An empirical study of two typical locality preserving linear discriminant analysis methods

Lishan Qiao^{1,2}, Limei Zhang^{1,2}, Songcan Chen^{1,*}

^{1} Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, 210016, Nanjing, P.R. China

^{2} Department of Mathematics Science, Liaocheng University, 252000, Liaocheng, P.R. China

Abstract: Laplacian Linear Discriminant Analysis (LapLDA) and Semi-supervised Discriminant Analysis (SDA) are two recently proposed LDA methods. They are developed independently with the aim to improve LDA by introducing a locality preserving regularization term, and have been shown their effectiveness experimentally on some benchmark datasets. However, both algorithms ignored comparison with much simpler methods such as Regularized Discriminant Analysis (RDA). In this paper, we make an empirical and supplemental study on LapLDA and SDA, and get somewhat counterintuitive results: 1) Although LapLDA can generally improve the classical LDA via resorting to a complex regularization term, it does not outperform RDA which is only based on the simplest Tikhonov regularizer; 2) To reevaluate the performance of SDA, we develop purposely a new and much simpler semi-supervised algorithm called Globality Preserving Discriminant Analysis (GPDA) and make a comparison with SDA. Surprisingly, we find that GPDA tends to achieve better performance. These two points drive us to reconsider whether one should use or how to use locality preserving strategy in practice. Finally, we discuss the reasons which lead to the possible failure of the locality preserving criterion and provide alternative strategies and suggestions to address these problems.

Key words: Linear Discriminant Analysis, Semi-supervised learning, Locality preserving regularizer, Graph construction

1 Introduction

Linear Discriminant Analysis (LDA) is a popular supervised dimensionality reduction (DR) method. It seeks a set of optimal linear projections by simultaneously maximizing the between-class dissimilarity and minimizing the within-class dissimilarity. On one hand, LDA is simple, efficient and effective for many practical applications such as face recognition[1]; but on the other hand, it does suffer from some problems: 1) the so-called Small Sample Size (SSS) problem, where the training sample size is small compared with the high feature dimensions; 2) LDA is a completely global method, and thereby can not naturally capture the local geometry of the data. To deal with the first problem, many extended LDA algorithms, such as PseudoLDA[2], PCA+LDA[1], NullLDA[3], LDA/QR[4], RDA[5] and 2DLDA[6], have been developed in the past decade years. To deal with the second problem, a simple but feasible strategy is to incorporate a data-dependent regularization term into the original objective function of LDA, and the representative methods have Chen et al.'s Laplacian Linear Discriminant Analysis (LapLDA)[7]

[*] Corresponding author: Tel: +86-25-84896481 Ext. 12221; Fax: +86-25-84892400; E-mail: s.chen@nuaa.edu.cn (S. Chen)

and Cai et al.’s Semi-supervised Discriminant Analysis (SDA)[8]. The former is supervised, while the latter is semi-supervised. But they share a nearly similar objective function which integrates Fisher criterion with a locality preserving regularizer and attempt to get the best of both worlds. In fact, in the recent years, many researchers have independently developed some similar DR methods such as Song et al.’s semi-supervised LDA (SSLDA)[9] which is actually equivalent to SDA.

Undoubtedly, it is important and attractive to develop new and effective DR algorithms. However, on the other hand we believe that further comparative study on the existing methods is also quite valuable, since it may correct some misunderstanding for users, guide practitioners in choosing appropriate methods, and help to design better algorithms. In this paper, we make an empirical discussion on two recently proposed and relatively popular locality preserving LDA methods, i.e., the aforementioned LapLDA and SDA. All experiments are based on the same benchmark datasets which have been employed to verify their own effectiveness. The main insights and contributions of this paper include:

1) According to [7], LapLDA can generally outperform LDA by introducing a locality preserving regularization term. However, our further experiments show that LapLDA does not outperform RDA, a regularized counterpart of the LDA, which is developed earlier and only resorts to much simpler Tikhonov regularizer. This indicates that the role of locality preserving term is not so important as [7] claimed, which drives us to reconsider its limitation or inappropriateness in characterizing the real geometric structure in data.
2) SDA is designed especially for semi-supervised scenario and shares the similar locality preserving regularizer as in LapLDA. Also, the authors of [8] experimentally validated that their SDA benefits from such local regularization term and outperforms many state-of-the-art methods such as LapSVM. To reevaluate its performance, we develop purposely a new and much simpler semi-supervised algorithm called Globality Preserving Discriminant Analysis (GPDA) and make a comparison with SDA. Consequently, such a relatively simple algorithm tends to achieve better performance, even though only the global structure of the data is considered.
3) Different from many existing studies [7-8, 10] which mainly focus on the advantages of locality preserving strategy, we discuss the reasons why the locality preserving algorithms may not work well. Moreover, motivated by the above insights, we provide some alternative strategies and suggestions to address this problem.

