It is the paper published as

Author: J. Gao, P. Kwan and X. Huang

Title: Comprehensive Analysis for the Local Fisher Discriminant Analysis

Journal: International Journal of Pattern Recognition and Artificial Intelligence

ISSN: 0218-0014 1793-6381

Year: 2009

Volume: 23

Issue: 6

Pages: 1129-1143

Abstract: Using data local information, the recently proposed local Fisher Discriminant Analysis (LFDA) algorithm (Sugiyama, 2007) provides a new way of handling the multimodal issues within classes where the conventional Fisher Discriminant Analysis(FDA) algorithm fails. Like the FDA algorithm -€" its global counterpart FDA algorithm, the LFDA suffers when it is applied to the higher dimensional data sets.In this paper we propose a new formulation by which a robust algorithm can beformed. The new algorithm offers more robust results for higher dimensional datasets when compared with the LFDA in most cases. By extensive simulation studies,we have demonstrated the practical usefulness and robustness of our new algorithmin data visualization.

Author Address: xhuang@csu.edu.au

URL: http://dx.doi.org/10.1142/S0218001409007478

http://journals.wspc.com.sg/ijprai/ijprai.shtml

Comprehensive Analysis for the Local Fisher Discriminant Analysis

Junbin Gao[1]*, Paul W. H. Kwan[2] and Xiaodi Huang[3]
[1]*School of Computer Science*
Charles Sturt University, Bathurst, NSW 2795, Australia
jbgao@csu.edu.au
[2]*School of Science and Technology*
University of New England, Armidale, NSW 2351, Australia
kwan@turing.une.edu.au
[3]*School of Business and Information Technology*
Charles Sturt University, Albury, NSW 2640, Australia
xhuang@csu.edu.au

To appear in
International Journal of Pattern Recognition and Artificial Intelligence

Abstract

Using data local information, the recently proposed local Fisher Discriminant Analysis (LFDA) algorithm (Sugiyama, 2007) provides a new way of handling the multimodal issues within classes where the conventional Fisher Discriminant Analysis (FDA) algorithm fails. Like the FDA algorithm — its global counterpart FDA algorithm, the LFDA suffers when it is applied to the higher dimensional data sets. In this paper we propose a new formulation by which a robust algorithm can be formed. The new algorithm offers more robust results for higher dimensional data sets when compared with the LFDA in most cases. By extensive simulation studies, we have demonstrated the practical usefulness and robustness of our new algorithm in data visualization.

1 Introduction

As a vital tool for data exploration, Dimensionality Reduction (DR) is used in areas such as pattern recognition, including face recognition, hand writing recognition, fault detection and classification, and hyperspectral imagery. DR has also been used in other areas

*The author to whom all correspondences should be addressed.

such as robotics (Ham et al., 2005), information retrieval (He and Niyogi, 2004), biometrics (Raytchev et al., 2006; Mekuz et al., 2005), and bioinformatics (Miguel et al., 2002; Okun et al., 2005). Essentially, DR attempts to find a low dimensional representation of a dataset which exists in a high dimensional space. The low dimensional representation is then easily used for comparison, or for classification in other pattern recognition techniques. DR is also applied to select useful features from high dimensional data, see (Uchyigit and Clark, 2007; Caillault and Viard-Gaudin, 2007).

In the last twenty years numerous DR algorithms have been developed (van der Maaten et al., 2007), and research has yielded increasingly sophisticated techniques for the above-mentioned applications. Based on their observations and analysis, Guo et al. (2006) proposed the new twin kernel embedding (TKE) which is a novel type of DR algorithms. In (Guo et al., 2008) they further presented a unified framework for DR, based on their initial work in TKE. We note here that many of the newer algorithms employ kernel machine learning (Schölkopf and Smola, 2002). Many such algorithms have been very successful in different areas.

Two of the most commonly used techniques are Fisher discriminant analysis (FDA) and principal component analysis (PCA). They have been extensively used in pattern recognition. Both PCA and FDA have been applied to different learning tasks like clustering and classification. PCA learns a kind of subspaces where the maximum covariance of all training data is preserved. The eigenfaces method using the PCA technique has been widely used in facial structure analysis (Turk and Pentland, 1991). FDA is one of the most popular dimensionality reduction techniques used in classification formulation (Fisher, 1936; Fukunaga, 1990; Duda et al., 2001). FDA projects high-dimensional data onto a low-dimensional space where the data is reshaped to maximize class separability. The conventional FDA algorithm finds an embedding transformation in such a way that the between-class scatter is maximized while the within-class scatter is minimized or the ratio of the between-class scatter to within-class scatter is maximized. The two most commonly used measures are the trace criterion and the determinant criterion (Duda et al., 2001).

It is well-known that FDA may suffer if samples in a class form several separate clusters. The reason for this is that the mapping from high dimensional space to low dimensional space is actually driven only by the global layout of the data while their local details are overseen by the FDA algorithm. To overcome these drawbacks, Sugiyama (2007) proposed the idea of local FDA that evaluates the between-class scatter and the within-class scatter at a local level. The idea comes out of the analysis of the unsupervised nature of another dimensionality reduction algorithm — Locality Preserving Projection (LLP), see (He and Niyogi, 2004). FDA works well in most scenarios, however it tends to give undesirable results if samples in a class form several separate clusters. To overcome this drawback, the local FDA has been proposed in (Sugiyama, 2007). Almost at the same time Zhao et al. (2007) proposed a similar algorithm.

