same ordered sequence $\eta_{(i)}$ of the projected centroid locations $\eta_i = w^T\mu_i$ of C classes. Then, the region $\mathcal{A}$ is a convex polyhedron.*

After the within-class whitening preprocessing in the input feature space, we assume the conditional distribution is homoscedastic Gaussian, i.e. $N_i(\mu_i, \Sigma_i)$ and $\Sigma_i = I$. Under this assumption, the error rate of Bayes optimal classification with ordered classes $\mu_1, \ldots, \mu_C$ can be expressed as follows.

$$J(w) = \frac{2}{C}\sum_{i=1}^{C-1}\Phi\left(\frac{-w^T\mu_{i,i+1}}{2}\right) \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution, $w^T\mu_{i,i+1} = w^T(\mu_{i+1} - \mu_i)$ is the projected distance between the neighborly ordered classes. The gradient of the error function is

$$\frac{\partial J}{\partial w} = -\frac{1}{C}\sum_{i=1}^{C-1}\frac{1}{\sqrt{2\pi}}e^{(w^T\mu_{i,i+1})^2/4}\mu_{i,i+1} \tag{5}$$

The Heissian of the error function is

$$\frac{\partial^2 J}{\partial^2 w} = \frac{1}{4C\sqrt{2\pi}}\sum_{i=1}^{C-1}e^{(w^T\mu_{i,i+1})^2/4}(w^T\mu_{i,i+1})\mu_{i,i+1}\mu_{i,i+1}^T \tag{6}$$

Because the Hessian matrix is semi-definite in $\mathcal{A}$, the objective function $J(w)$ is apparently a convex function that can be easily minimized.

This theoretically elegant criterion has two practical limitations. First, this criterion becomes suboptimal when the distribution of real data are far from Gaussian. Second, more severely, this criterion is not applicable for some predefined orderings, on which the solution region is the origin. To address these limitations, we propose a novel maximum margin criterion for more general usage.

3.2. Maximum Margin Criterion

Let $y = (y_1, \ldots, y_{C-1})$, where $y_i = w^T\mu_{i,i+1}$, be a vector in $\mathbb{R}^{C-1}$ to represent the *distances* between the nearby class means given a certain ordered sequence of projected class means. Assuming that the classification errors are mainly caused by the k closest class pairs in the projected subspace, and it is natural to seek the optimal subspace in manner that maximizing the sum (average) of the k minimum distances between the nearby class pairs. To facilitate our discussion, define $\theta(y) = (y_{(1)}, y_{(2)}, \cdots, y_{(C-1)})$ to be the vector obtained by sorting the $C-1$ components of y in order, i.e., $y_{(1)} \le y_{(2)} \le \cdots \le y_{(C-1)}$. Finally, for $k = 1, \cdots, C-1$, we define $\Theta_k(y) = \sum_{i=1}^{k} y_{(i)}$, the sum of the k minimum components of y. Finally, the *max-k-min distance criterion* for class separation is formulated as follows.

$$\max_w \Theta_k(y) \tag{7}$$

- By settling $k = 1$, the criterion guarantees the closest pair of classes in the projected subspace is not too close. This is reasonable when the classification errors are mostly caused by a single closest pair of classes. However, it is possible that the classification errors are concurrently caused by several close pairs of classes. Thus, the general errors may be increased in order to separate the single worst pair of classes apart. In other words, mapping the worst pair of classes overlapping may improve on the overall performance. See the synthetic data test in Section 4.1 for example.

- By settling $k = C - 1$, the criterion maximizes the sum of all distance between all neighboring classes, which is solely dependent on the first and last elements of the ordered sequence of projected means. The algorithm actually maximizes the projected range of all classes, and this may be reasonable if all the between-class distances are similarly close. When the between-class distances are diverse, the criterion may emphasize the large between-class distances and ignore the small ones, which is similar to what classical LDA does.

- By properly selecting k from $\{1, 2, 3, \ldots, C-1\}$, the max-k-min distance criterion can be adaptive to various class distributions. It is difficult to define a criterion for selecting k without restrictive distributional assumption. In our experiments, we empirically select k by minimizing the classification error in cross-validation.