The rest of this paper is organized as follows: Section 2 briefly introduces several regularization based LDA methods which are closely associated with our topic. In section 3, we empirically compare LapLDA and SDA with two simple baseline methods. In section 4, we make further discussion on the locality preserving strategy. Finally, we conclude the paper in section 5.

2 Regularization based LDA methods

In this section, we briefly introduce two locality preserving LDA methods (i.e., LapLDA and

SDA), and two baseline methods including the existing RDA and a newly proposed GPDA which is designed just for revaluating the effectiveness of both LapLDA and SDA. All these methods are closely related to regularization technique which plays an important role in many popular machine learning algorithms such as SVM and LapSVM.

2.1 Regularized Discriminant Analysis (RDA)

It is well known that LDA seeks an optimal projection matrix by maximizing the following Fisher's criterion:

$$W^* = \arg\max_W \; tr(W^T S_b W) / tr(W^T S_t W) \tag{1}$$

where $tr(\cdot)$ denotes the trace operator, S_b and S_t are respectively the between-class scatter matrix and the total scatter matrix. Here, we use the total scatter matrix S_t instead of the within-class scatter matrix S_w in the Fisher's criterion to keep consistent with the form used in LapLDA and SDA. The optimal W^* are built by the eigenvectors corresponding to the eigen-problem: $S_t^{-1} S_b w = \lambda w$.

Despite its simplicity and effectiveness, LDA does suffer from some limitations such as the SSS problem which leads to the singularity of S_t and thus the failure of classical LDA algorithm. As described previously, many methods [1-6] have been developed to attack this problem. Among others, RDA seems to be the simplest one, which overcomes the singularity problem in virtue of typical Tikhonov regularizer [11]. Its objective function is defined as follows:

$$\max_W \; \frac{tr(W^T S_b W)}{tr(W^T S_t W) + \alpha \cdot tr(W^T W)} \tag{2}$$

where $\alpha > 0$ is a trade-off parameter. As a result, its optimal projection matrix W^* can be easily computed by $(S_t + \alpha I)^{-1} S_b w = \lambda w$, since $S_t + \alpha I$ is nonsingular now.

2.2 Laplacian Linear Discriminant Analysis (LapLDA)

As pointed out in [7], LDA is a completely global DR method, it thereby fails to capture the local structure in data. To handle this problem, Chen et al. developed the LapLDA which aims to capture the global and local structure of the given data simultaneously by integrating LDA with a locality preserving regularizer. Given a set of **labeled** training samples $X = [x_1, x_2, \cdots, x_n] \in R^{m\times n}$ including n data points from *m*-dimensional space, LapLDA seeks an optimal projection matrix $W^* \in R^{m\times d}$ ($d < m$) by the following objective function [7]:

$$\min_W \; \| X^T W - Y \|_F^2 + \alpha \cdot tr(W^T X L X^T W) \tag{3}$$

where Y is a class indicator matrix defined in [12], $\|\cdot\|_F$ denotes Frobenius norm, α is a trade-off parameter, $L = D - S$ is the graph Laplacian whose corresponding adjacency weight matrix $S = (s_{ij})_{n\times n}$ is defined as follows:

$$s_{ij} = \begin{cases} \exp(-\| x_i - x_j \|^2 / 2\sigma^2), & if\ x_i\ is\ among\ kNN\ of\ x_j \\ & or\, if\ x_j\ is\ among\ kNN\ of\ x_i \\ 0 & ,\ otherwise \end{cases} \tag{4}$$

As a result, minimizing $tr(W^T XLX^T W)$ essentially aims to preserve the local geometry in data. Here, we call it locality preserving regularizer.

Note that LapLDA is proposed under a Least Square framework, it can also be easily recast as a trace ratio form under mild condition, according to the proved equivalence between multi-class LDA and multivariate linear regression [12]. For convenience of comparison and without loss of generality, we transform the objective function of LapLDA to the following form:

$$\max_W \frac{tr(W^T S_b W)}{tr(W^T S_t W) + \alpha \cdot tr(W^T XLX^T W)} \tag{5}$$

Similarly, the optimal projection matrix can also be obtained by solving a generalized eigen-equation: $(S_t + \alpha XLX^T)^{-1} S_b w = \lambda w$.