The formulation of the local FDA uses the inverse of the local within-class scatter matrix which may be singular. Because of this, the algorithm may fail when only a small number of labeled data are available or the data dimension is very high. Both of these cases are quite common in modern applications. Although this can be avoided by applying PCA first,and then the local FDA, some important information may get lost as

a result of applying PCA. This paper solves such a problem by introducing an equivalent formulation to replace the matrix inverse with its pseudo-inverse.

In the next section, we describe the local Fisher Discriminant Analysis (LFDA). In Section 3 a new formulation of local FDA is introduced and a family of solutions is then provided with mathematical proof, followed by a numerical algorithm designed to solve the formulated optimization problem introduced in Section 3. In Section 4, we present the experimental results to evaluate the robust local FDA algorithm we proposed. Finally, we present our conclusions in Section 5.

2 Local Fisher Discriminant Analysis

In this paper we use bold small letters for column vectors (samples or data) and capital letters for a matrix. Let $\mathbf{x}_i^k$ be the i-th datum in the k-th class of K different classes. The number of data vectors in the k-th class is n_k. Let $N = \sum_{k=1}^{K} n_k$ be the total number of data elements and let $\mathbf{X} = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_1}^1, \ldots, \mathbf{x}_1^K, \ldots, \mathbf{x}_{n_K}^K\}$ be the given dataset of N vectors in a high dimensional Euclidean space $\mathbb{R}^d$. Denote the matrix whose columns consist of the data vectors $\mathbf{x}_i^k$ as $\mathbf{X}$. That is, $\mathbf{X} = [\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_1}^1, \ldots, \mathbf{x}_1^K, \ldots, \mathbf{x}_{n_K}^K] \in \mathbf{R}^{d \times N}$. If we don't want to distinguish the class of the data, we simply write $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$.

One of the objectives of DR algorithms is to find a suitable mapping ϕ which maps each high dimensional vector $\mathbf{x}$ to a lower dimensional vector $\mathbf{y} = \phi(\mathbf{x}) \in \mathbb{R}^l$ where $l << d$. In linear algorithms like the FDA the desired mapping is a linear transformation $F \in \mathbb{R}^{d \times l}$ such that $\mathbf{y} = F^T\mathbf{x}$.

To introduce the local FDA, let us define the local within-class scatter matrix S_w and the local between-class scatter matrix S_b as follows,

$$S_w = \frac{1}{2} \sum_{i,j=1}^{N} W_{i,j}^w (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{2.1}$$

$$S_b = \frac{1}{2} \sum_{i,j=1}^{N} W_{i,j}^b (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{2.2}$$

where

$$W_{i,j}^w = \begin{cases} A_{i,j}/n_k & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are same} \\ 0 & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not same} \end{cases} \tag{2.3}$$

$$W_{i,j}^b = \begin{cases} A_{i,j}(1/N - 1/n_k) & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are same} \\ 1/N & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not same} \end{cases} \tag{2.4}$$

In the above definition $A_{i,j}$ is considered as the weight for the sample pair $(\mathbf{x}_i, \mathbf{x}_j)$. The weight has encoded the local relation of the pair. In the actual implementation the weight value is specified according to

$$A_{i,j} = \exp\left\{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_i \sigma_j}\right\} \tag{2.5}$$

where σ_i is the local scaling around $\mathbf{x}_i$ defined by $\sigma_i = \|\mathbf{x}_i - \mathbf{x}_i^{(k)}\|$ where $\mathbf{x}_i^{(k)}$ is the k-th nearest neighbor of $\mathbf{x}_i$, say $k = 7$. The smaller the weight value $A_{i,j}$, the farther away the pair $(\mathbf{x}_i, \mathbf{x}_j)$.

The local FDA transformation matrix $F \in \mathbb{R}^{d\times l}$ is defined as

$$F^* = \underset{F\in\mathbb{R}^{d\times l}}{\operatorname{argmax}} \left[\operatorname{tr}((F^T S_w F)^{-1} F^T S_b F)\right] \tag{2.6}$$

That is, we seek for the "best" mapping F from the d-dimensional data space to a lower l-dimensional projected space ($l \leq d$) such that nearby data pairs in the same class are located close and the data pairs in different classes are separated from each other; far apart data pairs in the same class are not imposed to be close.

The objective function in eq. (2.6) has the same formulation as the objective function used in the conventional FDA. It has been proved, see (Fukunaga, 1990), that the solution F^* can be obtained by solving a generalized eigenvalue problem of S_b and S_w in which the columns of F^* are given by the generalized eigenvectors.

