Locality Adaptive Discriminant Analysis *

Xuelong Li, Mulin Chen, Feiping Nie, Qi Wang†
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
Xuelong_li@opt.ac.cn, chenmulin@mail.nwpu.edu.cn, feipingnie@gmail.com, crabwq@nwpu.edu.cn

Abstract

Linear Discriminant Analysis (LDA) is a popular technique for supervised dimensionality reduction, and its performance is satisfying when dealing with Gaussian distributed data. However, the neglect of local data structure makes LDA inapplicable to many real-world situations. So some works focus on the discriminant analysis between neighbor points, which can be easily affected by the noise in the original data space. In this paper, we propose a new supervised dimensionality reduction method, *Locality Adaptive Discriminant Analysis* (LADA), to learn a representative subspace of the data. Compared to LDA and its variants, the proposed method has three salient advantages: (1) it finds the principle projection directions without imposing any assumption on the data distribution; (2) it's able to exploit the local manifold structure of data in the desired subspace; (3) it exploits the points' neighbor relationship automatically without introducing any additional parameter to be tuned. Performance on synthetic datasets and real-world benchmark datasets demonstrate the superiority of the proposed method.

1 Introduction

Dimensionality reduction is fundamentally important for analyzing high-dimensional data, and have received sufficient attention in the filed of artificial intelligence [Zhang *et al.*, 2011; Peng *et al.*, 2016]. The goal of dimensionality reduction is to embed the data into a low-dimensional subspace, while retaining the desired discriminant information. Among the numerous dimensionality reduction algorithms, *Principal Component Analysis* (PCA) [Wold *et al.*, 1987] and *Linear Discriminant Analysis* (LDA) [Friedman and Kandel, 1999] are the most widely used techniques. Here we mainly focus on the LDA, since it outperforms PCA in many cases.

As the most popular supervised dimensionality reduction method, LDA aims to find a linear transformation matrix W which minimizes the separation within each class and simultaneously maximizes the discrepancy between different classes. However, it has three major disadvantages. First, LDA suffers from the over-reducing problem [Wan *et al.*, 2015]. Denoting the number of classes as C, the rank of the between-class scatter matrix S_b is at most $C-1$ [Wan *et al.*, 2015]. As a result, LDA could find at most $C-1$ projection directions, which are insufficient for tasks with just a few classes, such as binary classification. Second, the *Small Sample Size* (SSS) problem [Lu *et al.*, 2005] often occurs. When the dimension of data exceeds the number of training samples, the within-class scatter matrix S_w becomes singular, which makes LDA unsuitable for the data with very high dimensionality. Third, LDA assumes that the input data obeys the Gaussian distribution globally. However, in real world applications, the data may be multimodally distributed, which means that each class has a unique distribution. LDA fails in these occasions because it can't capture the underlying data structure in the local area.

In the past several decades, plenty of methods are proposed to address the above drawbacks. Among them, a number of works [Sharma and Paliwal, 2015; Kumar and Agrawal, 2016; Wan *et al.*, 2015] have been conducted to avoid the over-reducing and SSS problems, and the achieved performance is satisfying. However, the third problem of LDA is not well-solved yet. Though many algorithms are developed to investigate the local data structure by applying LDA to neighbor points, they share the same problem that the neighbors found in the original data space are not reliable to reveal the intrinsic local structure, especially when the noise is large.

In this paper, a new supervised dimensionality reduction method, *Locality Adaptive Discriminant Analysis* (LADA), is proposed to investigate the geometry of local data structure. Similar to the existing locality-aware approaches, the proposed LADA focuses on the data points with close relationship. The major difference is that LADA exploits the points' local relationship adaptively in the learned subspace, and doesn't involve a kNN procedure. Moreover, benefiting from the new objective function and the optimization strategy, LADA replaces S_b with the covariance matrix, and it doesn't calculate the inverse of the within-class scatter matrix, so the over-reducing and SSS problems don't exist naturally. The salient merits of LADA are summarized as follows:

1. LADA doesn't rely on any arbitrary assumption about

*This work is supported by the National Natural Science Foundation of China under Grant 61379094.

†Qi Wang is the corresponding authors.

the data distribution, and doesn't have the over-reducing and SSS problems.

2. LADA has the capability to capture the neighbor relationship of data points in the desired subspace, and exploit the intrinsic local structure of data manifold.