2.3 Semi-supervised Discriminant Analysis (SDA)

SDA shares similar objective function as LapLDA, but it mainly focuses on semi-supervised scenario[1], and simultaneously integrates Fisher criterion, locality preserving regularizer and Tikhonov regularizer. Given a set of **partially labeled** training samples $X = [X_l, X_u]$, where X_l and X_u are the labeled and the unlabeled sample sets respectively, the objective function of SDA is defined as follows:

$$\max_W \frac{tr(W^T S_b W)}{tr(W^T S_t W) + \alpha \cdot tr(W^T XLX^T W) + \beta \cdot tr(W^T W)} \tag{6}$$

where S_b and S_t are calculated using the labeled samples X_l, while the locality preserving regularizer $tr(W^T XLX^T W)$ is calculated using both the labeled and unlabeled samples X. α and β are two trade-off parameters. It is easy to see that if $\alpha = \beta = 0$, SDA becomes the standard LDA; if $\alpha = 0, \beta \neq 0$, it becomes RDA; and if $\alpha \neq 0, \beta = 0$, it becomes the semi-supervised version of LapLDA.

2.4 Globality Preserving Discriminant Analysis (GPDA)

Revisiting the above mentioned three regularization-based LDA methods, we can see that they are formally similar to each other with specific discriminant criteria and data-dependent (or data-independent) regularizers. Motivated by this observation and to reexamine the performance

[1] In contrast, original LapLDA is developed in supervised form, though it can naturally work in both supervised and semi-supervised scenarios.

of these locality preserving LDAs, we purposely design a simple semi-supervised DR algorithm called Globality Preserving Discriminant Analysis (GPDA). Of course, it can naturally work in full supervised scenario.

Instead of preserving the local geometry as imposed in LapLDA and SDA, GPDA aims to just preserve the discriminant information as well as the global structure in data. More specifically, we are given a set of **partially labeled** training samples $X = [X_l, X_u]$ as in SDA. The discriminant information is captured by Fisher criterion based on those labeled samples X_l, while the global structure is reflected by maximizing the variance of both labeled and unlabeled samples X. Without loss of generality, we assume that the samples have been centralized. Then, we define objective function of GPDA as follows:

$$\max_W \frac{tr(W^T S_b W) + \alpha \cdot tr(W^T X X^T W)}{tr(W^T S_t W)} \tag{7}$$

where, S_b and S_t are respectively the between-class scatter matrix and total scatter matrix calculated using labeled samples; XX^T is the sample covariance matrix calculated by both labeled and unlabeled samples, and thus maximizing $tr(W^T XX^T W)$ plays a role in reflecting the global structure in data. We call this term globality preserving regularizer to distinguish it from locality preserving one. It is worthwhile to point out that the locality preserving regularizer corresponds to a ***minimization*** problem essentially motivated by LPP [10], while the globality preserving regularizer corresponds to a ***maximization*** problem essentially motivated by PCA [13].

Naturally, we can add the Tikhonov regularizer $tr(W^T W)$ to the objective function of GPDA to overcome the singularity of S_t for SSS problem, however, we instead prefer to performing GPDA in the PCA transformed subspace, since according to [14], this transform is computationally efficient and does not lose discriminating information.

3 An empirical study on LapLDA and SDA

The experiments in [7] and [8] have verified that the LapLDA and SDA benefit from their used locality preserving regularizers which characterize the specific geometric structure of the data. Surprisingly, the authors of [7] and [8], however, seem to ignore comparing the LapLDA and SDA with simpler methods such as RDA without locality preserving term. In this section, we conduct further experiments and get somewhat counterintuitive results.

3.1 Datasets and Experimental setting

For convenience and impartiality in comparison, we employ the same datasets, experimental setting and parameter selection strategy as in [7] and [8]. Following their schemes, we consider two groups of different experiments. The first group focuses on supervised scenario (subsection 3.2) based on 6 benchmark datasets shown in Table. 1. The same datasets are used to evaluate the effectiveness of LapLDA in [7]. The second group focuses on semi-supervised scenario (subsection 3.3) based on the CMU PIE face database. This dataset has been used to evaluate the effectiveness of SDA in [8]. The original PIE database includes 68 subjects with 41,368 face images as a whole. According to [8], we choose a subset (i.e. Pose C27)[2] with frontal pose and varying illumination, which leaves us 43 images per subject. All the images are cropped to 32x32 pixels, and the gray level values are rescaled to unit interval. Figure 1 shows some sample images from the first subject. Since SDA mainly focuses on semi-supervised learning, 30 images including only 1 labeled and 29 unlabeled are randomly selected from each subject as the training set, the rest as test set. As a result, this essentially brings about a single (labeled) training image face recognition problem [15]. In what follows, we simply call it "single training image face recognition problem" just for keeping consistent with the terminology used in SDA [8].

Table 1. The benchmark datasets and their corresponding partitions used in [7].

