Journal of Machine Learning Research 8 (2007) 1027-1061 Submitted 3/06; Revised 12/06; Published 5/07

Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis[*]

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

Editor: Sam Roweis

Abstract

Reducing the dimensionality of data without losing intrinsic information is an important preprocessing step in high-dimensional data analysis. Fisher discriminant analysis (FDA) is a traditional technique for supervised dimensionality reduction, but it tends to give undesired results if samples in a class are *multimodal*. An unsupervised dimensionality reduction method called locality-preserving projection (LPP) can work well with multimodal data due to its locality preserving property. However, since LPP does not take the label information into account, it is not necessarily useful in supervised learning scenarios. In this paper, we propose a new linear supervised dimensionality reduction method called *local Fisher discriminant analysis* (LFDA), which effectively combines the ideas of FDA and LPP. LFDA has an analytic form of the embedding transformation and the solution can be easily computed just by solving a generalized eigenvalue problem. We demonstrate the practical usefulness and high scalability of the LFDA method in data visualization and classification tasks through extensive simulation studies. We also show that LFDA can be extended to non-linear dimensionality reduction scenarios by applying the kernel trick.

Keywords: dimensionality reduction, supervised learning, Fisher discriminant analysis, locality preserving projection, affinity matrix

1. Introduction

The goal of dimensionality reduction is to embed high-dimensional data samples in a low-dimensional space so that most of 'intrinsic information' contained in the data is preserved (e.g., Roweis and Saul, 2000; Tenenbaum et al., 2000; Hinton and Salakhutdinov, 2006). Once dimensionality reduction is carried out appropriately, the compact representation of the data can be used for various succeeding tasks such as visualization, classification, etc. In this paper, we consider the *supervised* dimensionality reduction problem, that is, samples are accompanied with class labels.

Fisher discriminant analysis (FDA) (Fisher, 1936; Fukunaga, 1990) is a popular method for linear supervised dimensionality reduction.[1] FDA seeks for an embedding transformation such

[*] An efficient MATLAB implementation of local Fisher discriminant analysis is available from the author's website: 'http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA/'.

[1] FDA may refer to the classification method which first projects data samples onto a one-dimensional subspace and then classifies the samples by thresholding (Fisher, 1936; Duda et al., 2001). The one-dimensional embedding space used here is obtained as the maximizer of the so-called *Fisher criterion*. This Fisher criterion can be used for dimensionality reduction onto a space with dimension more than one in multi-class problems (Fukunaga, 1990). With some abuse, we refer to the dimensionality reduction method based on the Fisher criterion as FDA (see Section 2.2 for detail).

that the between-class scatter is maximized and the within-class scatter is minimized. FDA is a traditional but useful method for dimensionality reduction. However, it tends to give undesired results if samples in a class form several separate clusters (i.e., *multimodal*) (see, e.g., Fukunaga, 1990).

Within-class multimodality can be observed in many practical applications. For example, in disease diagnosis, the distribution of medial checkup samples of sick patients could be multimodal since there may be several different causes even for a single disease. In a traditional task of hand-written digit recognition, within-class multimodality appears if digits are classified into, for example, even and odd numbers. More generally, solving multi-class classification problems by a set of two-class 'one-versus-rest' problems naturally induces within-class multimodality. For this reason, there is a universal need for reducing the dimensionality of multimodal data.

In order to reduce the dimensionality of multimodal data appropriately, it is important to preserve the local structure of the data. *Locality-preserving projection* (LPP) (He and Niyogi, 2004) meets this requirement; LPP seeks for an embedding transformation such that nearby data pairs in the original space close in the embedding space. Thus LPP can reduce the dimensionality of multimodal data without losing the local structure. However, LPP is an unsupervised dimensionality reduction method and does not take the label information into account. Therefore, it does not necessarily work appropriately in supervised dimensionality reduction scenarios.

In this paper, we propose a new dimensionality reduction method called *local Fisher discriminant analysis* (LFDA). LFDA effectively combines the ideas of FDA and LPP, that is, LFDA maximizes between-class separability and preserves *within-class local structure* at the same time. Thus LFDA is useful for dimensionality reduction of multimodal labeled data.

The original FDA provides a meaningful result only when the dimensionality of the embedding space is smaller than the number of classes because of the rank deficiency of the between-class scatter matrix (Fukunaga, 1990). This is an essential limitation of FDA in dimensionality reduction. On the other hand, the proposed LFDA does not generally suffer from this problem and can be employed for dimensionality reduction into an *arbitrary* dimensional space. Furthermore, LFDA inherits an excellent property from FDA—it has an *analytic* form of the embedding matrix and the solution can be easily computed just by solving a generalized eigenvalue problem. This is an advantage over recently proposed supervised dimensionality reduction methods (e.g., Goldberger et al., 2005; Globerson and Roweis, 2006). Furthermore, LFDA can be naturally extended to non-linear dimensionality reduction scenarios by applying the *kernel trick* (Schölkopf and Smola, 2002).