3. LADA doesn't involve a kNN processing, so no additional efforts are needed to tune the parameter k.

2 Review of LDA and its Locality-Aware Variants

In this section, we briefly review the classical LDA. There are many variants of LDA [Suzuki and Sugiyama, 2013; Flamary *et al.*, 2016; Ren *et al.*, 2015; Xu *et al.*, 2017], we mainly discuss the some locality-aware techniques which are proposed to investigate the local data structure.

2.1 Linear Discriminant Analysis

Given the data matrix $\mathrm{X} = [x_1, x_2, \cdots, x_n], x_j \in \mathbb{R}^{d\times 1}$ with C classes, the purpose of LDA is to learn a linear transformation matrix $W \in \mathbb{R}^{d\times m}$ ($m \ll d$) to map the d-dimensional data x_j to a m-dimensional vector:

$$y_i = W^T x_i. \tag{1}$$

LDA supposes that a optimal transformation should push the data points from different classes far away from each other while pulling those within the same class close to each other. So the objective of LDA can be written as

$$\max_W \frac{\sum_{i=1}^{C} n_i ||W^T(\mu^i - \mu)||_2^2}{\sum_{i=1}^{C}\sum_{j=1}^{n_i} ||W^T(x_j^i - \mu^i)||_2^2}, \tag{2}$$

where n_i is the number of samples in class i, μ^i is the mean of the samples in class i, μ is the mean of all the samples, and x_j^i is the j-th sample in class i. Denote the between-class scatter matrix S_b and the within-class scatter matrix S_w as

$$S_b = \sum_{i=1}^{C} n_i(\mu^i - \mu)(\mu^i - \mu)^T, \tag{3}$$

$$S_w = \sum_{i=1}^{C}\sum_{j=1}^{n_i} (x_j^i - \mu^i)(x_j^i - \mu^i)^T, \tag{4}$$

then the problem 2 can be rewritten into a concise form:

$$\max_W \frac{tr(W^T S_b W)}{tr(W^T S_w W)}, \tag{5}$$

where $tr()$ indicates the trace operator. Due to the complexity to solve the above trace ratio problem, many researchers transform it into a ratio trace form,

$$\max_W tr(\frac{W^T S_b W}{W^T S_w W}). \tag{6}$$

From the objective function, it can be clearly seen that LDA just emphasizes the global relationship of data, which makes it unable to discover the local structure. So some locality-aware variants are proposed to address this drawback.

2.2 Locality-Aware Variants of LDA

To deal with multimodally distributed data, it's essential to investigate the local structure of data manifold [Nie *et al.*, 2009]. For this purpose, some researchers perform discriminant analysis in the local data space instead.

Bressan and Vitria [2003] redefined the between-class scatter matrix S_b as the distances between the data points and their extra-class nearest neighbors, and defined the within-class scatter matrix S_w as those for the intra-class nearest neighbors. Sugiyama [2006] learned a maximum margin between the extra-class and intra-class k neighbors, and transformed the objective into a ratio trace form, which leads to the suboptimal solution [Jia *et al.*, 2009]. Cai *et al.* [2007] found the k nearest neighbors of each point, and used the neighbors to replace the data center in original LDA. Similar to LSDA, Nie *et al.* [2007] also learned a maximum margin between the extra-class and intra-class k neighbors, but it formulated the objective function as a trace ratio problem and solved it in an efficient way, so it achieved relatively better performance. Weinberger and Saul [2009] put forward a cost function to penalize the large distances between a point and it k nearest neighbors. Different from the above methods, Fan *et al.* [2011] found the k nearest neighbors of a test sample from the training set, and learned a transformation matrix for each test sample separately, so it's time-consuming when the number of test samples is large.

A shortcoming shared by all these methods is that they find the neighbors of points based on their distances in the original data space, which is unreliable. The intrinsically similar points may be far away from each other in the original space, especially for data with large noise. So these methods are sensitive to the data noise.

3 Locality Adaptive Discriminant Analysis

In this section, the Locality Adaptive Discriminant Analysis (LADA) method for dimensionality reduction is presented. First, the objective function of LADA is described and theoretically analyzed. Then, an adaptive learning strategy is designed to obtain the optimal solution.

3.1 Problem Formulation

In real-world applications, such as face classification, the input data may be multimodally distributed. So it's essential to capture the local structure of data manifold. Our goal is to learn an optimal transformation matrix W to pull the similar points together while pushing the dissimilar ones far away from each other.