Datasets	Training	Test	Sample size	Dim	Class
USPS	750	2250	3000	256	10
20Newsgroups[3]	240	960	1200	8298	4
Waveform	300	900	1200	40	3
Satimage	300	3300	3600	36	6
Letter(a-m)	260	3640	3900	16	13
Soybean	150	412	562	35	15

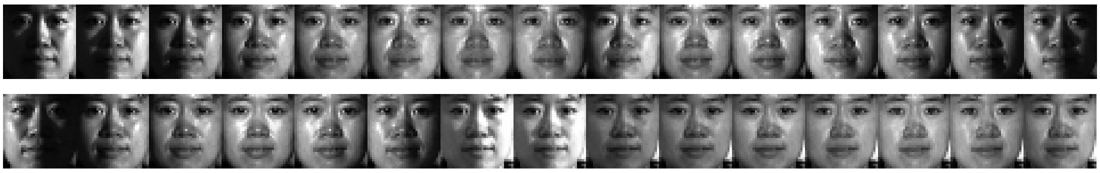

Figure 1. Some face images from the PIE face database.

For the two groups of experiments, the random partition on each data set is repeated 30 times and the classification accuracies (based on 1NN classifier) are averaged as the ultimate performance. One can also refer to [7] and [8] for detailed information about the datasets and experimental setting. We use these datasets mainly for two facts: 1) these datasets cover a wide range of sample sizes and feature dimensions, and consider different application fields such as face recognition, text classification; 2) all the datasets with the same experimental settings are used in [7-8] for evaluating LapLDA and SDA, and thus it is convenient and impartial for comparison.

3.2 Experiments in supervised scenario: LapLDA vs. both RDA and GPDA

Parameter setting. The regularized parameter α in the three methods is determined by 5-fold cross-validation. For original LapLDA, the parameters k and σ in the adjacency weight matrix S are artificially fixed [7]. Here, we also attempt to assign appropriate values for these parameters by 5-fold cross-validation. For 20Newsgroups and Satimage datasets, the total scatter

[2] http://www.cs.uiuc.edu/homes/dengcai2/Data/FaceData.html

[3] For 20Newsgroups dataset, the literature [7] does not give a definite description, we use the first 4 classes and the first 8298 features here.

matrix S_t may be singular due to heavily concentrated eigenvalue distribution (see Figure 2). Therefore, for these two data sets we perform LapLDA and GPDA on the PCA transformed subspace where 98 percent energy of the data is kept.

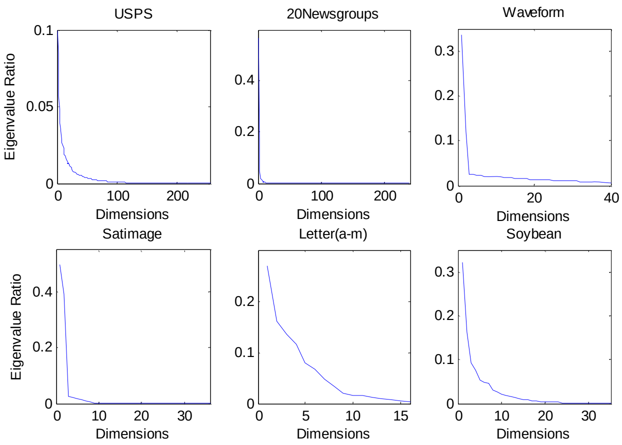

Figure 2. The eigenvalue distributions on 6 benchmark datasets. For 20Newsgroups and Satimage datasets, their eigenvalues mainly concentrate on relatively small feature dimensions. This may incur the singularity of total scatter matrix. Therefore, we perform LapLDA and GPDA on the PCA transformed subspace which keeps 98% energy of the data.

Experiment results. Now we perform original LapLDA (i.e., LapLDA with fixed parameter k and σ), LapLDAcv (i.e., LapLDA with optimal selected k and σ by cross-validation), GPDA and RDA on the 6 benchmark datasets and report their corresponding classification accuracies in Table 2. For LapLDAcv, we can not get the experimental results on USPS and 20Newsgroups datasets due to high computational cost of cross-validation.

Table 2 The classification accuracies related to LapLDA, LapLDAcv, GPDA and RDA. The mean and standard deviation are based on 30 random partitions. (*For 20Newsgourps and Satimage datasets, we perform LapLDA and GPDA on the PCA transformed subspace where 98 percent energy of the training data is kept.)

Datasets	LapLDA[7]		LapLDAcv		GPDA		RDA	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
USPS	84.58	0.84	-	-	82.50	0.71	**90.91**	0.63
20Newsgroups*	80.33	2.17	-	-	82.45	1.35	**84.08**	1.06
Waveform	80.74	1.83	81.25	1.57	74.27	1.95	**81.29**	1.88
Satimage*	83.67	0.79	83.98	1.33	83.72	0.79	**84.34**	0.83
Letter(a-m)	80.40	1.48	**80.96**	1.25	79.82	1.34	80.50	1.45
Soybean	86.89	2.54	87.54	2.33	86.37	2.12	**88.37**	2.18

According to these results, we can get the following observations:

1) The locality preserving LapLDA does not come up to very simple RDA on most of the used benchmark datasets, even though it resorts to much more complex regularizer and parameter selection procedure. Especially for the high-dimensional problems, RDA can achieve more outstanding performance than LapLDA. For example, in classification accuracy on USPS dataset, RDA achieves 90.91%, while LapLDA only 84.58%.