Let S_t be the local total scatter matrix defined by

$$S_t = S_b + S_w.$$

It is easy to show that

$$S_t = \frac{1}{2}\sum_{i,j=1}^{N} W_{i,j}^t (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{2.7}$$

where $W_{i,j}^t$ is given by

$$W_{i,j}^t = \begin{cases} A_{i,j}/N & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are same} \\ 1/N & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not same} \end{cases}$$

3 The Family of Solutions to Local FDA

The conventional FDA extracts only at most $K-1$ meaningful features since its between-class scatter matrix S_b has a rank at most $K-1$ (Fukunaga, 1990). In fact, the local between-class scatter matrix generally has a much higher rank with less eigenvalue multiplicity as pointed out in (Sugiyama, 2007).

3.1 Formulation of the new Local FDA

The algorithm for solving (2.6) assumes the non-singularity of S_w which limits its applications to high-dimensional data. When $d > N$, the matrix S_w would be singular. Although a local FDA algorithm can be preceded by the conventional PCA suggested by (Jin et al., 2001), we should bear in mind that the PCA stage may lose some useful information. Following the idea presented in (Ye and Xiong, 2006), we present the following formulation to replace (2.6) by using pseudo-inverse of the total scatter matrix S_t

$$F^* = \underset{F\in\mathbb{R}^{d\times l}}{\operatorname{argmax}} \left[\operatorname{tr}((F^T S_t F)^{+} F^T S_b F)\right] \tag{3.1}$$

When S_w is non-singular, it can be proved that the above optimal problem is equivalent to the original formulation of the local FDA (2.6).

Lemma 3.1 *The local total scatter matrix S_t and the local within-class scatter matrix S_w are positive semi-definite while the local between-class scatter matrix S_b is also positive semi-definite when $A_{i,j}$ is defined by* (2.5).

Proof: It is obvious that both S_t and S_w are positive semi-definite because weight coefficients $W^t_{i,j}$ and $W^w_{i,j}$ are non-negative.

For the positive semi-definiteness of S_b, let us consider the between-class scatter matrix of the conventional FDA which is defined as

$$S_b^{FDA} = \frac{1}{2}\sum_{i,j=1}^{N} W_{i,j}^{FDA}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{3.2}$$

where

$$W_{i,j}^{FDA} = \begin{cases} 1/N - 1/n_k & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are same} \\ 1/N & \text{if the labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not same} \end{cases}$$

Immediately we can see

$$S_b - S_b^{FDA} = \frac{1}{2}\sum_{k=1}^{K}\sum_{y_i=y_j=k}(A_{i,j}-1)(1/N - 1/n_k)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$$

Note that $A_{i,j} \leq 1$ when $A_{i,j}$s are given by (2.5). Thus the coefficients $(A_{i,j}-1)(1/N - 1/n_k)$ are non-negative because $N \geq n_k$. Hence the matrix $S_b - S_b^{FDA}$ is positive semi-definite. We already know the conventional FDA between-class scatter matrix is positive semi-definite, so is S_b. This completes the proof for the Lemma.

Lemma 3.2 *Let S_t, S_w and S_b be defined as the above, and $t = rank(S_t)$. Then there exists a nonsingular matrix $X \in \mathbb{R}^{d\times d}$, such that*

$$X^T S_w X = D_w = diag(A_t, 0_{d-t}) \tag{3.3}$$

$$X^T S_b X = D_b = diag(B_t, 0_{d-t}) \tag{3.4}$$

where both A_t and B_t are diagonal satisfying $A_t + B_t = I_t$.

Proof: According to Lemma 3.1, both S_b and S_w are positive semi-definite, hence their SVDs have the form of $U_b\Sigma_b U_b^T$ and $U_w\Sigma_w U_w^T$ where both Σ_b and Σ_w are diagonal matrices with nonnegative diagonal elements.

Consider both matrices $\Sigma_b^{1/2}U_b^T$ and $\Sigma_w^{1/2}U_w^T$. By the generalized singular value decomposition (GSVD) (Golub and van Loan, 1996; Paige and Saunders, 1981), there exist orthogonal matrices $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{d\times d}$ and a nonsingular matrix $X \in \mathbb{R}^{d\times d}$ such that

$$U^T\Sigma_b^{1/2}U_b^T X = \text{diag}(\alpha_1, ..., \alpha_t, 0_{d-t}) \tag{3.5}$$

$$V^T\Sigma_w^{1/2}U_w^T X = \text{diag}(\beta_1, ..., \beta_t, 0_{d-t}) \tag{3.6}$$

where $t = \text{rank}(S_t) = \text{rank}(S_b + S_w)$, $1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \alpha_t \geq 0$, $0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_t \leq 1$, and $\alpha_i^2 + \beta_i^2 = 1$ $(i = 1, 2, ..., t)$.

It follows that

$$\begin{aligned} X^T S_w X =& X^T U_w \Sigma_w^{1/2} U U^T \Sigma_w^{1/2} U_w^T X \\ =& \text{diag}(\beta_1^2, ..., \beta_t^2, 0_{d-t}) \end{aligned} \tag{3.7}$$

and similarly

$$X^T S_b X = \text{diag}(\alpha_1^2, ..., \alpha_t^2, 0_{d-t}) \tag{3.8}$$

Particularly

$$X^T S_t X = \text{diag}(\alpha_1^2 + \beta_1^2, ..., \alpha_t^2 + \beta_t^2, 0_{d-t}) = \text{diag}(I_t, 0_{d-t}) \tag{3.9}$$

This completes the proof of the lemma.