Given the data points $\mathrm{X} = [x_1, x_2, \cdots, x_n], x_j \in \mathbb{R}^{d\times 1}$, the objective function is defined as

$$\begin{aligned} &\min_{W,s} \frac{\sum_{i=1}^{C} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \left\|W^T(x_j^i - x_k^i)\right\|_2^2}{\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\|W^T(x_j - x_k)\|_2^2} \\ &s.t. W^T W = I, \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0, \end{aligned} \tag{7}$$

where n is the number of samples, s is a weighted matrix, s_{jk}^i means the weight between the j-th and k-th sample in

class i, and the remaining definitions are the same as those in LDA. In the above function, s is squared to avoid the trivial solution. Note that, x_j is the j-th sample in the whole data set, and it's different from x_j^i.

In problem (7), the weighted matrix s is introduced to capture the local relationship between data points. The constraints on s avoids the case that some rows of s are all zeros. Supposing the transformation W is already obtained, s_{jk}^i will be large if the transformed distance $||W^T(x_j^i - x_k^i)||_2^2$ is small, which means x_j^i and x_k^i are similar in the learned subspace. In the next step, if we fix s and optimize W again, the objective function will emphasize the similar points in the previously learned subspace. Consequently, the points' relationship in the desired subspace can be learned by optimizing s and W iteratively.

3.2 Optimization Strategy

Here an adaptive learning strategy is presented to solve problem (7). First, the weight of the points in the class i is initialized as $\frac{1}{n_i}$, and the weight of points from different classes is set to 0. Then the optimal solution can be computed by solving W and s iteratively.

When s is fixed, denoting $\tilde{S}_t$ and $\tilde{S}_w$ as

$$\tilde{S}_t = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(x_j - x_k)(x_j - x_k)^T, \quad (8)$$

$$\tilde{S}_w = \sum_{i=1}^{C} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 (x_j^i - x_k^i)(x_j^i - x_k^i)^T, \quad (9)$$

then problem (7) becomes

$$\min_{W^TW=I} \frac{tr(W^T\tilde{S}_wW)}{tr(W^T\tilde{S}_tW)}, \quad (10)$$

where $tr()$ indicates the trace operator. The above trace ratio problem can be efficiently solved by the optimization algorithm in [Nie *et al.*, 2007].

When W is fixed, the objective function (7) can be reduced to

$$\begin{aligned} &\min_{s} \sum_{i=1}^{C}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 ||W^T(x_j^i - x_k^i)||_2^2 \\ &s.t. \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0, \end{aligned} \quad (11)$$

which is equivalent to the following problem

$$\begin{aligned} &\min_{s_j^i} \sum_{k=1}^{n_i} {s_{jk}^i}^2 ||W^T(x_j^i - x_k^i)||_2^2 \\ &s.t. \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0, \end{aligned} \quad (12)$$

where s_j^i is a column vector with its k-th element equal to s_{jk}^i. Denoting a column vector α equal to s_j^i, and denoting v_k equal to $||W^T(x_j^i - x_k^i)||_2^2$, the above problem is simplified to

$$\min_{\alpha^T\mathbf{1}=1,\alpha\geq 0} \sum_{k=1}^{n_i} \alpha_k^2 v_k. \quad (13)$$

Defining a diagonal matrix V with V_{kk} equal to v_k, problem (13) becomes

$$\min_{\alpha^T 1=1,\alpha\geq 0} \alpha^T V\alpha. \quad (14)$$

Without the second constraint $\alpha \geq 0$, the Lagrangian function of problem (14) is

$$\mathcal{L}(\alpha,\eta) = \alpha^T V\alpha - \eta(\alpha^T\mathbf{1} - 1), \quad (15)$$

where η is the Lagrangian multiplier. Taking the derivative of Eq. (15) w.r.t. α and setting it to zero, we get

$$2V\alpha - \eta\mathbf{1} = 0. \quad (16)$$

Together with the constraint $\alpha^T\mathbf{1} = 1$, the α can be computed as

$$\alpha_k = \frac{1}{v_k} \times (\sum_{t=1}^{n_i} \frac{1}{v_t})^{-1}. \quad (17)$$

Fortunately, the above α satisfies the constraint $\alpha \geq 0$, so it's also the optimal solution to the problem (14). Accordingly, the optimal solution to the problem (11) is

$$s_{jk}^i = \frac{1}{||W^T(x_j^i - x_k^i)||_2^2} \times (\sum_{t=1}^{n_i} \frac{1}{||W^T(x_j^i - x_t^i)||_2^2})^{-1}. \quad (18)$$

By optimizing W and s iteratively, our method is capable of quantifying the data points' local relationship in the desired subspace. Unlike existing locality-aware algorithms, our method is totally self-weighted, and saves the efforts to tune parameters. The complete algorithm is shown in Alg. 1.