The rest of this paper is organized as follows. In Section 2, we formulate the linear dimensionality reduction problem, briefly review FDA and LPP, and illustrate how they typically behave. In Section 3, we define LFDA and show its fundamental properties. In Section 4, we discuss the relation between LFDA and other methods. In Section 5, we numerically evaluate the performance of LFDA and existing methods in visualization and classification tasks using benchmark data sets. Finally, we give concluding remarks and future prospects in Section 6.

2. Linear Dimensionality Reduction

In this section, we formulate the problem of linear dimensionality reduction and review existing methods.

2.1 Formulation

Let $x_i \in \mathbb{R}^d$ $(i = 1, 2, \ldots, n)$ be d-dimensional samples and $y_i \in \{1, 2, \ldots, c\}$ be associated class labels, where n is the number of samples and c is the number of classes. Let n_ℓ be the number of samples in class ℓ:

$$\sum_{\ell=1}^{c} n_\ell = n.$$

Let X be the matrix of all samples:

$$X \equiv (x_1|x_2|\cdots|x_n).$$

Let $z_i \in \mathbb{R}^r$ $(1 \leq r \leq d)$ be low-dimensional representations of x_i, where r is the reduced dimension (i.e., the dimension of the embedding space). Effectively we consider d to be large and r to be small, but not limited to such cases.

For the moment, we focus on linear dimensionality reduction, that is, using a $d \times r$ transformation matrix T, the embedded samples z_i are given by

$$z_i = T^\top x_i,$$

where $^\top$ denotes the transpose of a matrix or vector. In Section 3.4, we extend our discussion to the non-linear dimensionality reduction scenarios where the mapping from x_i to z_i is non-linear.

2.2 Fisher Discriminant Analysis for Dimensionality Reduction

One of the most popular dimensionality reduction techniques is *Fisher discriminant analysis* (FDA) (Fisher, 1936; Fukunaga, 1990; Duda et al., 2001). Here we briefly describe the definition of FDA.

Let $S^{(w)}$ and $S^{(b)}$ be the *within-class scatter matrix* and the *between-class scatter matrix*:

$$S^{(w)} \equiv \sum_{\ell=1}^{c} \sum_{i:y_i=\ell} (x_i - \mu_\ell)(x_i - \mu_\ell)^\top, \tag{1}$$

$$S^{(b)} \equiv \sum_{\ell=1}^{c} n_\ell (\mu_\ell - \mu)(\mu_\ell - \mu)^\top, \tag{2}$$

where $\sum_{i:y_i=\ell}$ denotes the summation over i such that $y_i = \ell$, μ_ℓ is the mean of the samples in class ℓ, and μ is the mean of all samples:

$$\mu_\ell \equiv \frac{1}{n_\ell} \sum_{i:y_i=\ell} x_i,$$

$$\mu \equiv \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{\ell=1}^{c} n_\ell \mu_\ell.$$

We assume that $S^{(w)}$ has full rank. The FDA transformation matrix T_{FDA} is defined as follows:[2]

$$T_{FDA} \equiv \underset{T \in \mathbb{R}^{d \times r}}{\operatorname{argmax}} \left[\mathrm{tr}\left((T^\top S^{(w)} T)^{-1} T^\top S^{(b)} T \right) \right]. \tag{3}$$

[2] The following definition is also used in the literature (e.g., Fukunaga, 1990) and yields the same solution.

$$T_{FDA} = \underset{T \in \mathbb{R}^{d \times r}}{\operatorname{argmax}} \left[\frac{\det\left(T^\top S^{(b)} T\right)}{\det\left(T^\top S^{(w)} T\right)} \right],$$

where $\det(\cdot)$ denotes the determinant of a matrix.

That is, FDA seeks a transformation matrix T such that the between-class scatter is 'maximized' while the within-class scatter is 'minimized'. In the above formulation, we implicitly assumed that $T^\top S^{(w)} T$ is invertible. This implies that the above optimization is subject to

$$\mathrm{rank}\,(T) = r.$$

Let $\{\varphi_k\}_{k=1}^{d}$ be the generalized eigenvectors associated with the generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of the following generalized eigenvalue problem:

$$S^{(b)}\varphi = \lambda S^{(w)}\varphi.$$

Then a solution T_{FDA} of the above maximization problem is analytically given by

$$T_{FDA} = (\varphi_1|\varphi_2|\cdots|\varphi_r).$$

Note that the solution is not unique and the following simple constraint is sometimes imposed additionally (Fukunaga, 1990).

$$T_{FDA}^\top S^{(w)} T_{FDA} = I_r,$$

where I_r is the identity matrix on $\mathbb{R}^r$. This constraint makes the within-class scatter in the embedding space *sphered*.