Similar to (Ye, 2005), we can find out all the solutions to the optimization problem given by (3.1). The solutions are characterized by a set of nonsingular matrices.

Theorem 3.3 *Let the matrix X be defined in Lemma 3.2 and $b = rank(S_b)$. The solution to the optimization problem defined in* (3.1) *is given by any matrix F in the family $\mathcal{F} = \{F | F = X_b\, M$, where X_b is the matrix consisting of the first b columns of X and M is an arbitrary nonsingular matrix$\}$.*

Proof: By Lemma 3.2, there exists a nonsingular matrix X such that

$$X^T S_t X = D_t = \text{diag}(I_t, 0_{d-t}) \tag{3.10}$$

Consider the objective function defined in (3.1) where F is the variable to be optimized (a $d \times l$ matrix). Let us denote by $\widehat{F} = X^{-1} F$, then

$$F^T S_t F = F^T (X^{-1})^T (X^T S_t X) X^{-1} F = \widehat{F}^T \text{diag}(I_t, 0_{d-t}) \widehat{F}, \tag{3.11}$$

$$F^T S_b F = F^T (X^{-1})^T (X^T S_b X) X^{-1} F = \widehat{F}^T \text{diag}(\alpha_1^2, ..., \alpha_t^2, 0_{d-t}) \widehat{F} \tag{3.12}$$

where $1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_b > \alpha_{b+1} = \cdots = \alpha_t = 0$.

Denote $\Sigma_\alpha = \text{diag}(\alpha_1^2, ..., \alpha_b^2, 0_{t-b})$. Let $\widehat{F} = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}$ with $F_1 \in \mathbb{R}^{t \times l}$ and $F_2 \in \mathbb{R}^{(d-t) \times l}$. The objective function in (3.1) can be represented as

$$\left[\text{tr}((F_1^T F_1)^+ F_1^T \Sigma_\alpha F_1)\right]$$

Since $(F_1^T F_1)^+ = F_1^+ (F_1^+)^T$, then

$$\text{tr}((F_1^T F_1)^+ F_1^T \Sigma_\alpha F_1) = \text{tr}((F_1 F_1^+)^T \Sigma_\alpha (F_1 F_1^+)) \tag{3.13}$$

Denote $\mu = \text{rank}(F_1)$ and write F_1 in its SVD form, $F_1 = U \begin{pmatrix} \Sigma_\mu & 0 \\ 0 & 0 \end{pmatrix} V^T$, where U and V are orthogonal and Σ_μ has nonzero diagonal elements. Immediately we have $F_1^+ = V \begin{pmatrix} \Sigma_\mu^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T$. Hence $F_1 F_1^+ = U \begin{pmatrix} I_\mu & 0 \\ 0 & 0 \end{pmatrix} U^T$. It follows that

$$\begin{aligned} \text{tr}((F_1F_1^+)^T\Sigma_\alpha(F_1F_1^+)) &= \text{tr}\left(U \begin{pmatrix} I_\mu & 0 \\ 0 & 0 \end{pmatrix} U^T\Sigma_\alpha U \begin{pmatrix} I_\mu & 0 \\ 0 & 0 \end{pmatrix} U^T\right) \\ &= \text{tr}\left(\begin{pmatrix} I_\mu & 0 \\ 0 & 0 \end{pmatrix} U^T\Sigma_\alpha U \begin{pmatrix} I_\mu & 0 \\ 0 & 0 \end{pmatrix}\right) \\ &= \text{tr}(U_\mu^T\Sigma_\alpha U_\mu) \le \alpha_1^2 + \cdots + \alpha_\mu^2. \end{aligned} \tag{3.14}$$

where U_μ is the matrix consisting of the first μ columns of U. That is, U_μ is column orthogonal. If we let U_μ take a particular form of $U_\mu = \begin{pmatrix} W \\ 0 \end{pmatrix}$ where $W \in \mathbb{R}^{\mu\times\mu}$ with $\mu = l = b$ is an orthogonal matrix, then we can see

$$\text{tr}(U_\mu^T\Sigma_\alpha U_\mu) = \alpha_1^2 + \cdots + \alpha_\mu^2. \tag{3.15}$$

For the special choice of U_μ as above, we have

$$F_1 = U_\mu\Sigma_b V^T = \begin{pmatrix} W\Sigma_b V^T \\ 0 \end{pmatrix}. \tag{3.16}$$

where Σ_b is the first $b \times b$ subblock of Σ_α. We also note that the value of the objective function in (3.1) is independent of F_2. Thus we may write the optimal solution of (3.1) as

$$\widehat{F} = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} W\Sigma_b V^T \\ 0 \end{pmatrix}. \tag{3.17}$$

Since the orthogonal matrices W, V, and the diagonal matrix Σ_b are arbitrary. Hence $M = W\Sigma_b V^T$ is an arbitrary nonsingular matrix. It follows that $F = X\widehat{F} = X_b M$ for any nonsingular M maximizes the objective function. This completes the proof of the theorem.