Algorithm 1 The algorithm of LADA for dimensionality reduction

Input: Data matrix $\mathrm{X} = [x_1, x_2, \cdots, x_n], x_j \in \mathbb{R}^{d\times 1}$, desired dimension m

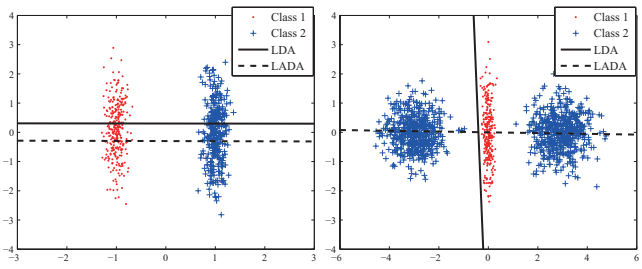
1: Initialize weight matrix s
2: **repeat**
3: compute the optimal W of problem (10)
4: update s with Eq. (18)
5: **until** Converge
6: $Y = W^TX$

Output: $\mathrm{Y} = [y_1, y_2, \cdots, y_n], y_j \in \mathbb{R}^{m\times 1}$

The objective is monotonically decreased in each iteration, and converges to the lower bound finally. In addition, for LADA, the over-reducing problem doesn't exist because $\tilde{S}_b$ is of full rank. And the Small Sample Size problem is also avoided because our learning algorithm doesn't calculate the inverse of $\tilde{S}_w$.

4 Connection to LDA

In this section, we show the close connection between the proposed LADA and the original LDA. LADA and LDA share the similar goals to maximize the between-class scatter matrix while minimizing the within-class scatter matrix. In fact, when the s_{jk}^i in problem (7) is set as $1/n_i$, LADA becomes equivalent to LDA. A theorem is proposed to support this statement.



(a) Unimodal (b) Multimodal

Figure 1: Projection directions of LDA (dashed) and LADA. LDA finds the correct direction when the class density is unimodal (a), but fails on mutimodally distributed data (b). LADA works well on both cases.

Theorem 1. *When s_{jk}^i equals to $\frac{1}{n_i}$, $\tilde{S}_w$ equals to $2 \times S_w$.*

Proof. If $s_{jk}^i = \frac{1}{n_i}$, $\tilde{S}_w$ can be derived into the following form

$$\begin{aligned}
&\tilde{S}_w \\
&= \sum_{i=1}^{C} n_i \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} (\tfrac{1}{n_i})^2 (x_j^i - x_k^i)(x_j^i - x_k^i)^T \\
&= \sum_{i=1}^{C} \tfrac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} (x_j^i x_j^{iT} + x_k^i x_k^{iT} - 2x_j^i x_k^{iT}) \\
&= \sum_{i=1}^{C} (\sum_{j=1}^{n_i} x_j^i x_j^{iT} + \sum_{j=1}^{n_i} x_k^i x_k^{iT} - \tfrac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} 2x_j^i x_k^{iT}) \\
&= 2 \sum_{i=1}^{C} (\sum_{j=1}^{n_i} x_j^i x_j^{iT} - \tfrac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} x_j^i x_k^{iT}) \\
&= 2 \sum_{i=1}^{C} (\sum_{j=1}^{n_i} x_j^i x_j^{iT} - \tfrac{1}{n_i} \sum_{j=1}^{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT}).
\end{aligned} \tag{19}$$

And S_w can be derived as

$$\begin{aligned}
&S_w \\
&= \sum_{i=1}^{C} \sum_{j=1}^{n_i} (x_j^i - \mu^i)(x_j^i - \mu^i)^T \\
&= \sum_{i=1}^{C} \sum_{j=1}^{n_i} (x_j^i - \tfrac{1}{n_i} \sum_{k=1}^{n_i} x_k^i)(x_j^i - \tfrac{1}{n_i} \sum_{k=1}^{n_i} x_k^i)^T \\
&= \sum_{i=1}^{C} \sum_{j=1}^{n_i} (x_j^i x_j^{iT} - \tfrac{2}{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT} + \tfrac{1}{{n_i}^2} \sum_{j=1}^{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT}) \\
&= \sum_{i=1}^{C} (\sum_{j=1}^{n_i} x_j^i x_j^{iT} - \tfrac{2}{n_i} \sum_{j=1}^{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT} + \tfrac{1}{n_i} \sum_{j=1}^{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT}) \\
&= \sum_{i=1}^{C} (\sum_{j=1}^{n_i} x_j^i x_j^{iT} - \tfrac{1}{n_i} \sum_{j=1}^{n_i} x_j^i \sum_{k=1}^{n_i} x_k^{iT}).
\end{aligned} \tag{20}$$