2) On some datasets, locality preserving LapLDA achieves better performance than globality preserving GPDA. This illustrates that local information may be more important than global information for discriminating tasks. However, GPDA can outperform LapLDA on some high-dimensional dataset such as 20Newsgroups. In section 4, we will give detailed discussion on the reasons why locality preserving regularizer may not work so well, especially for high-dimensional problems.

3.3 Experiments in semi-supervised scenario: SDA vs. GPDA

Single training image face recognition problem is one of classical challenges to appearance-based face recognition. Many typical DR methods (e.g. PCA and LDA) will suffer serious performance drop or even fail to work under such scenario [15]. Considering that one may easily gather unlabeled samples sometimes, Cai et al. proposed semi-supervised SDA algorithm for this task, and validated its effectiveness in comparison with many state-of-the-art semi-supervised methods such as LapSVM[16] for single (labeled) training image face recognition problem. Here, we compare SDA with GPDA based on the same problem. It is worthwhile to point out that the within-class scatter S_w is equal to zero if only one available labeled sample in each class, and thus the between-class scatter matrix S_b is exactly equal to the total scatter matrix S_t. As a result, with simple formulation we can get compact objective functions for SDA and GPDA respectively:

$$W^{SDA} = \arg\max_{W} \;\; tr(W^T S_t W) / tr(W^T XLX^T W) \tag{8}$$

$$W^{GPDA} = \arg\max_{W} \;\; tr(W^T XX^T W) / tr(W^T S_t W) \tag{9}$$

Parameter setting. For original SDA, we use the code and parameters provided by Deng Cai [8]. Concretely, the trade-off parameters are set to $\alpha = 0.1, \beta = 0.01$, the neighborhood size k on the graph is set to k =2, the adjacency weight is computed by ***cosine*** distance. To avoid singularity problem, we perform compact SDA and GPDA on the PCA transformed $l-1$ dimensional subspace, where l is the number of the labeled training samples.

In [8], the authors just experiment with fixed 29 unlabeled samples per subject. Here, we attempt different unlabeled sample sizes of 1, 4, 9, 19, 29, since it is generally uneasy to collect abundant unlabeled training samples for single training image face recognition application [15]. The experimental results of original SDA, compact SDA and GPDA are shown in Table 3.

Table 3. Recognition error rates of original SDA, compact SDA and GPDA. The mean and standard deviation is based on 30 random partitions. The results are shown in the form of “mean ± standard deviation% (dimensions)”.

	Recognition error rates with different unlabeled training sample size per subject				
	1	4	9	19	29
SDA[8]	70.2 ± 1.9(67)	65.1 ± 2.3(67)	56.1 ± 2.6(67)	45.0 ± 3.5(67)	40.5 ± 2.7(67)
compact SDA	63.8 ± 2.9(67)	58.2 ± 5.1(67)	46.8 ± 3.6(65)	35.9 ± 3.1(65)	**32.7 ± 2.6(65)**
GPDA	**37.5 ± 3.1(58)**	**36.2 ± 2.7(57)**	**34.1 ± 2.1(53)**	**33.6 ± 2.4(52)**	**32.7 ± 2.3(54)**

From the experimental results, we get the following observations:

1) Globality preserving GPDA tends to achieve better performance than the locality preserving SDA for single (labeled) training image face recognition problem. Incorporating the experimental results on 20Newsgroups dataset, we arrive at a conclusion that the locality preserving regularizer does not necessarily come up to globality preserving one, especially for high-dimensional problems.

2) With rapid decrease of unlabeled samples, the locality preserving SDA will suffer serious performance drop. In contrast, the globality preserving GPDA shows relatively stable performance. For instance, GPDA gets lower recognition error rate even though only one labeled and one unlabeled samples per class are available.

3) In addition, the compact version of SDA can consistently achieve better performance than its original version. This is mainly owing to the use of prior knowledge (i.e., only one labeled sample per class is available) and the previous PCA transformation.

4 Further discussion on locality preserving strategy

According to the previous two groups of experiments, we notice that the locality preserving LDA algorithms, i.e., LapLDA and SDA, may not come up to the simpler RDA and GPDA, especially for high-dimensional problems. Different from many existing works[7-8, 10] which tend to emphasize the advantages of locality preserving strategy, in this section, we instead discuss the reasons why it may not work so well. Moreover, we provide alternative strategies and suggestions to address this problem.