The between-class scatter matrix $S^{(b)}$ has at most rank $c-1$ (Fukunaga, 1990). This implies that the multiplicity of $\lambda = 0$ is at least $d-c+1$. Therefore, FDA can find at most $c-1$ meaningful features; the remaining features found by FDA are arbitrary. This is an essential limitation of FDA for dimensionality reduction and is very restrictive in practice.

2.3 Locality-Preserving Projection

Another dimensionality reduction technique that is relevant to the current setting is *locality-preserving projection* (LPP) (He and Niyogi, 2004). Here we review LPP.

Let A be an *affinity matrix*, that is, the n-dimensional matrix with the (i,j)-th element $A_{i,j}$ being the affinity between x_i and x_j. We assume that $A_{i,j} \in [0,1]$; $A_{i,j}$ is large if x_i and x_j are 'close' and $A_{i,j}$ is small if x_i and x_j are 'far apart'. There are several different manners of defining A. We briefly describe typical definitions in Appendix D. The LPP transformation matrix T_{LPP} is defined as follows:[3]

$$T_{LPP} \equiv \underset{T\in\mathbb{R}^{d\times r}}{\mathrm{argmin}} \left(\frac{1}{2} \sum_{i,j=1}^{n} A_{i,j} \|T^\top x_i - T^\top x_j\|^2 \right)$$
$$\text{subject to } T^\top X D X^\top T = I_r, \tag{4}$$

where D is the n-dimensional diagonal matrix with i-th diagonal element being

$$D_{i,i} \equiv \sum_{j=1}^{n} A_{i,j}.$$

[3] The matrix D in the constraint (4) is motivated by a geometric argument (Belkin and Niyogi, 2003). However, it is sometimes dropped for the sake of simplicity (Ham et al., 2004).

Eq. (4) implies that LPP looks for a transformation matrix T such that *nearby* data pairs in the original space $\mathbb{R}^d$ are kept close in the embedding space. The constraint (4) is imposed for avoiding degeneracy.

Let $\{\psi_k\}_{k=1}^{d}$ be the generalized eigenvectors associated with the generalized eigenvalues $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d$ of the following generalized eigenvalue problem:

$$XLX^\top\psi = \gamma XDX^\top\psi,$$

where

$$L \equiv D - A.$$

L is called the *graph-Laplacian matrix* in the spectral graph theory (Chung, 1997), where A is seen as the *adjacency matrix* of a graph. He and Niyogi (2004) showed that a solution of Eq. (4) is given by

$$T_{LPP} = (\psi_d|\psi_{d-1}|\cdots|\psi_{d-r+1}).$$

2.4 Typical Behavior of FDA and LPP

Dimensionality reduction results obtained by FDA and LPP are illustrated in Figure 1 (LFDA will be defined and explained in Section 3)—two-dimensional two-class data samples are embedded into a one-dimensional space. In LPP, the affinity matrix A is determined by the *local scaling method* (Zelnik-Manor and Perona, 2005, see also Appendix D.4).

For the simplest data set depicted in Figure 1(a), both FDA and LPP nicely separate the samples in different classes (‘$\circ$’ and ‘$\times$’) from each other. For the data set depicted in Figure 1(b), FDA still works well, but LPP mixes samples in different classes into a single cluster. This is caused by the unsupervised nature of LPP. On the other hand, for the data set depicted in Figure 1(c), LPP works well but FDA collapses the samples in different classes into a single cluster. The reason for the failure of FDA is that the ‘levels’ of the between-class scatter and the within-class scatter are not evaluated in an intuitively natural way because of the two separate clusters in ‘$\circ$’-class (see also Fukunaga, 1990).