To facilitate our discussion, we define $(z)_+ = \max(z, 0)$, $y_i = w^T\mu_{i,i+1}$ be the distance between neighboring (projected) prototypes. Define $\theta(y) = (\theta_1(y), \theta_2(y), \cdots, \theta_{C-1}(y))$ to be the vector obtained by sorting the $C-1$ components of y in nondecreasing order, i.e., $\theta_1(y) \le \theta_2(y) \le \cdots \le \theta_{C-1}(y)$. Finally, for $k = 1, \cdots, C-1$, define $\Theta_k(y) = \sum_{i=1}^{k}\theta_i(y)$, the sum of the k minimum components of y. In light of the formulation in [19], it is easy to justify following lemma and theorem.

Lemma 1 ([19]). *For any vector $y \in \mathbb{R}^{C-1}$, and $k =$*

$1, \cdots, C-1$

$$\Theta_k(y) = \frac{1}{C-1}\left(k\sum_{i=1}^{C-1} y_i - \max_{t\in\mathbb{R}} \sum_{i=1}^{C-1}(C-1-k)(t-y_i)_+ + k(y_i-t)_+\right) \tag{8}$$

Moreover, $t^* = \theta_k(y)$ is an optimizer of the above minimization problem.

Theorem 2 ([19]). *Given the collection of linear function, $\{g_i(w) = w^T(\mu_{i+1} - \mu_i)\}_{i=1}^{C-1}$, the problem of maximizing $\Theta_k(g(w))$, the sum of the k smallest function is convex, which can be formulated as a linear program as follows*

$$\begin{aligned} &\max \left(kt - \textstyle\sum_{i=1}^{C-1}\xi_i\right) \\ &\textit{subject to} \\ &\xi_i \geq t - g_i(w),\ i = 1, \ldots, C-1, \\ &\xi_i \geq 0,\ i = 1, \ldots, C-1, \\ &w^T w \leq 1 \end{aligned} \tag{9}$$

3.3. Applications

3.3.1 LRA Based Zero-shot Learning

Consider N classes of interests. In the training stage, S of the these classes are 'seen' classes for which training images are provided, while the remaining $U = N - S$ classes are 'unseen', for which no training images are provided. The S classes are described by the ranking order of the presence of a certain attribute. On the other hand, the U unseen classes are described relative to one or two seen classes for a subset of the attributes. For example, seen classes $c_j^{(u)}$ can be described as $c_i^{(s)} \succ c_j^{(u)} \succ c_k^{(s)}$ for attribute a_m, or $c_i^{(s)} \succ c_j^{(u)}$, or $c_j^{(u)} \succ c_k^{(s)}$, where $c_i^{(s)}$ and $c_k^{(s)}$ are seen classes. In the testing stage, a novel sample is to be classified into any of the N classes.

Predicting the real-valued rank of all samples in the training set allows us to transform the samples from $x_i \in \mathbb{R}^n$ (observation space) to $\tilde{x}_i \in \mathbb{R}^M$ (attribute space), in such a way that each sample i is represented by an M-dimensional vector $\tilde{x}_i$ storing its ranking score for all M attributes. Then, the Gaussian distribution based maximum likelihood classification is performed in the attribute space. In the training stage, generative models $c_i^{(s)} \sim \mathcal{N}\left(\mu_i^{(s)}, \Sigma_i^{(s)}\right)$ of each of the S seen classes are first computed, then the models of the unseen classes are selected by the rules defined in the well-known relative attributes method [20].

Given a test sample x, the M-dimensional ranking score vector $\tilde{x}$ is first computed, and then the class label is assigned by the maximum likelihood rule

$$c^* = \arg\max_{i\in\{1,\ldots,N\}} P\left(\tilde{x}|\mu_i, \Sigma_i\right) \tag{10}$$