4.1 Why locality preserving criterion may not work well

Naturally, how to characterize the “locality” is at the heart of the locality preserving algorithms. For most of the existing locality preserving methods, this reduces to a graph construction problem[17] which generally relies on the nearest neighbor criterion as in Eq. (4). However, such construction manners may be a hidden trouble incurring the failure of locality preserving strategy, due to that it suffers from several serious problems as follows:

Issue 1: Curse of dimensionality. Generally speaking, the locality preserving LDA algorithms such as LapLDA and SDA are mainly motivated by manifold learning, assuming that the data lie on or near a low-dimensional manifold embedded in the ambient space. However, to characterize the manifold structure, the sample sizes are required to grow exponentially with the intrinsic dimensions of the manifold. As pointed out in [18], this makes the local learning strategy suffer from the so-called curse of dimensionality. For example, recent research shows the face subspace is estimated to have at least 100 dimensions[19]. As a result, many locality based manifold learning techniques perform well on some artificial datasets such as Swiss roll, but do not work well for real-world tasks[20]. This further explains why the locality preserving LapLDA and SDA can not perform well for high-dimensional problem in our above experiments. It is well known that dimensionality reduction is mainly motivated to overcome the “curse of dimensionality”, but unfortunately locality preserving criterion itself suffers from such a curse. This seems to be a paradox.

Issue 2: Sensitivity to noise and outlier. As described before, the "locality" is generally determined by the nearest neighbor criterion for most of the current local DR or semi-supervised learning algorithms [10, 17, 21-22]. However, this leads to that the performance of those methods generally relies heavily on how well the nearest neighbor criterion works in the original high-dimensional space[23]. Since our ultimate goal is classification, we expect that the constructed graph contains as much discriminating information as possible. That is, two data points are linked by an edge if they are likely from the same class. However, based on the nearest neighbor criterion, the graph may link the sample points from different classes, especially in high-dimensional and noisy scenario. For example, the distance between the face images from the same subject may be larger than the distance between the ones from the different subjects due to varying lighting as shown in Figure 3. Obviously, without sufficient training samples, the locality preserving criterion may make the results worse in this case for ultimate classification task. This further illustrates why the SDA algorithm can not even outperform globality preserving GPDA on PIE face database.

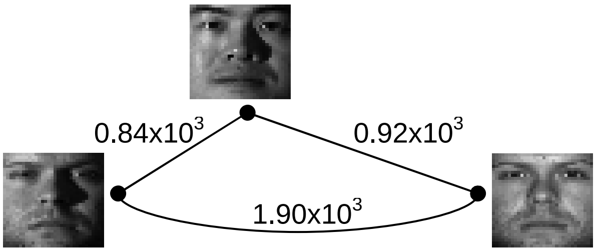

Figure 3. Three face images under varying lighting. The numbers denote the Euclidean distances (wrt. the gray values) among the face images. It is easy to see that the distance between the images from the same person is larger than that from the different persons.

Issue 3: Difficulty of parameter selection. Another problem for the locality preserving strategy is the difficulty of parameter selection involved in adjacency graph construction. It is well known that the graph is characterized by the free parameters k and σ. According to recent research[24], these parameters affect the ultimate classification accuracy significantly. However, assigning proper values for these free parameters is generally uneasy. One way to address this problem is resorting to the cross-validation technique, but it is time-consuming and wasteful of training data. What's worse is that currently there seems no reliable approach for parameter selection if only few labeled samples are available[25].

4.2 Suggestions to address these issues

Suggestion 1: Use of prior knowledge. In pattern recognition and machine learning field, incorporating domain knowledge is seen as an important way to improve the generalization. In what follows, we give two examples based on LapLDA and SDA respectively to show that the use of proper prior knowledge is helpful for improving their performance.

Example (1) LapLDA based on class-specific adjacency graph. For supervised methods, we can naturally construct class-specific graph since labeled information is fully available. That is, we link two data points if they satisfy Eq.(4) and simultaneously belong to the same class. This strategy has been used in many supervised versions of local DR methods [23, 26]. Here, we attempt to perform LapLDA based on such class-specific adjacency graph. Table 4 shows the classification accuracies of original LapLDA and the class-specific LapLDA. Surprisingly, the class-specific graph can improve the performance, especially on high-dimensional datasets, even though the original LapLDA has considered class information by Fisher criterion. This illustrates that the supervised information may be more important than local structure information for classification tasks. In class-specific graph, the supervised information plays a role in pruning the error links among the data points from different classes.

Table 4. The classification accuracies of original LapLDA and class-specific LapLDA. The latter denotes LapLDA with newly constructed graph based on supervised information.