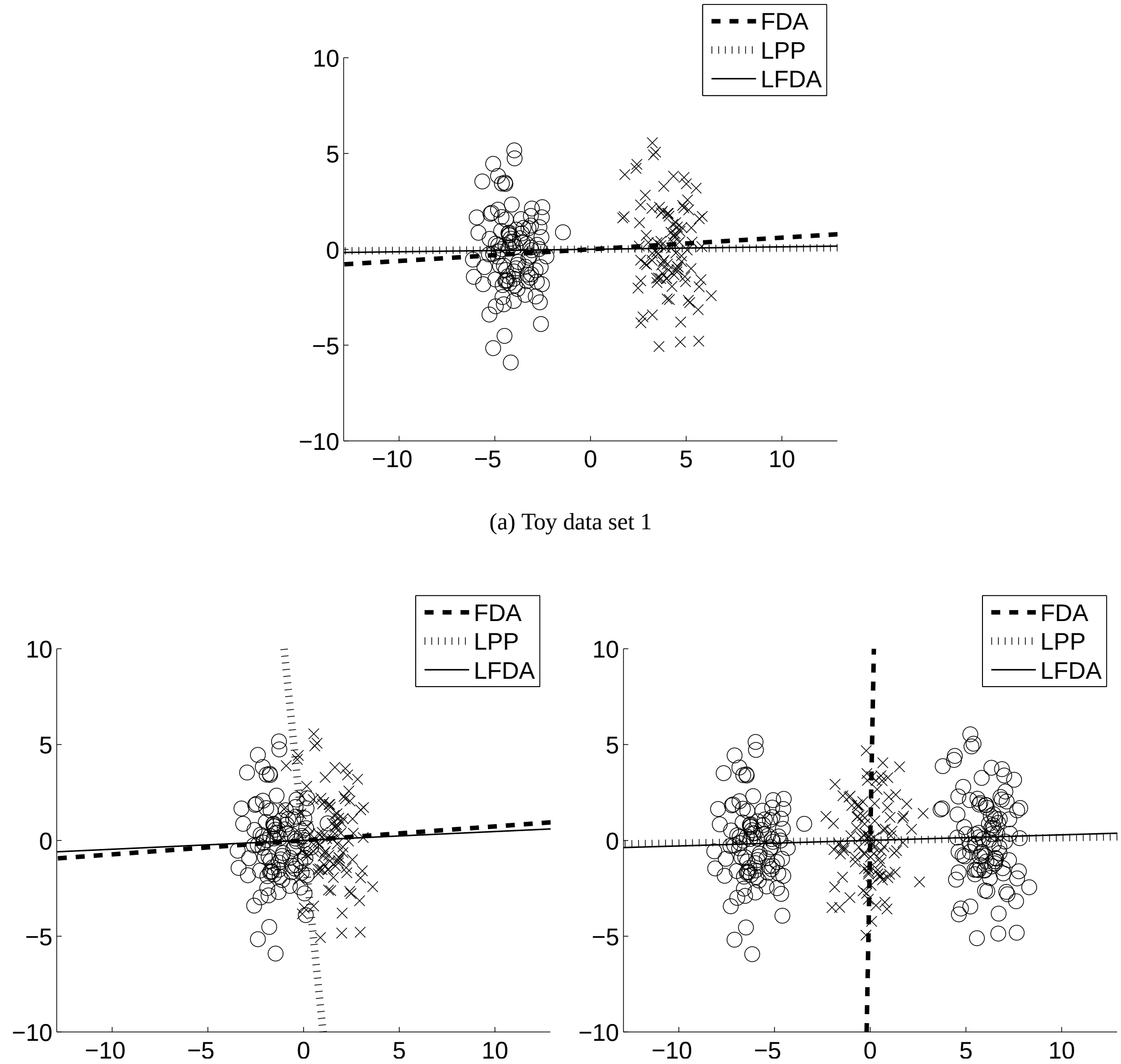
3. Local Fisher Discriminant Analysis

As illustrated in Figure 1, FDA can perform poorly if samples in a class form several separate clusters (i.e., *multimodal*). In other words, the undesired behavior of FDA is caused by the *globality* when evaluating the within-class scatter and the between-class scatter (e.g., Figure 1(c)). On the other hand, because of the unsupervised nature of LPP, it can overlap samples in different classes if they are close in the original high-dimensional space $\mathbb{R}^d$ (e.g., Figure 1(b)).

To overcome these problems, we propose combining the ideas of FDA and LPP; more specifically, we evaluate the levels of the between-class scatter and the within-class scatter in a *local* manner. This allows us to attain between-class separation and within-class local structure preservation at the same time. We call our new method *local Fisher discriminant analysis* (LFDA).

3.1 Reformulating FDA

In order to introduce LFDA, let us first reformulate FDA in a *pairwise* manner.



(a) Toy data set 1

(b) Toy data set 2

(c) Toy data set 3

Figure 1: Examples of dimensionality reduction by FDA, LPP and LFDA. Two-dimensional two-class samples are embedded into a one-dimensional space. The line in the figure denotes the one-dimensional embedding space (which the data samples are projected on) obtained by each method.