3.3.2 LRA Based Dimension Reduction for Classification

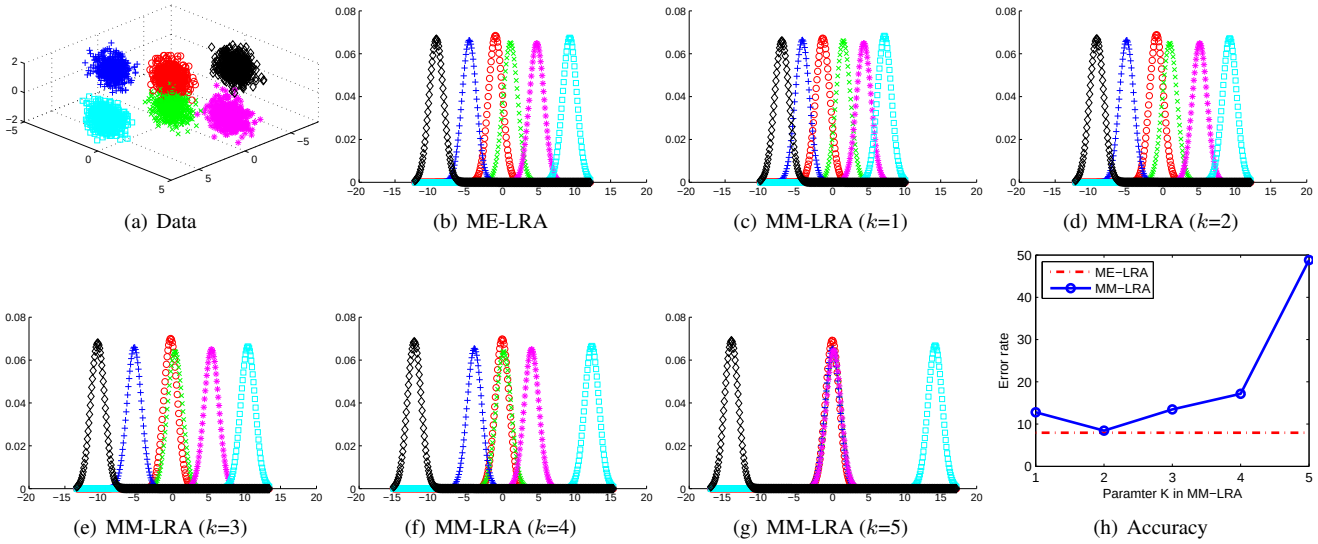
Theorem 2 provides an efficient method to find the max-k-min margin solution for any given ordered sequence of the projected centroids $\eta_{(1)} \leq \eta_{(2)} \leq \cdots \leq \eta_{(C)}$ by solving a linear programming problem. The remaining problem is to determine which of all possible sequences provides the optimal solution, where the sum of k minimum distances is maximized. Apparently, the number of possible sequences is $C!$, and two mirrored sequences, e.g. $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_C$ and $\eta_C \leq \eta_{C-1} \leq \cdots \leq \eta_1$ result in the same solution. Moreover, one can filter out the feasible sequence by detecting whether the convex region $\mathcal{A}$ is the origin. In general, one need to solve smaller number of linear programming problems with $C!/2$ being the upper limit.

The one dimensional Max-K-Min projection algorithm first searches the possible subproblems with feasible class orderings, then solves them separately according to Theorem 2, and finally finds the global optimum by comparing the results of all subproblems. Specifically, the algorithm can be summarized as follows:

- First, find the set Q of possible orderings of the class means. This is easily achieved by selecting all those sequences for which $\mathcal{A}$ is larger than the origin.
- Second, for each ordering q_i in Q find that $w^{(i)} \in \mathcal{A}$, which minimizes the sum of the k minimum distances by using a linear programming algorithm (We use CVX to solve the LP problem in our experiments).
- Finally, the optimal solution w to our problem is given by

$$w = \arg\max_{w^{(i)}} \Theta_k\left(g(w^{(i)})\right) \tag{11}$$

To find a subspace solution of more than one dimension, we recursively apply our algorithm to the null space of the previously obtained subspace. After applying the algorithm described in the previous subsection, one obtains the first subspace solution, which is denoted as w_1. The optimal k is selected by cross-validation with highest accuracy. The null space of this first projection is denoted $W_1^\perp$. Now, we can re-apply the same algorithm within this null space $W_1^\perp$, obtaining the second optimal dimension w_2. To do this, we will first need to project the class means onto $W_1^\perp$ and then calculate the next solution on this null space. The 2-dimensional subspace where the sum of k minimum distance is minimized is then given by the projection matrix $W_2 = (w_1, w_2)$ with the usual constraint $w_1^T w_2 = 0$. In this way, our algorithm can be recursively applied to find that d-dimensional solution from any m-dimensional space, with $d < \min(m, C-1)$.