Datasets	Original LapLDA		Class-specific LapLDA	
	Mean	Std.	Mean	Std.
USPS	84.58	0.84	**86.36**	0.72
20Newsgroups	80.33	2.17	**82.49**	1.28
Waveform	80.74	1.83	81.04	1.90
Satimage	83.67	0.79	83.67	0.79
Letter(a-m)	80.40	1.48	80.39	1.48
Soybean	86.89	2.54	86.93	2.55

Example (2) SDA on illumination-insensitive subspace. In unsupervised or semi-supervised scenario, there is no sufficient labeled information available, and thus it is generally uneasy to construct class-specific graph. However, if the domain prior knowledge is reasonably considered, one may still improve the existing algorithms significantly. In fact, our previous experiments in section 3.3 have shown that we can formulate a compact SDA and help to improve the original SDA if single label prior is reasonably used. Here, we take the above mentioned PIE face database again as an example to demonstrate that the performance of SDA can be further improved by employing the illumination prior provided by this database. Revisiting the face images shown in Fig.1 and Fig. 3, one may intuitively find that much of the variation among the face images is due to illumination change which does not generally correspond to important discriminating information. Therefore, we attempt to perform SDA on PCA transformed subspace by discarding several most significant principal components since they generally correspond to variation in lighting. Not only can this mitigate the curse of dimensionality, but also avoid the singularity problem due to insufficient samples. This strategy has been suggested to improve the performance of Eigenface [13], we apply it here under semi-supervised setting. As shown in Figure 3, the minimal error rate of SDA drops from 40.5% to 27.8% on such pruned subspace.

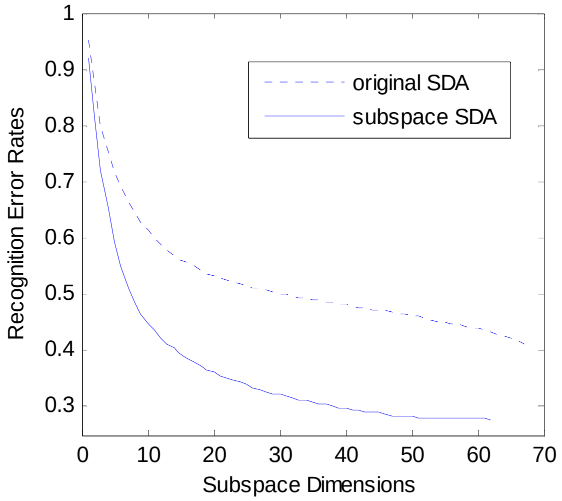

Fig. 3 The recognition error rates of original SDA and its pruned subspace version.

Suggestion 2: Non-local strategy. According to our previous experiments, the proposed globality preserving GPDA may achieve better performance than the locality preserving LDA algorithms sometimes, especially for high-dimensional problems. Therefore, we can consider handling those issues incurred in locality preserving methods in virtue of non-local strategy. Not only can this bypass the curse of dimensionality to a certain extent, but also mitigate the difficulty of parameter selection. In fact, a few years ago, Bengio et al.[27-28] suggested non-local learning as a new research topic. Here, we provide two specific non-local strategies according to our recent studies.

(1) Graph construction based on sparse representation. Different from the traditional adjacency graph which relies on the nearest neighbor criterion, recently we attempt to construct non-local graph based on minimizing a L1 regularization-related objective function [29]. It can potentially link the sample points far from each other, and is validated more discriminative than the nearest neighbor criterion on several publicly available face databases.

(2) Optimizing the projection directions and adjacency weight matrix simultaneously in a unified objective function (to be brief, it means that one minimizes the locality preserving term $tr(W^T XLX^T W)$ wrt. W and L simultaneously). Although this leads to a non-convex optimizing problem, fortunately we can solve it by alternating iterative technique and get suboptimal solutions. As a result, the weight values corresponding to a certain point do not necessarily rely on its neighbors in input space, especially for insufficiently sampling case. We will give detailed discussion about this topic in a forthcoming paper.

5 Conclusions

In this paper, we make an empirical discussion on two recently proposed locality preserving LDA methods, LapLDA and SDA. Through further experiments, we find that these extended LDA algorithms can not generally outperform relatively simple methods such as Regularized Discriminant Analysis (RDA) and the purposely-designed globality preserving discriminant analysis (GPDA). This drives us to reconsider if one should use or how to use the locality preserving strategy in practice. In general, our suggestion is: for the problems which have a few degrees of freedom and thus follow the low-dimensional manifold hypothesis, one should consider locality preserving strategy; however, for the others (e.g., face images especially under uncontrolled condition) which have tens or hundreds of degrees of freedom, the non-local strategy may be a more appropriate option.

It is worthwhile to note that this paper is just an empirically comparative study based on the benchmark datasets recently used to evaluate the effectiveness of LapLDA and SDA. Naturally, an in-depth theoretical analysis of the reasons why these localized methods may not work so well is important and thus imperatively required. This will be our next research goal.