Lemma 1 *$S^{(w)}$ and $S^{(b)}$ defined by Eqs.* (1) *and* (2) *can be expressed as*

$$S^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(w)}(x_i - x_j)(x_i - x_j)^\top, \tag{5}$$

$$S^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(b)}(x_i - x_j)(x_i - x_j)^\top, \tag{6}$$

where

$$W_{i,j}^{(w)} \equiv \begin{cases} 1/n_\ell & \text{if } y_i = y_j = \ell, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{7}$$

$$W_{i,j}^{(b)} \equiv \begin{cases} 1/n - 1/n_\ell & \text{if } y_i = y_j = \ell, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \tag{8}$$

A proof of Lemma 1 is given in Appendix A. Note that $1/n - 1/n_\ell$ in Eq. (8) is negative while $1/n_\ell$ and $1/n$ in Eqs. (7) and (8) are positive. This implies that if the data pairs in the same class are made close, the within-class scatter matrix $S^{(w)}$ gets 'small' and the between-class scatter matrix $S^{(b)}$ gets 'large'. On the other hand, if the data pairs in different classes are separated from each other, the between-class scatter matrix $S^{(b)}$ gets 'large'. Therefore, we may interpret FDA as keeping the sample pairs in the same class close and the sample pairs in different classes apart. A more formal discussion on the above interpretation is given in Appendix B.

3.2 Definition and Typical Behavior of LFDA

Based on the above pairwise expression, let us define the *local* within-class scatter matrix $\widetilde{S}^{(w)}$ and the *local* between-class scatter matrix $\widetilde{S}^{(b)}$ as follows.

$$\widetilde{S}^{(w)} \equiv \frac{1}{2}\sum_{i,j=1}^{n} \widetilde{W}_{i,j}^{(w)}(x_i - x_j)(x_i - x_j)^\top, \tag{9}$$

$$\widetilde{S}^{(b)} \equiv \frac{1}{2}\sum_{i,j=1}^{n} \widetilde{W}_{i,j}^{(b)}(x_i - x_j)(x_i - x_j)^\top,$$

where

$$\widetilde{W}_{i,j}^{(w)} \equiv \begin{cases} A_{i,j}/n_\ell & \text{if } y_i = y_j = \ell, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{10}$$

$$\widetilde{W}_{i,j}^{(b)} \equiv \begin{cases} A_{i,j}(1/n - 1/n_\ell) & \text{if } y_i = y_j = \ell, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \tag{11}$$

Namely, according to the affinity $A_{i,j}$, we weight the values for the sample pairs in the same class. This means that *far apart* sample pairs in the same class have less influence on $\widetilde{S}^{(w)}$ and $\widetilde{S}^{(b)}$. Note that we do *not* weight the values for the sample pairs in different classes since we want to separate them from each other *irrespective* of the affinity in the original space. From here on, we denote the local counterparts of matrices by symbols with *tilde*.

We define the LFDA transformation matrix T_{LFDA} as

$$T_{LFDA} \equiv \underset{T\in\mathbb{R}^{d\times r}}{\operatorname{argmax}} \left[\operatorname{tr}\left((T^\top \widetilde{S}^{(w)} T)^{-1} T^\top \widetilde{S}^{(b)} T \right) \right]. \tag{12}$$

That is, we look for a transformation matrix T such that *nearby* data pairs in the same class are made 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.

Eq. (12) is of the same form as Eq. (3). Therefore, we can similarly compute an *analytic* form of T_{LFDA} by solving a generalized eigenvalue problem of $\widetilde{S}^{(b)}$ and $\widetilde{S}^{(w)}$. An efficient implementation of LFDA is summarized as a pseudo code in Figure 2 (see Appendix C for detail).

Toy examples of dimensionality reduction by LFDA are illustrated in Figure 1. We used the local scaling method for computing the affinity matrix A (see Appendix D.4). Note that we perform the nearest neighbor search in the local scaling method in a *classwise* manner since we do not need the affinity values for the sample pairs in different classes (see Eqs. 10 and 11). This highly contributes to reducing the computational cost (see Appendix C). Figure 1 shows that LFDA gives desirable results for all three data sets, that is, LFDA can compensate for the drawbacks of FDA and LPP by effectively combining the ideas of FDA and LPP.

If the affinity value $A_{i,j}$ is set to 1 for all sample pairs (i.e., all pairs are 'equally close' to each other), $\widetilde{S}^{(w)}$ and $\widetilde{S}^{(b)}$ agree with $S^{(w)}$ and $S^{(b)}$, respectively, and LFDA is reduced to the original FDA. Therefore, LFDA may be regarded as a natural localized variant of FDA.

3.3 Properties of LFDA

Here we discuss fundamental properties of LFDA.