(a) Data (b) ME-LRA (c) MM-LRA (k=1) (d) MM-LRA (k=2)

(e) MM-LRA (k=3) (f) MM-LRA (k=4) (g) MM-LRA (k=5) (h) Accuracy

Figure 2. The LRA learning of ranking function of six normal distributions with the predefined ordering (black $\prec$ blue $\prec$ red $\prec$ green $\prec$ magenta $\prec$ light-blue). (a) three-dimensional scatter plot of six classes. (b) Bayes optimal 1D result . (c–g) show the class distributions in the max-k-min optimal 1D subspace with $k = \{1, 2, 3, 4, 5\}$. (h) Dependence of error rate (using the nearest mean classification) on the parameter k of the MM-LRA using the homogeneous Gaussian data set.

4. Experiments

4.1. Synthetic Data Test

This simulation is intended to serve as an illustrative example which involves 6 homoscedastic Gaussian distributions embedded in a 3-dimensional space with means located at: $(0, 0, 1)^T$, $(0, 0, -1)^T$, $(0, 4, 0)^T$, $(0, -4, 0)^T$, $(7, 0, 0)^T$, $(-7, 0, 0)^T$, respectively, and covariance matrices equal to one fourth of the identity matrix, i.e., $\Sigma = I/4$. In our simulation, we randomly generated 500 samples from each of these distributions. The class distributions of these samples were shown in Fig. 2(a)

The solution of the ME-LRA is shown in Fig. 2(b), which results in a about 8% error rate, which is the lower bound of the error rate that one can be obtained in this data set. Fig. 2(c–g) show the 1D representation obtained with the MM-LRA algorithm for $k = \{1, 2, 3, 4, 5\}$. When the algorithm maximizes the minimum distance with $k = 1$, the intervals between neighboring class are seen to be uniform so that the separation of all class pairs is considered. This observation is consistent with Bian and Tao [2]. However, $k = 1$ is not optimal for classification because the global extent of the whole data is limited in this case. In the other limit, when the algorithm maximizes the global extent of the data with $k = 5$, four of the six classes are largely overlapped so the classification error becomes as high as 49%. This result is similar to the result of the classical LDA that maximizes the scatter of the class mean. When $k = 2$, MM-LRA considers both the class separation and the global extent, and reaches the lowest classification error, i.e. 8%.

Fig. 2(h) shows the classification errors in the one-dimensional subspace derived by different choice of k, and one can see from the figure that $k = 2$ produces the best performance. In general, the choice of k affects the classification error rate to large extent, and thus, for practical applications, it is important to use the cross-validation method to select a suitable k for each subspace dimension. Note that ME-LRA provides theoretically optimal solution on homoscedastic Gaussian data, and MM-LRA approximate this optimal result by controlling both class separation and global scale. Further, we can expect MM-LRA may be more universally valid than ME-LRA since it does not impose any distributional assumption during the optimization.

4.2. Ranking Learning Results

Outdoor Scene Recognition (OSR) Dataset, which contains 2688 images from 8 classes, and a subset of Public Figure Face Database (PubFig), which contains 800 images from 8 random identities (100 images per class) are used to evaluate our approaches. A concatenation of the gist descriptor and a 45-dimensional Lab color histogram is used as our image features. The reference [20] provides more details about the datasets, which include the binary memberships and relative orderings of categories by attributes. These were collected using the judgements of a colleague unfamiliar with the details of that work.