Acknowledgments

This work was partly supported by Natural Science Foundation of China and Jiangsu under grant No. 60973097, 60773061 and BK2008381. Also thank Deng Cai for providing the code and parameter setting of SDA.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, *Eigenfaces vs. Fisherfaces: recognition using class specific linear projection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, **19**(7): 711-720.

[2] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, *Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data.* Applied Statistics, 1995, **44**(1): 101-115.

[3] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, *A new LDA-based face recognition system which can solve the small sample size problem.* Pattern Recognition, 2000, **33**(10): 1713-1726.

[4] J. P. Ye and Q. Li, *A two-stage linear discriminant analysis via QR-decomposition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, **27**(6): 929-941.

[5] J. H. Friedman, *Regularized discriminant analysis.* Journal of the American Statistical Association, 1989, **84**(405): 165-175.

[6] J. P. Ye, R. Janardan, and Q. Li, *Two-dimensional linear discriminant analysis. in* Neural Information Processing Systems (NIPS), 2004.

[7] J. H. Chen, J. P. Ye, and Q. Li, *Integrating global and local structures: A least squares framework for dimensionality reduction. in* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[8] D. Cai, X. F. He, and J. W. Han, *Semi-supervised discriminant analysis. in* IEEE International Conference on Computer Vision (ICCV), 2007.

[9] Y. Q. Song, F. P. Nie, C. S. Zhang, and S. M. Xiang, *A unified framework for semi-supervised dimensionality reduction.* Pattern Recognition, 2008, **41**(9): 2789-2799.

[10] X. F. He and P. Niyogi, *Locality preserving projections. in* Neural Information Processing

Systems (NIPS), 2003.

[11] A. N. Tikhonov and A. V. A., *Solution of Ill-posed Problems*, 1977, Washington: Winston & Sons.

[12] J. P. Ye, *Least squares linear discriminant analysis*. *in* International Conference on Machine Learning (ICML), 2007.

[13] M. Turk and A. Pentland, *Eigenfaces for Recognition.* Journal of Cognitive Neuroscience, 1991, **3**(1): 71-86.

[14] J. Yang and J. Y. Yang, *Why can LDA be performed in PCA transformed space?* Pattern Recognition, 2003, **36**(2): 563-566.

[15] X. Y. Tan, S. C. Chen, Z. H. Zhou, and F. Y. Zhang, *Face recognition from a single image per person: A survey.* Pattern Recognition, 2006, **39**(9): 1725-1745.

[16] M. Belkin, P. Niyogi, and V. Sindhwani, *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.* Journal of Machine Learning Research, 2006, **7**: 2399-2434.

[17] X. Zhu, *Semi-supervised learning literature survey*. Technical Report, 2008.

[18] Y. Bengio, O. Delalleau, and N. L. Roux, *The Curse of Highly Variable Functions for Local Kernel Machines*. *in* Neural Information Processing Systems (NIPS), 2006.

[19] M. Meytlis and L. Sirovich, *On the dimensionality of face space.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, **29**(7): 1262-1267.

[20] L. J. P. Van-der-Maaten, E. O. Postma, and H. J. Van-den-Herik, *Dimensionality Reduction: A Comparative Review*. submit to Journal of Machine Learning Research (http://ict.ewi.tudelft.nl/~lvandermaaten/Publications_files/JMLR_Paper.pdf), 2009-10.

[21] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000, **290**(5500): 2323-2326.

[22] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural Computation, 2003, **15**(6): 1373-1396.

[23] H. T. Chen, H. W. Chang, and T. L. Liu, *Local discriminant embedding and its variants*. *in* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[24] F. Wang and C. S. Zhang, *Label propagation through linear Neighborhoods.* IEEE Transactions on Knowledge and Data Engineering, 2008, **20**(1): 55-67.

[25] D. Y. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, *Learning with local and global consistency*. *in* Neural Information Processing Systems (NIPS), 2004.

[26] D. Xu, S. C. Yan, D. C. Tao, S. Lin, and H. J. Zhang, *Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval.* IEEE Transactions on Image Processing, 2007, **16**(11): 2811-2821.

[27] Y. Bengio, M. Monperrus, and H. Larochelle, *Nonlocal Estimation of Manifold Structure.* Neural Computation, 2006, **18**(10): 2509-2528.

[28] Y. Bengio, O. Delalleau, and N. L. Roux, *Label Propagation and Quadratic Criterion*. *in Semi-Supervised Learning*, 2006, MIT press.

[29] L. Qiao, S. Chen, and X. Tan, *Sparsity preserving projections with applications to face recognition.* Pattern Recogn., 2010, **43**(1): 331-341.