From Eq. (19) and (20), we can see $\tilde{S}_w = 2S_w$. □

According to the above proof, the scatter matrix $\tilde{S}_t$ can be written as

$$\begin{aligned}
&\tilde{S}_t \\
&= \tfrac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (x_j - x_k)(x_j - x_k)^T \\
&= 2 \sum_{j=1}^{n} (x_j - \mu)(x_j - \mu)^T.
\end{aligned} \tag{21}$$

It's easy to prove that $S_w + S_b = \sum_{j=1}^{n} (x_j - \mu)(x_j - \mu)^T$, so we have

$$\tilde{S}_t = 2(S_w + S_b). \tag{22}$$

Thus, when $s_{jk}^i = \frac{1}{n_i}$, the objective function of LADA can be transformed into

$$\min_{W^T W = I} \frac{tr(W^T S_w W)}{tr(W^T (S_w + S_b) W)}, \tag{23}$$

which is equivalent to the problem (5) in LDA.

However, compared with LDA, the proposed method doesn't pull the far away points within the same class closer, and it emphasizes more on the local structure of data manifold. So LADA is able to handle data with multimodal distribution. To illustrate this point, a toy example is provided. As shown in Figure 1(a), both LDA and LADA find the correct projection directions on unimodally distributed data. However, when the data distribution is multimodal, as shown in Figure 1(b), LDA fails while LADA still works well.

5 Experiments

In this section, the proposed LADA is evaluated on a toy dataset and five real-world datasets. Throughout the experiments, we let the competitors utilize their respective optimal parameters.

5.1 Performance on Toy Dataset

In this work, two toy datasets are built to validate the effectiveness of LADA.

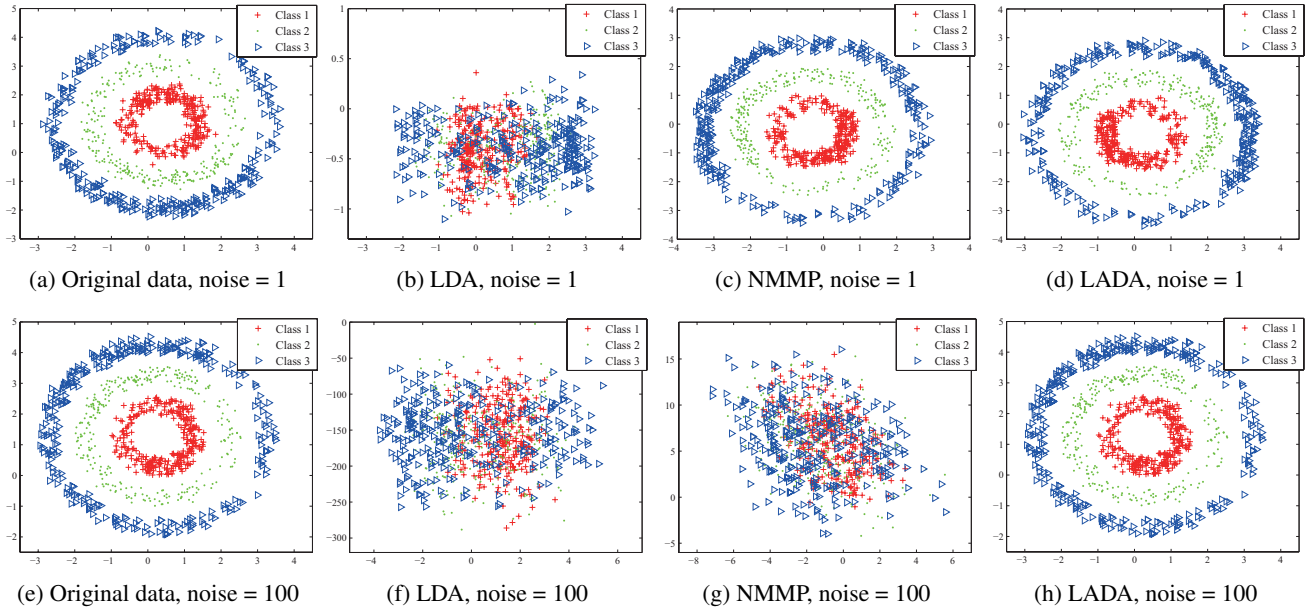
Dataset: As Figure 2 visualizes, each dataset consists of data from three classes. In the first two dimensions, the data points are distributed in concentric circles, while the other eight dimensions are noises, which are randomly generated in the range of 0 and c. For the two datasets, the noise c is set as 1 and 100 respectively.