For each attribute, we use 30 training images per class, and the rest for testing. For an image-pair (i, j), in the test set, we evaluate the learnt ranking function, and if $w^T x_i > w^T x_j$, we predict $i \succ j$, else $i \prec j$. We implement LDA, linear binary SVM, modified ranking SVM to learn the ranking function, to compare them with our pur-

Table 1. The ranking function's accuracy on various attributes on Outdoor Scene Recognition and Public Figure Face Databases

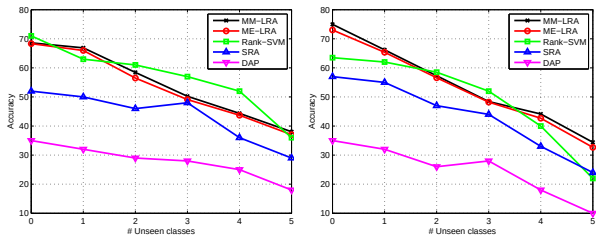
Attributes	LDA	Binary SVM	Rank SVM	ME-LRA	MM-LRA
natural	92.25	90.74	**94.36**	94.03	94.04
open	68.48	84.54	**90.97**	89.88	89.92
perspective	78.87	78.22	**85.78**	85.17	85.45
large-objects	74.57	69.85	**86.36**	85.37	85.87
diagonal-plane	81.29	81.82	**87.52**	86.89	87.23
close-depth	72.96	86.89	**88.70**	83.38	86.56
masculine-looking	75.83	70.05	81.00	82.17	**82.47**
white	60.58	64.37	77.31	77.46	**78.46**
Young	78.40	74.48	81.05	81.46	**82.02**
Smiling	71.67	68.97	79.66	79.73	**80.65**
Chubby	60.94	61.65	76.14	75.85	**77.24**
visible-forehand	79.25	75.20	**87.91**	86.74	87.42
bushy-eyebrows	67.79	69.28	78.89	80.08	**80.57**
narrow-eyes	57.35	74.80	**80.72**	79.08	80.04
pointy-nose	58.11	68.75	74.84	77.35	**78.64**
big-lips	64.11	73.88	78.07	79.56	**80.12**
round-face	73.44	72.69	80.46	81.79	**81.96**
Average	71.52	74.48	82.93	82.71	**83.45**

posed Minimum Error-LRA and Maximum-Margin-LRA. As shown in Table 1, the learnt ranking function's accuracies of LRA are similar to the ranking SVM, confirming the effectiveness of LRA on class-level relative attribute modeling. As expected, LRA is significantly better than LDA by modeling of the relative strength of attributes. MM-LRA is slightly better than ME-LRA, indicating the margin based criterion is more suitable for real data sets.

4.3. Zero-Shot Learning Results

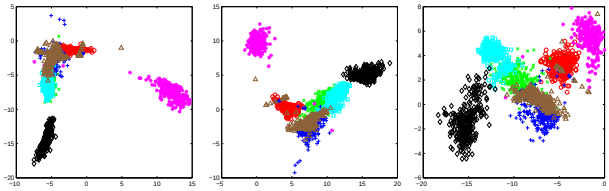
We compare our approach to three baselines: The first baseline is the direct attribute prediction (DAP) model of Lampert et al. [16], which trains linear SVMs by transferring the binary supervision to training samples from the seen categories. A test image x is assigned to a class according to a naive bayes rule, where the posterior of each attribute is approximated by the sigmoid function. The second method, "score based relative attributes (SRA)" also based on the scores of the linear SVMs as features, but use generative modeling of seen classes and relative descriptions of unseen classes as our approach. The final baseline, the relative attribute (RA) method, uses the scores of a modified ranking SVMs to construct the attribute space, which has demonstrated state-of-the-art zero-shot learning performance [20].

We use 30 training images per class, and the rest for testing, and report the mean accuracy over 10 random train/test and seen/unseen splits. We study zero-shot learning accuracy . Fig. 4.3 show the zero-shot classification accuracy as the number of unseen classes increases, and one can see that the proposed LRA is similar to the ranking SVM method, indicating the idea of LRA is effective to solve the zero-shot learning problem. The difference between MM-LRA and ME-LRA is not notable on both data sets. In general, we can expect that LRA is more scalable than ranking SVM to the problems with large number of classes, as it conducts analysis only on the class centroids.