In summary, the theorem has characterized all the solutions to the new local FDA (nLFDA) problem defined by (3.1). Thus one can choose an appropriate solution from $F = X\widehat{F} = X_b M$ by specifying a special non-singular matrix M according to different requirements in applications. Particularly when $M = I_b$ the resulting mapping $F_b = X_b$ satisfies $F_b^T S_t F_b = I_b$; in other words, the components given by the algorithm are S_t-orthogonal. Let us consider X_b's QR decomposition $X_b = QR$ and take $M = R^{-1}$ where Q is a matrix with orthogonal columns, then the resulting mapping $F = Q$. Thus the components defined in this way are orthogonal.

3.2 Algorithm Design for the nLFDA

The new local FDA algorithm is very similar to the original local FDA algorithm. The local FDA is solved based on a generalized eigenvalue problem. The software can be

found at `http://sugiyama-www.cs.titech.ac.jp/~sugi/Software/LFDA/` in which an efficient algorithm is provided to compute local within-class scatter and between-class scatter. Our algorithm design makes use of the way of computing those scatters and then employs the MATLAB's function of generalized singular value decomposition `gsvd` to compute the nonsingular matrix X in Lemma 3.2. With X it is easy to make the transform F as suggested by Theorem 3.3.

The main steps are to

1. Input the labeled samples $\mathbf{x}_i^k$ where $i = 1, ..., n_k$ and $k = 1, ..., K$ and the targeted dimension l;
2. Calculate the local within-class scatter S_w and the between-class scatter S_b according to the algorithm in (Sugiyama, 2007);
3. Compute SVDs $U_w \Lambda_w V_w' = S_w$ and $U_b \Lambda_b V_b' = S_b$
4. Find X by applying the matlab function `gsvd` to $\Lambda_w^{1/2} V_w'$ and $\Lambda_b^{1/2} V_b'$
5. Construct F according to Theorem 3.3, e.g. by taking $M = I$;
6. Output the projected data by using transform F^T.

Generally speaking both the LFDA and nLFDA would have similar performance in general cases, however the nLFDA is more robust than the LFDA because the algorithm for the nLFDA makes use of generalized SVD to automatically handle the singularity of the within-class scatter matrix which may fail the LFDA algorithm.

4 Experiment Evaluation

4.1 Synthetic Data

We first demonstrate that the performance of the new local FDA (nLFDA) algorithm is comparable to Sugiyama's local FDA (LFDA) algorithm (Sugiyama, 2007) by using a synthetic data set. Dimensionality reduction results obtained by nLFDA, LFDA, LDA (conventional FDA), LPP and PCA are illustrated in Figure 1 in which two class data samples in the 2-dimensional space are projected onto a one dimensional space determined by those algorithms, respectively. Both nLFDA and LFDA algorithms nicely separate the samples in difference classes in the projected one dimensional space with only a slight difference. Actually both algorithms perform similarly in most scenarios. Similarly both LPP and PCA discover the similar separation direction as the first principal direction due to their unsupervised nature. However the conventional FDA fails on this data set due to the multimodality within the same class.

4.2 Data Visualization

In the following experiments we consider the performance of both nLFDA and LFDA, and other existing algorithms in visualizing high dimensional data sets.

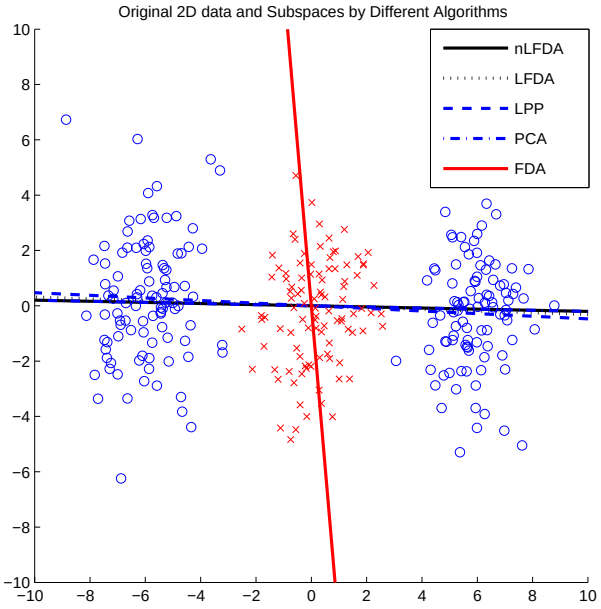

Figure 1: Examples of dimensionality reduction by nLFDA, LFDA, LDA (conventional FDA), LPP and PCA. Two class samples in 2-dimensional space are projected onto a one dimensional spaces by all the algorithms. The line in the figure denotes the one dimensional discriminant space obtained by each method

The experiments are carried out on the *Letter recognition* data set and *Iris* data set which are available from the UCI machine learning repository (Asuncion and Newman, 2007). The *Letter recognition* data set has sample dimension of 16 and 26 categories, i.e., 'A' to 'Z'. The *Iris* data set contains 149 samples whose dimension is 4. The samples belong to three classes: Setosa, Virginica and Versicolour.