Competitors: The proposed LADA is compared with LDA [Friedman and Kandel, 1999], and a state-of-the-art locality-aware method Neighborhood MinMax Projections (NMMP) [Nie *et al.*, 2007].

Performance: Figure 2 shows the two-dimensional subspaces found by LDA, NMMP and LADA. It's manifest that the subspace of LADA preserves the manifold structure of original data with more discriminability. LDA focuses on the global aspect, and imposes the far away points within the same class to be close to each other, so it can't capture the local data structure. Because of the investigation of local structure, NMMP performs well when the noise is small. However, since NMMP relies on the points' distances in the original data space, consequently, it fails when the noise is large. The proposed LADA adaptively captures the points' local relationship in the learned subspace, so it shows robustness to the noise and learns the discriminant subspace successfully.

5.2 Performance on Real-World Dataset

In this part, experiments are conducted on various real-world datasets to demonstrate the usefulness of LADA. First, the input data is transformed into a low-dimensional subspace.



(a) Original data, noise = 1 (b) LDA, noise = 1 (c) NMMP, noise = 1 (d) LADA, noise = 1

(e) Original data, noise = 100 (f) LDA, noise = 100 (g) NMMP, noise = 100 (h) LADA, noise = 100

Figure 2: (a) and (e) are the first two dimensions of the two datasets. (b)-(e) and (f)-(h) show the two-dimensional subspaces learned by LDA, NMMP and LADA on the two datasets. The performance of LDA is unsatisfying due to the neglect of local manifold structure. NMMP fails when noise is 100, because it can't find the correct neighbor points. Under different noises, LADA is able to find the discriminant subspace while preserving the manifold structure.

	USPS	Mnist	Yale Face	AR Face	ALLAML
class number	5	6	15	50	2
training number	180	6000	120	350	20
testing number	5244	30012	45	950	52
input dimensionality	256	784	6400	4980	7129
dimensionality after PCA	79	283	88	508	66

Table 1: Description of the real-world datasets.

	USPS	Mnist	Yale Face	AR Face	ALLAML
Baseline	91.6±0.68	96.4±6.01	71.9±6.19	41.6±1.38	80.6±5.49
LDA	83.1±2.26	19.1±8.35	72.4±5.40	52.3±2.33	58.9±9.56
NDA	87.1±1.25	<u>96.3±9.56</u>	72.3±5.31	71.7±9.93	78.3±6.38
LLDA	86.3±0.99	81.3±7.63	76.9±5.43	57.8±2.73	76.4±7.05
LSDA	86.7±1.74	95.9±8.34	59.0±6.97	49.6±2.17	80.5±7.79
LFDA	84.6±1.58	75.8±9.40	74.2±5.87	70.6±8.59	76.3±8.20
NMMP	**90.7±0.86**	78.6±7.94	<u>92.4±2.89</u>	<u>73.3±1.76</u>	<u>84.6±6.60</u>
LADA	<u>90.4±0.87</u>	**96.8±6.33**	**93.8±3.52**	**83.4±1.60**	**86.2±5.60**

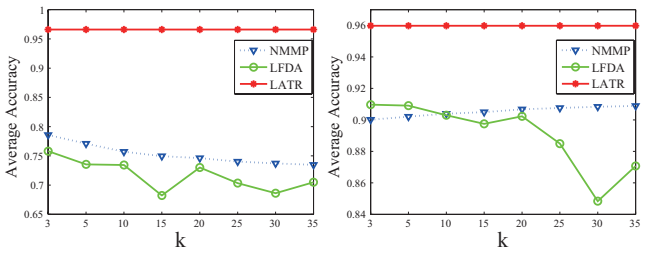
Table 2: Average classification accuracy over 30 random splits of different datasets (mean±standard deviation%). Best results are in bold face, and the second-best results are underlined. The results show that the proposed LADA achieves satisfying performance on different kinds of data.