First, we give an interpretation of LFDA in terms of the 'pointwise scatter'. $\widetilde{S}^{(w)}$ can be expressed as

$$\widetilde{S}^{(w)} = \frac{1}{2}\sum_{i=1}^{n} \frac{1}{n_{y_i}} \widetilde{P}_i^{(w)},$$

where n_{y_i} is the number of samples in the class to which the sample x_i belongs and $\widetilde{P}_i^{(w)}$ is the *pointwise* local within-class scatter matrix around x_i:

$$\widetilde{P}_i^{(w)} \equiv \sum_{j:y_j=y_i} A_{i,j}(x_j - x_i)(x_j - x_i)^\top.$$

Therefore, 'minimizing' $\widetilde{S}^{(w)}$ corresponds to minimizing the weighted sum of the pointwise local within-class scatter matrices over all samples. $\widetilde{S}^{(b)}$ can also be expressed in a similar way as

$$\widetilde{S}^{(b)} = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{1}{n_{y_i}}\right)\widetilde{P}_i^{(w)} + \frac{1}{2n}\sum_{i=1}^{n} P_i^{(b)}, \tag{13}$$

where $P_i^{(b)}$ is the *pointwise* between-class scatter matrix around x_i:

$$P_i^{(b)} \equiv \sum_{j:y_j\neq y_i} (x_j - x_i)(x_j - x_i)^\top.$$

```
    Input:    Labeled samples ${(x_i, y_i) \mid x_i \in \mathbb{R}^d, y_i \in \{1,2,\ldots,c\}\}_{i=1}^{n}$
              Dimensionality of embedding space $r$ $(1 \le r \le d)$
    Output:   $d \times r$ transformation matrix $T_{LFDA}$

1:  $\widetilde{S}^{(b)} \longleftarrow 0_{d\times d}$;
2:  $\widetilde{S}^{(w)} \longleftarrow 0_{d\times d}$;
3:  for $\ell = 1,2,\ldots,c$      % Compute scatter matrices in a classwise manner
4:      $\{\underline{x}_i\}_{i=1}^{n_\ell} \longleftarrow \{x_j\}_{j:y_j=\ell}$;
5:      for $i = 1,2,\ldots,n_\ell$      % Determine local scaling
6:          $\underline{x}_i^{(7)} \longleftarrow$ 7th nearest neighbor of $\underline{x}_i$ among $\{\underline{x}_j\}_{j=1}^{n_\ell}$;
7:          $\underline{\sigma}_i \longleftarrow \|\underline{x}_i - \underline{x}_i^{(7)}\|$;
8:      end
9:      for $i, j = 1,2,\ldots,n_\ell$      % Define affinity matrix
10:         $\underline{A}_{i,j} \longleftarrow \exp(-\|\underline{x}_i - \underline{x}_j\|^2/(\underline{\sigma}_i\underline{\sigma}_j))$;
11:     end
12:     $\underline{X} \longleftarrow (\underline{x}_1|\underline{x}_2|\cdots|\underline{x}_{n_\ell})$;
13:     $\underline{G} \longleftarrow \underline{X}\mathrm{diag}(\underline{A}1_{n_\ell})\underline{X}^\top - \underline{X}\,\underline{A}\,\underline{X}^\top$;
14:     $\widetilde{S}^{(b)} \longleftarrow \widetilde{S}^{(b)} + \underline{G}/n + (1 - n_\ell/n)\underline{X}\,\underline{X}^\top + \underline{X}1_{n_\ell}(\underline{X}1_{n_\ell})^\top/n$;
15:     $\widetilde{S}^{(w)} \longleftarrow \widetilde{S}^{(w)} + \underline{G}/n_\ell$;
16: end
17: $\widetilde{S}^{(b)} \longleftarrow \widetilde{S}^{(b)} - X1_n(X1_n)^\top/n - \widetilde{S}^{(w)}$;
18: $\{\widetilde{\lambda}_k, \widetilde{\varphi}_k\}_{k=1}^{r} \longleftarrow$ generalized eigenvalues and normalized eigenvectors of
                  $\widetilde{S}^{(b)}\widetilde{\varphi} = \widetilde{\lambda}\widetilde{S}^{(w)}\widetilde{\varphi}$;      % $\widetilde{\lambda}_1 \ge \widetilde{\lambda}_2 \ge \cdots \ge \widetilde{\lambda}_d$
19: $T_{LFDA} = (\sqrt{\widetilde{\lambda}_1}\widetilde{\varphi}_1|\sqrt{\widetilde{\lambda}_2}\widetilde{\varphi}_2|\cdots|\sqrt{\widetilde{\lambda}_r}\widetilde{\varphi}_r)$;
```