We test nLFDA, LFDA, LDA (conventional FDA) and PCA. Figure 2 shows the samples of the *Iris* data set in the two-dimensional space found by each algorithm. The horizontal axis is the first feature found by each algorithm, while the vertical axis is the second feature. For this data set all the algorithms have nicely revealed the categories of the data.

In the second experiment, within four classes we select 789 samples from 'A' class, 766 samples from 'B' class, 736 samples from 'C' and 775 samples from 'F' class. We note that the first two components found by both FDA and PCA algorithms do not separate categorized samples well. In Figure 3 we therefore depict the samples of the *Letter recognition* data set in the three-dimensional space found by each algorithm. In this experiment, the conventional FDA algorithm finds only two discriminant components, hence we use a random direction as the third component in the picture. The view used in each figure was found by rotating the space until a satisfactory visual effect was achieved. From the figures it is clearly seen that majority of samples from classes 'B' and 'F' are mixed in the result revealed by the FDA algorithm while only the samples from class 'A' are clearly separated from other class samples which are overlapped. Both nLFDA and LFDA revealed nice class structures in four classes although some samples from different

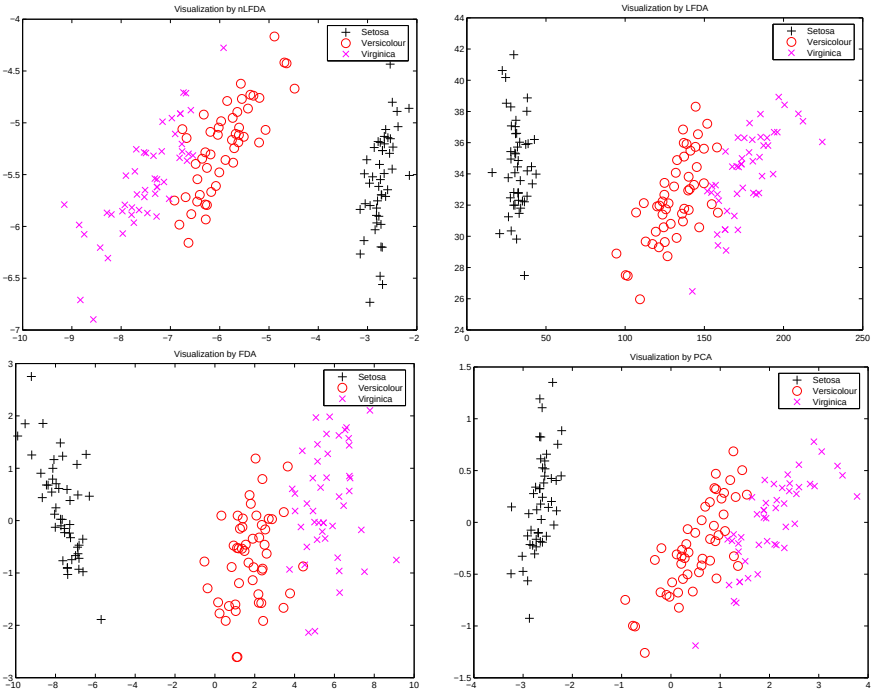

Figure 2: The result of visualizing the *Iris* data set

classes are overlapped.

To evaluate the performance of the new algorithm on visualizing higher dimensional data sets we conduct another experiment on the facial image data. The data set used in this experiment is the Olivetti's facial images which can be obtained from `http://www.cs.toronto.edu/~roweis/data.html`. The data set has 400 gray scale facial images of 40 people, and 10 images for each person. Each image consists of $4096 = 64 \times 64$ pixels and each pixel takes an integer intensity value between 0 and 255. In this experiment the image is downsampled to 32×32 pixels so that the experiment can be conducted on an ordinary desktop machine without memory problems. The pixel values are scaled to between 0 and 1. The downsampling results in a data set of dimension 1024.

In this experiment, we decompose the images into two classes according to whether the subject wears glasses or not. Our task is to project the facial images onto a two dimensional space so that the subjects with and without glasses are separated from each other. As the dimension (1024) of image data is larger than the size of the data set (400), both FDA and LFDA algorithms fail for this data set. The embedded results produced by nLFDA and PCA are shown in Figure 4, where circle/plus symbols indicate faces with/without glasses. The figure shows that nLFDA perfectly separates the labeled samples while PCA offers no clues about the labeled data.