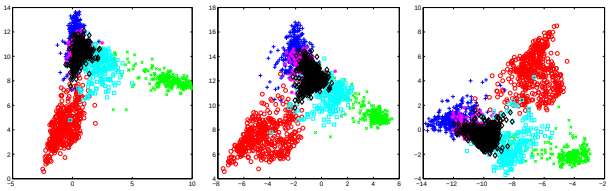
(a) OSR (b) PubFig

Figure 3. Zero-shot learning performance as the proportion of unseen classes increases. Total number of classes is constant at 8.



(a) LDA (b) BLDA (c) LRA

Figure 4. Projections of the Image Segmentation data (testing set) onto the two most discriminant feature vectors found by (a) LDA, (b) Bayes optimal LDA, (c) LRA.



(a) LDA (b) BLDA (c) LRA

Figure 5. Projections of the LandSat Satellite data (testing set) onto the two most discriminant feature vectors found by (a) LDA, (b) Bayes optimal LDA, (c) LRA.

4.4. Dimension Reduction with A Small Number of Classes

We evaluate the effectiveness of LRA[1] on two data sets from the UCI Machine Learning Repository. The "Image Segmentation" data set consists of 2,310 measurements with 19 attributes from seven classes: brickface, sky, foliage, cement, window, path, and grass. The data set is divided into a training set of 210 images (30 samples per class) and a 2,100 testing images (300 samples per class). The "Landsat Satellite" data set consists of 6,435 measure-

[1]The minimum error criterion has been used by BLDA [13] for classification. Therefore, we only test LRA with maximum margin criterion in the following classification experiments, and compare its performance with BLDA.

Table 2. Average Classification Error Rate by Nearest Mean Classifier (upper table) and Nearest Neighbor Classifier (lower table) on Image Segmentation Data Set

Dim	1	2	3	4	5	6	Rank
LDA	49.39	31.21	15.71	11.04	9.74	8.53	5.60
GMSS	31.43	16.84	17.19	11.77	9.70	8.53	4.20
HMSS	32.47	21.21	15.54	10.95	9.91	8.53	4.40
aPAC	44.85	30.09	17.32	11.65	9.44	8.53	5.20
MMDA	49.05	19.70	14.68	12.90	8.87	8.53	4.20
BLDA	28.53	19.35	15.02	11.82	9.26	8.53	3.40
LRA	**27.14**	**15.71**	**12.21**	**9.52**	**8.74**	8.53	**1.00**

Dim	1	2	3	4	5	6	Rank
LDA	42.38	21.17	8.63	3.20	3.29	2.73	5.80
GMSS	31.22	14.16	7.84	3.85	3.90	2.73	5.20
HMSS	30.81	14.46	7.88	3.24	3.12	2.73	4.20
aPAC	41.90	21.82	7.10	3.20	3.16	2.73	5.00
MMDA	34.61	17.14	6.88	3.14	3.03	2.73	3.40
BLDA	**26.84**	13.77	5.93	3.38	3.16	2.73	3.20
LRA	27.97	**9.91**	**4.33**	**2.94**	**2.60**	2.73	**1.20**

Table 3. Average Classification Error Rate by Nearest Mean Classifier (upper table) and Nearest Neighbor Classifier (lower table) on Landsat Satellite Data Set

Dim	1	2	3	4	5	Rank
LDA	51.48	26.92	16.95	16.29	15.79	5.00
GMSS	30.75	27.49	19.27	16.46	15.79	5.75
HMSS	34.16	23.98	17.86	16.18	15.79	4.50
aPAC	34.62	19.58	**16.75**	16.27	15.79	3.00
MMDA	45.53	23.01	17.34	**15.85**	15.79	4.00
BLDA	**30.07**	21.31	16.97	16.29	15.79	3.25
LRA	31.53	**17.20**	17.04	15.89	15.79	**2.50**

Dim	1	2	3	4	5	Rank
LDA	54.72	28.16	18.71	15.87	14.25	7.00
GMSS	38.25	27.22	17.94	15.66	14.25	5.25
HMSS	36.76	25.69	18.07	15.01	14.25	4.75
aPAC	40.45	21.77	16.29	14.27	14.25	2.75
MMDA	48.73	24.64	17.92	14.46	14.25	4.00
BLDA	**35.77**	24.94	16.55	14.98	14.25	3.00
LRA	36.61	**20.46**	**16.03**	**13.98**	14.25	**1.25**

ments with 36 attributes from six classes. The set includes 4,435 training samples and 2,000 testing samples.