Then the nearest neighbor classifier is used to classify the obtained low-dimensional data. For each dataset, we randomly choose several samples for training, and use the remaining samples for testing. After 30 random splits, the averaged classification accuracies [Wang *et al.*, 2017] and standard deviations are reported.

Datasets: The proposed LADA is evaluated on five standard benchmarks, USPS [Hull, 1994], Mnist [LeCun *et al.*, 1998], Yale Face [Georghiades *et al.*, 2001], AR Face [Ding and Martinez, 2010], and AMLALL [Golub *et al.*, 1999]. USPS is a handwritten digit image dataset, and the size of each image is 16×16. We use the digits 1, 2, 3, 4 and 5 as the five classes for classification. The Mnist dataset consists of 70000 handwritten digit images, and the image size is 28×28. Digits 1, 2, 3, 4, 5 and 6 are employed in the experiments. Yale Face dataset contains 165 face images of 15 individuals under different facial expressions and facial details [Nie *et al.*, 2007]. Each image is of size 112×92, and we down-sample them into 80×80 in the simulation. AR face includes over 4000 color images of 126 persons' faces (70 males and 56 males). Each individual has 26 image, and the image size is 165×120. In our experiments, we choose the images of the first 50 males for classification. The color images are down-sampled into 83×60, and transformed into 256 gray levels. The ALLAML dataset consists of 7129 probes of human genes for cancer classification. There are 47 samples from the acute lymphoblastic leukemia (ALL) type and 25 samples from the acute myeloid leukemia (AML) type. A brief description of the datasets is shown in Table 1. For all the datasets, Principal Component Analysis (PCA) [Wold *et al.*, 1987] is performed as the preprocessing step to speed up, and the desired projection direction number is set as 60.

Competitors: To verify the superority of LADA, LDA [Friedman and Kandel, 1999] and five state-of-the-art locality-aware methods Nonparametric discriminant analysis (NDA) [Bressan and Vitria, 2003], Local Linear Discriminant Analysis (LLDA) [Fan *et al.*, 2011], Locality Sensitive Discriminant Analysis (LSDA) [Cai *et al.*, 2007], Local Fisher Discriminant Analysis (LFDA) [Sugiyama, 2006] and Neighborhood MinMax Projections (NMMP) [Nie *et al.*, 2007] are taken for comparison. Moreover, The classification result directly performed after PCA is taken as the baseline.

Performance: Table 2 illustrates the averaged classification accuracies and standard deviations of different methods. It can be seen that the proposed LADA achieves the best performance on Mnist, Yale Face, AR Face and ALLAML, and obtains the second best result on USPS. LDA and LLDA both fail on the Mnist, AR Face and ALLAML because the over-reducing problem occurs, so they can only find limited number of projection directions, which are insufficient to classify the samples correctly. Even though, LLDA works a little better than LDA because it exploits the local data structure. NMMP outperforms the other neighbor-based methods in most cases because it finds the global optimal solution by solving a trace ratio problem, but it's inferior to the proposed LADA because of data noise. On the AR Face dataset, the baseline result is unsatisfying, which indicates the large data noise. Thus, the good performance of LADA validates its robustness to noise.



(a) Mnist1 (b) Mnist2

Figure 3: Average classification accuracy on (a) Mnist1 and (b) Mnist2 with varying k. It can be seen that the performance of NMMP and LFDA change with k.

In addition, all of the above locality-aware methods rely on a kNN processing. However, the choice of the parameter k may affect the final results, so these methods are not so practical as LADA. For a better interpretation, we compare the performance of NMMP and LFDA under different occasions. The experiments are conducted on two subsets of Mnist dataset, as shown in Figure 3, Mnist1 contains the digits images of 1, 2, 3, 4, 5 and 6, and Mnist2 includes the images of 7, 8, 9 and 0. Figure 3 shows that the performance of NMMP and LFDA change with k, so the decision of a proper k is the basis of these two methods. However, even on Mnist1 and Mnist2, which belong to the same dataset, the optimal k of NMMP is different. Therefore, it's impractical to chose a k that works well for various applications. The proposed LADA produces good results steadily because it doesn't depend on any additional parameter.