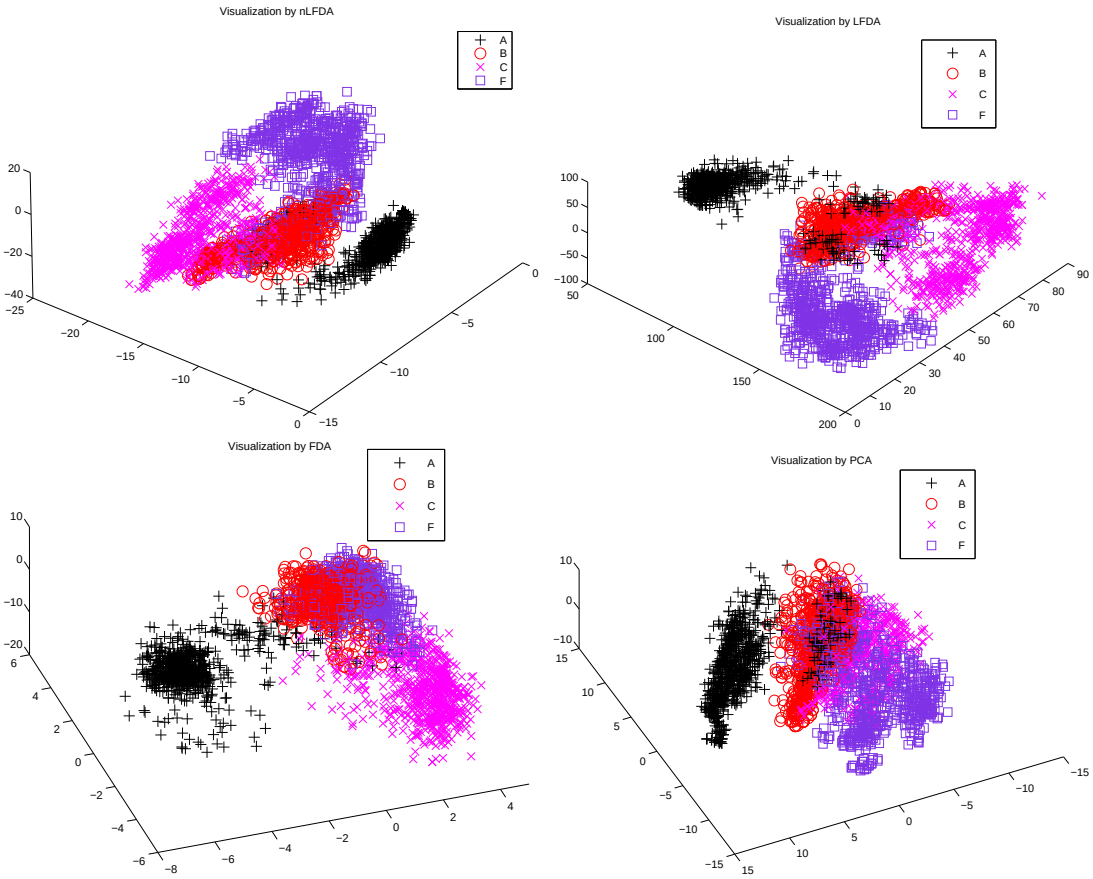

Figure 3: The result of visualizing the *Letter recognition* data set

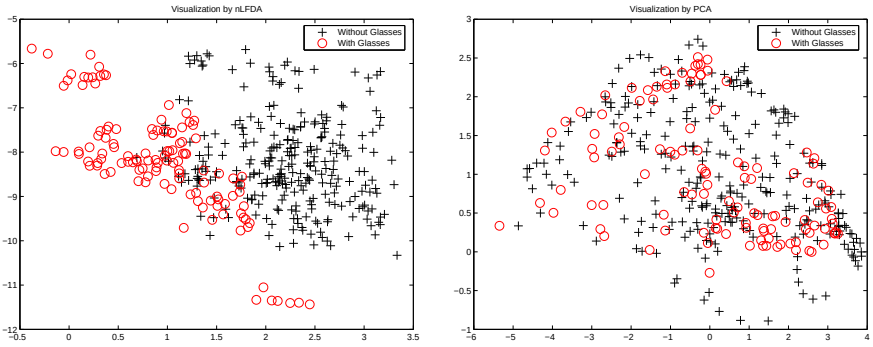

Figure 4: The result of visualizing Olivetti's facial image data set

Data Sets	nLFDA	LFDA	FDA	PCA
Iris	3.17% (0.0253)	3.17% (0.0253)	4.67% (0.0274)	3.67% (0.0304)
Letters	13.1% (0.0161)	13.0% (0.0161)	16.6% (0.0112)	32.8% (0.0167)
USPS Digits	7.74% (0.0032)	7.74% (0.0032)	7.24% (0.0041)	7.73% (0.0025)
Olivetti's faces	16.4% (0.0497)	N/A	N/A	24.4% (0.0296)

Table 1: Comparison on classification accuracy between these algorithms using 1-nearest neighbor criterion (the standard deviation appeared in parenthesis)

4.3 Classification

Here we apply the proposed new algorithm nLFDA, LFDA, LDA (conventional FDA) and PCA to classification tasks.

There are several measures for quantitatively evaluating separability of projected data samples in different classes (Fukunaga, 1990). Here we use a simple one: misclassification rate by a 1-nearest-neighbor criterion. We know the one-nearest-neighbor metric depends on the distance of projected data samples. To get consistent results, in our experiments we use an orthogonal transform F as defined by Theorem 3.3, i.e., we take F as the orthogonal part of QR decomposition of X.