Fig 4 and 5 show the 2D representations of the testing samples obtained with LDA, BLDA, and LRA algorithms, one can see from the figure that the two data sets display non-Gaussianity and heteroscedasticity. The representations of LDA display severe class separation problem, especially on the Image Segmentation data set where three of the seven classes are totally overlapped. Although BLDA reduce the class overlapping of LDA to some extent, its 2D representations seem to be far from optimal on these two data sets whose class distributions are not Gaussian. In contrast, LRA provides dramatically separable 2D representations which are different from those of LDA and BLDA. For quantitative comparison, we measure the classification error using nearest mean classifier in these 2D subspaces. As expected, LRA achieves notably lower error rates than LDA and BLDA in both data sets, which indicates the proposed Max-K-Min criterion is better than the Bayes criterion based on restrictive distributional assumptions, when applied to the real data sets.

For comprehensive comparison, we further combine the training and test set together to conduct the five-fold cross-validation tests and evaluate the classification performance with varying dimensions. Six top-level linear discriminative dimension reduction methods, namely LDA [11][21], GMSS [24], HMSS [1], aPAC [17], BLDA [13], and MMDA [2], together with the proposed LRA, were compared. Tables 1 and 2 summarize the average classification error rate of all methods by using the nearest mean and the nearest neighbor classifier, respectively. The parameters of LRA are selected by 10-fold cross validation on the training data set.

For comparison purpose, we calculate an average rank of the used methods by averaging the rank of their performances on the dimensionalities from 1 to $C-2$ (columns of the tables). It should be noted that all the tested methods have the same performance when the dimensionality equals to $C-1$, because all discriminative information is contained exactly by the C class means and the C dimensional mean vectors are embedded exactly in a $(C-1)$-dimensional subspace. However, when subspace dimensionality is less than $C-1$, the sophisticated methods generally improve classical LDA. From these results, one can see from the tables that LRA consistently has the best average ranking in all the four test cases. In particular, on the nearest mean classification of the Image Segmentation data set, LRA achieves the average ranking of 1, suggesting it obtains the best solutions on class separation on all feature dimensions. The consistently better results of LRA over MMDA confirm our intuition: Max-K-Min criterion improve Max-Min criterion by considering both class separation and global scaling.

4.5. Classification on A Large Number of Classes

This experiment uses a well-know character recognition data set, the United States Postal Services (USPS) database, which contains 9,298 handwriting character measurements of 10 classes. The database is divided into two separated parts: a training set with 7,291 measurements and a test set with 2,007 measurements. Each measurement is a 256 dimensional vector. We use entire USPS database to evaluate the performances of LRA, and compare it against several top level DA methods.

In this experiment, LRA is applied in 20-dimensional subspace derived by ERE, so that the error rates of LRA and ERE are identical when the dimension is 20. As the digit classification involves 10 classes, the number of possible ordered sequences of projected means is upper bounded by $10!/2 = 1814400$. To improve the efficiency of BLDA and LRA, we use only the sequence of the projected means de-

Table 4. Classification Error Rate by NN Classifier on USPS Database

Dim	3	5	7	9	15	20
LDA	38.27	16.29	11.86	10.96	–	–
aPAC [8]	31.44	16.79	11.41	11.06	–	–
HDA [14]	35.18	25.16	19.88	17.34	–	–
WFLDA [12]	43.10	23.42	14.15	10.96	–	–
LFLDA [23]	34.73	18.09	13.35	11.01	8.82	7.42
ODA [3]	39.83	26.17	16.36	11.62	10.60	9.70
MODA [3]	39.50	28.50	15.76	10.32	10.27	9.37
GMSS[24]	**28.75**	15.74	11.16	9.87	5.98	5.83
ERE [15]	39.01	20.23	15.60	11.36	6.73	5.33
ERE+BLDA	30.34	16.29	10.26	8.92	6.48	5.33
ERE+LRA	29.60	**14.85**	**10.16**	**8.87**	**6.43**	5.33

termined by the dominant principal component of the class means, instead of searching through all possible sequences. Although this approximated method only considers a single sequence of ordered class means, the resulting LRA/BLDA still improves on ERE to a large extent. In particular, the average ranking of ERE is boosted from 9.25 to 1.25 by using LRA for subspace selection, which indicates that LRA has potential to enhance many other dimension reduction methods.