Figure 2: Efficient implementation of LFDA (see Appendix C for detail). The affinity matrix is computed by the local scaling method (see Appendix D.4). Matrices and vectors denoted with underline are classwise counterparts of the original ones. $0_{d\times d}$ denotes the $d \times d$ matrix with zeros, 1_{n_ℓ} denotes the n_ℓ-dimensional vector with ones, and $\mathrm{diag}(\underline{A}1_{n_\ell})$ denotes the diagonal matrix with diagonal elements $\underline{A}1_{n_\ell}$. The generalized eigenvectors in line 18 are normalized by Eq. (14), which is often automatically carried out by an eigensolver. The weighting scheme of the eigenvectors in line 19 is explained in Section 3.3. A possible bottleneck of the above implementation is the nearest neighbor search in line 6. This could be alleviated by incorporating the prior knowledge of the data structure or by approximation (see Saul and Roweis, 2003, and references therein). Another possible bottleneck is the computation of $\underline{X}\,\underline{A}\,\underline{X}^\top$ in line 13, which could be eased by sparsely defining the affinity matrix (see Appendix D). A MATLAB implementation is available from '`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA/`'.

Note that $P_i^{(b)}$ does not include the localization factor $A_{i,j}$. Eq. (13) implies that 'maximizing' $\widetilde{S}^{(b)}$ corresponds to minimizing the weighted sum of the pointwise local within-class scatter matrices and maximizing the sum of the pointwise between-class scater matrices.

Next, we discuss the issue of eigenvalue multiplicity in LFDA. The original FDA allows us to extract at most $c-1$ meaningful features since the between-class scatter matrix $S^{(b)}$ has rank at most $c-1$ (Fukunaga, 1990). On the other hand, the local between-class scatter matrix $\widetilde{S}^{(b)}$ generally has a much higher rank with less eigenvalue multiplicity, thanks to the localization factor $A_{i,j}$ included in $\widetilde{W}^{(b)}$ (see Eq. 11). In the simulation shown in Section 5, $\widetilde{S}^{(b)}$ is always full rank for various data sets. Therefore, the proposed LFDA can be practically employed for dimensionality reduction into *any* dimensional spaces. This is a very important and significant improvement over the original FDA.

Finally, we discuss the invariance property of LFDA. The value of the LFDA criterion (12) is *invariant* under linear transformations, that is, for any r-dimensional invertible matrix H, $T_{LFDA}H$ is also a solution of Eq. (12). Therefore, the solution T_{LFDA} is not unique—the *range* of the transformation $H^\top T_{LFDA}^\top$ is uniquely determined, but the *distance metric* (Goldberger et al., 2005; Globerson and Roweis, 2006; Weinberger et al., 2006) in the embedding space can be arbitrary because of the arbitrariness of the matrix H. In practice, we propose determining the LFDA transformation matrix T_{LFDA} as follows. First, we rescale the generalized eigenvectors $\{\widetilde{\varphi}_k\}_{k=1}^d$ so that

$$\widetilde{\varphi}_k \widetilde{S}^{(w)} \widetilde{\varphi}_{k'} = \begin{cases} 1 & \text{if } k = k', \\ 0 & \text{if } k \neq k'. \end{cases} \tag{14}$$

Note that this rescaling is often automatically carried out by an eigensolver. Then we weight each generalized eigenvector by the square root of its associated generalized eigenvalue, that is,

$$T_{LFDA} = (\sqrt{\widetilde{\lambda}_1}\widetilde{\varphi}_1 | \sqrt{\widetilde{\lambda}_2}\widetilde{\varphi}_2 | \cdots | \sqrt{\widetilde{\lambda}_r}\widetilde{\varphi}_r), \tag{15}$$

where $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_d$. This weighting scheme weakens the influence of minor eigenvectors and is shown to work well in experiments (see Section 5).

3.4 Kernel LFDA for Non-Linear Dimensionality Reduction

Here we show how LFDA can be extended to non-linear dimensionality reduction scenarios.

As detailed in Appendix C, the generalized eigenvalue problem that needs to be solved in LFDA can be expressed as

$$X\widetilde{L}^{(b)}X^\top \widetilde{\varphi} = \widetilde{\lambda} X \widetilde{L}^{(w)} X^\top \widetilde{\varphi}, \tag{16}$$

where $\widetilde{L}^{(b)} = \widetilde{L}^{(m)} - \widetilde{L}^{(w)}$ and $\widetilde{L}^{(m)}$ and $\widetilde{L}^{(w)}$ are defined by Eqs. (33) and (35), respectively. Since $X^\top \widetilde{\varphi}$ in Eq. (16) belongs to the range of $X^\top$, it can be expressed by using some vector $\widetilde{\alpha} \in \mathbb{R}^n$ as

$$X^\top \widetilde{\varphi} = X^\top X \widetilde{\alpha} = K \widetilde{\alpha},$$

where K is the n-dimensional matrix with the (i,j)-th element being

$$K_{i,j} \equiv x_i^\top x_j.$$

Then multiplying Eq. (16) by $X^\top$ from the left-hand side yields

$$K\widetilde{L}^{(b)}K\widetilde{\alpha} = \widetilde{\lambda}K\widetilde{L}^{(w)}K\widetilde{\alpha}. \tag{17}$$

This implies that $\{x_i\}_{i=1}^n$ appear only in terms of their *inner products*. Therefore, we can obtain a non-linear variant of LFDA by the *kernel trick* (Vapnik, 1998; Schölkopf et al., 1998), which is explained below.