6 Conclusion

In this work, we propose a new supervised dimensionality reduction method called *Locality Adaptive Discriminant Analysis* (LADA). LADA focuses on the points which are intrinsically similar, and then pulls them close to each other after the linear transformation. So it is able to discover the underlying local structure of data manifold. Compared to classical LDA, the proposed LADA is more suitable to deal with multimodally distributed data. In comparison with existing locality-aware algorithms, LADA is more robust to the data noise, and saves the efforts to tune additional parameters. Furthermore, the well-known over-reducing and Small Sample Size problems don't exist naturally in our method. Experimental results on various datasets demonstrate that our method outperforms the state-of-the-art techniques. In the future work, we would like to extend our method to non-linear discriminant analysis by introducing a kernel function. In addition, it's also desirable to apply LADA in more real-world applications.

References

[Bressan and Vitria, 2003] Marco Bressan and Jordi Vitria. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(15):2743–2749, 2003.

[Cai *et al.*, 2007] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao. Locality sensitive discriminant analysis. In *International Joint Conference on Artificial Intelligence*, pages 708–713, 2007.

[Ding and Martinez, 2010] Liya Ding and Aleix M Martinez. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2022–2038, 2010.

[Fan *et al.*, 2011] Zizhu Fan, Yong Xu, and David Zhang. Local linear discriminant analysis framework using sample neighbors. *IEEE Transactions on Neural Networks*, 22(7):1119–1132, 2011.

[Flamary *et al.*, 2016] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *CoRR*, abs/1608.08063, 2016.

[Friedman and Kandel, 1999] Menahem Friedman and Abraham Kandel. *Introduction to Pattern Recognition - Statistical, Structural, Neural and Fuzzy Logic Approaches*, volume 32 of *Series in Machine Perception and Artificial Intelligence*. WorldScientific, 1999.

[Georghiades *et al.*, 2001] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

[Golub *et al.*, 1999] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[Jia *et al.*, 2009] Yangqing Jia, Feiping Nie, and Changshui Zhang. Trace ratio problem revisited. *IEEE Transaction on Neural Networks*, 20(4):729–735, 2009.

[Kumar and Agrawal, 2016] Nitin Kumar and RK Agrawal. Two-dimensional exponential discriminant analysis for small sample size in face recognition. *International Journal of Artificial Intelligence and Soft Computing*, 5(3):194–208, 2016.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lu *et al.*, 2005] Juwei Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181–191, 2005.

[Nie *et al.*, 2007] Feiping Nie, Shiming Xiang, and Changshui Zhang. Neighborhood minmax projections. In *International Joint Conference on Artificial Intelligence*, pages 993–998, 2007.

[Nie *et al.*, 2009] Feiping Nie, Shiming Xiang, Yangqiu Song, and Changshui Zhang. Extracting the optimal dimensionality for discriminant analysis. *Pattern Recognition*, 42(1):105–114, 2009.

[Peng *et al.*, 2016] Xi Peng, Jiwen Lu, Zhang Yi, and Yan Rui. Automatic subspace learning via principal coefficients embedding. *IEEE Trans. Cybern.*, PP(99):1–14, 2016.

[Ren *et al.*, 2015] Chuan-Xian Ren, Dao-Qing Dai, Xiaofei He, and Hong Yan. Sample weighting: An inherent approach for outlier suppressing discriminant analysis. *IEEE Trans. Knowl. Data Eng.*, 27(11):3070–3083, 2015.

[Sharma and Paliwal, 2015] Alok Sharma and Kuldip K Paliwal. Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics*, 6(3):443–454, 2015.

[Sugiyama, 2006] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *International conference on Machine learning*, pages 905–912, 2006.

[Suzuki and Sugiyama, 2013] Taiji Suzuki and Masashi Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 25(3):725–758, 2013.

[Wan *et al.*, 2015] Huan Wan, Gongde Guo, Hui Wang, and Xin Wei. A new linear discriminant analysis method to address the over-reducing problem. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 65–72, 2015.

[Wang *et al.*, 2017] Qi Wang, Mulin Chen, and Xuelong Li. Quantifying and detecting collective motion by manifold learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,*, pages 4292–4298, 2017.

[Weinberger and Saul, 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[Wold *et al.*, 1987] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[Xu *et al.*, 2017] Xiao-Lin Xu, Chuan-Xian Ren, Ran-Chao Wu, and Hong Yan. Sliced inverse regression with adaptive spectral sparsity for dimension reduction. *IEEE Trans. Cybernetics*, 47(3):759–771, 2017.

[Zhang *et al.*, 2011] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision*, pages 471–478, 2011.