5. Conclusion

We extend the classical linear discriminant analysis (LDA) technique to linear ranking analysis (LRA), by considering the ranking order of classes centroids on the projected subspace. Under the constrain on the ranking order of the classes, two criteria are proposed: 1) minimization of the classification error with the assumption that each class is homogenous Guassian distributed; 2) maximization of the sum (average) of the k minimum distances of all neighboring-class (centroid) pairs. Both criteria can be efficiently solved by the convex optimization for one-dimensional subspace. Greedy algorithm is applied to extend the results to the multi-dimensional subspace. Experimental results show that 1) LRA with both criteria achieves state-of-the-art performance on the tasks of ranking learning and zero-shot learning; and 2) the maximum margin criterion is a potential discriminative subspace selection method, which significantly reduces the class separation problem in comparing with several representative extensions of LDA.

6. Acknowledgements

This work was partially sponsored by NSFC (National Natural Science Foundation of China) under Grant No. 61375031, No. 61005025, No. 61002051, and No. 61273217, the Fundamental Research Funds for the Central Universities, Beijing Higher Education Young Elite Teacher Program, and the Program for New Century Excellent Talents in University.

References

[1] W. Bian and D. Tao. Harmonic mean for subspace selection. In *ICPR*, pages 1–4. IEEE, 2008.

[2] W. Bian and D. Tao. Max-min distance analysis by using sequential sdp relaxation for dimension reduction. *PAMI*, 33(5):1037–1050, 2011.

[3] F. De la Torre and T. Kanade. Multimodal oriented discriminant analysis. In *ICML*, pages 177–184. ACM, 2005.

[4] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng. Emulating biological strategies for uncontrolled face recognition. *Pattern Recognition*, 43(6):2210–2223, 2010.

[5] W. Deng, J. Hu, J. Guo, W. Cai, and D. Feng. Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach. *Pattern Recognition*, 43(5):1748–1762, 2010.

[6] W. Deng, J. Hu, J. Lu, and J. Guo. Transform-invariant pca: A unified approach to fully automatic face alignment, representation, and recognition. *PAMI*, 99(PrePrints):1, 2013.

[7] W. Deng, Y. Liu, J. Hu, and J. Guo. The small sample size problem of ica: A comparative study and analysis. *Pattern Recognition*, 45(12):4438–4450, 2012.

[8] R. Duin and M. Loog. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *PAMI*, 26(6):732–739, 2004.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[10] V. Ferrari and A. Zisserman. Learning visual attributes. 2007.

[11] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[12] K. Fukunaga. Introduction to statistical pattern recognition. *Electrical science*, 1972.

[13] O. Hamsici and A. Martinez. Bayes optimality in linear discriminant analysis. *PAMI*, 30(4):647–657, 2008.

[14] B. Jelinek. Review on heteroscedastic discriminant analysis. *unpublished report, Center for Advanced Vehicular Systems, Mississippi State Univ*, 2001.

[15] X. Jiang, B. Mandal, and A. Kot. Eigenfeature regularization and extraction in face recognition. *PAMI*, 30(3):383–394, 2008.

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009.

[17] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *PAMI*, 23(7):762–766, 2001.

[18] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *PAMI*, 22(6):623–627, 2000.

[19] W. Ogryczak and A. Tamir. Minimizing the sum of the¡ i¿ k¡/i¿ largest functions in linear time. *Information Processing Letters*, 85(3):117–122, 2003.

[20] D. Parikh and K. Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[21] C. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[22] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *Trends and Topics in Computer Vision*, pages 1–14. Springer, 2012.

[23] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

[24] D. Tao, X. Li, X. Wu, and S. Maybank. Geometric mean for subspace selection. *PAMI*, 31(2):260–274, 2009.

[25] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, pages 537–544. IEEE, 2009.