We use the same data sets in Section 4.2. In addition, we use the ten classes data set created from the USPS handwritten digit data set. For the data sets IRIS, the embedding dimensionality is chosen as 2, while for other three data sets, the embedding dimensionality is set to 9 in our experiments. Table 1 describes the mean and standard deviation of the misclassification rate by each method on 10 fold experiments where the data sets are split into training and testing sets at a ratio of 4 to 1. The table shows that the overall performance of nLFDA is comparable to that of the LFDA, the latter of which has excellent performance. The LFDA as well as FDA failed in the Olivetti's facial image data. In this case, we only have 400 data points while the dimensional of each face image is 1024. This results in a singular between-class scatter matrix. However the nLFDA is designed to tackle this undersample problem. The performance of the nLFDA on this data set is much better than that of PCA.

5 Conclusions

In this paper we modify the formulation of the local Fisher Discriminant Analysis (LFDA) (Sugiyama, 2007) so that the new formulation is more robust for high dimensional data sets. When the data dimension is larger than the number of the size of a data set, the LFDA will fail due to the singularity of the within-class scatter matrix. Instead of using the matrix inverse in the formulation the pseudo-inverse is adopted in the new formulation. The new algorithm has a comparable capacity as the LFDA in most cases while it offers a robust implementation for higher dimensional data sets. A family of solutions to the new formulation has been derived based on a comprehensive analysis. The algorithm based on this analysis is robust in all the experiments.

In this paper we focused on the linear dimensionality reduction of the new algorithm and its application to data visualization. Obviously the new algorithm can be combined with other classifiers such as the 1-nearest-neighbor classifier for classification problems.

Acknowledgements

The authors are grateful to anonymous reviewers for their constructive advices.

References

Asuncion, A. and D. Newman (2007). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Caillault, E. and C. Viard-Gaudin (2007). Mixed discriminant training of hybrid ann/hmm systems for online handwritten word recognition. *International Journal of Pattern Recognition and Artificial Intelligence 21*, 117–134.

Duda, R., P. Hart, and D. Stork (2001). *Pattern Classification* (2nd ed.). New York: John Wiley and Sons.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (2nd ed.). Boston: Academic Press.

Golub, G. and C. van Loan (1996). *Matrix Computations* (3 ed.). Maryland: The Johns Hopkins University Press.

Guo, Y., J. B. Gao, and P. W. Kwan (2006). Visualization of non-vectorial data using twin kernel embedding. In K. Ong, K. Smith-Miles, V. Lee, and W. Ng (Eds.), *Proceedings of the International Workshop on Integrating AI and Data Mining (AIDM 2006) in Australia*, pp. pp11–17. IEEE Computer Society.

Guo, Y., J. B. Gao, and P. W. Kwan (2008). Twin kernel embedding. *IEEE Transaction on Pattern Analysis and Machine Intelligence 30*, 1490–1495.

Ham, J., Y. Lin, and D. Lee (2005). Learning non-linear appearance manifolds for robot localization. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2971–2976.

He, X. and P. Niyogi (2004). Locality preserving projections. In S. Thrun, L. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems*, Volume 16, Cambridge, MA. MIT Press.

Jin, Z., J. Yang, Z. Hu, and Z. Lou (2001). Face recognition based on the uncorrelated discriminant transformation. *Pattern Recognition 34*, 1405–1416.

Mekuz, N., C. Bauckhage, and J. Tsotsos (2005). Face recognition with weighted locally linear embedding. In *Proceedings of the 2nd Canadian Conference on Computer and Robot Vision*, pp. 290–296.

Miguel, L., G. Phillips, and L. Kavraki (2002). A dimensionality reduction approach to modeling protein flexibility. In *Proceedings of International Conference on Computational Molecular Biology*, pp. 299–308.

Okun, O., H. Priisalu, and A. Alves (2005). Fast non-negative dimensionality reduction for protein fold recognition. In *Proceedings of European Conference on Machine Learning*, 665-672.

Paige, C. and M. Saunders (1981). Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis 18*, 398–405.

Raytchev, B., I. Yoda, and K. Sakaue (2006). Multiview face recognition by nonlinear dimensionality reduction and generalized linear models. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, Washington, DC, USA, pp. 625–630.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. Cambridge, Massachusetts: The MIT Press.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research 8*, 1027–1061.

Turk, M. and A. Pentland (1991). Eigenfaces for recognition. *Journal of Cognitive Neuro Science 3*(1), 71–86.

Uchyigit, G. and K. Clark (2007). A new feature selection method for text classification. *International Journal of Pattern Recognition and Artificial Intelligence 21*, 423–438.

van der Maaten, L., E. O. Postma, and H. van den Herick (2007). Dimensionality reduction: A comparative review. http://www.cs.unimaas.nl/l.vandermaaten/dr/DR_draft.pdf.

Ye, J. (2005). Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research 6*, 483–502.

Ye, J. and T. Xiong (2006). Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research 7*, 1183–1204.

Zhao, D., Z. Lin, R. Xiao, and X. Tang (2007). Linear laplacian discrimination for feature extraction. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'07)*.