Let us consider a non-linear mapping $\phi(x)$ from $\mathbb{R}^d$ to a *reproducing kernel Hilbert space* $\mathcal{H}$ (Aronszajn, 1950). Let $K(x,x')$ be the reproducing kernel of $\mathcal{H}$. A typical choice of the kernel function would be the Gaussian kernel:

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right),$$

with $\sigma > 0$. For other choices, see, for example, Wahba (1990), Vapnik (1998), and Schölkopf and Smola (2002). Because of the reproducing property of $K(x,x')$, K is now the *kernel matrix*, that is, the (i,j)-th element is given by

$$K_{i,j} = \langle\phi(x_i),\phi(x_j)\rangle = K(x_i,x_j),$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathcal{H}$.

It can be confirmed that $\widetilde{L}^{(w)}$ is always degenerated (since $\widetilde{L}^{(w)}(1,1,\ldots,1)^\top$ always vanishes; see Eq. 35 for detail). Therefore, $K\widetilde{L}^{(w)}K$ is always degenerated and we cannot directly solve the generalized eigenvalue problem (17). To cope with this problem, we propose regularizing $K\widetilde{L}^{(w)}K$ and solving the following generalized eigenvalue problem instead (cf. Friedman, 1989).

$$K\widetilde{L}^{(b)}K\widetilde{\alpha} = \widetilde{\lambda}(K\widetilde{L}^{(w)}K + \varepsilon I_n)\widetilde{\alpha}, \tag{18}$$

where ε is a small constant. Let $\{\widetilde{\alpha}_k\}_{k=1}^n$ be the generalized eigenvectors associated with the generalized eigenvalues $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n$ of Eq. (18). Then the embedded image of $\phi(x')$ in $\mathcal{H}$ is given by

$$(\sqrt{\widetilde{\lambda}_1}\widetilde{\alpha}_1|\sqrt{\widetilde{\lambda}_2}\widetilde{\alpha}_2|\cdots|\sqrt{\widetilde{\lambda}_r}\widetilde{\alpha}_r)^\top \begin{pmatrix} K(x_1,x') \\ K(x_2,x') \\ \vdots \\ K(x_n,x') \end{pmatrix}.$$

We call this kernelized variant of LFDA *kernel LFDA* (KLFDA).

Recently, kernel functions for non-vectorial structured data such as strings, trees, and graphs have been proposed (see, e.g., Lodhi et al., 2002; Duffy and Collins, 2002; Kashima and Koyanagi, 2002; Kondor and Lafferty, 2002; Kashima et al., 2003; Gärtner et al., 2003; Gärtner, 2003). Since KLFDA uses the samples only via the kernel function $K(x,x')$, it allows us to reduce the dimensionality of such non-vectorial data.

4. Comparison with Related Methods

In this section, we discuss the relation between the proposed LFDA and other methods.

4.1 Dimensionality Reduction Using Local Discriminant Information

A *discriminant adaptive nearest neighbor* (DANN) classifier (Hastie and Tibshirani, 1996a) employs an adapted distance metric at each test point for classification. Based on a similar idea, they also proposed a global supervised dimensionality reduction method using *local discriminant information* (LDI) in the same paper. We refer to this supervised dimensionality reduction method as LDI. The main idea of LDI is to localize FDA—which is very similar to the proposed LFDA. Here we discuss the relation between LDI and LFDA.

In LDI, the data samples $\{x_i\}_{i=1}^n$ are first sphered according to the within-class scatter matrix $S^{(w)}$, that is, for $i = 1, 2, \ldots, n$,

$$\overline{x}_i \equiv (S^{(w)})^{-\frac{1}{2}} x_i.$$

Let $\overline{A}_{i,j}$ be the weight of sample $\overline{x}_j$ around $\overline{x}_i$ defined by

$$\overline{A}_{i,j} \equiv \begin{cases} \left[1 - \left(\frac{\|\overline{x}_i - \overline{x}_j\|}{\|\overline{x}_i - \overline{x}_i^{(K)}\|}\right)^3\right]^3 & \text{if } \|\overline{x}_i - \overline{x}_j\| < \|\overline{x}_i - \overline{x}_i^{(K)}\|, \\ 0 & \text{otherwise.} \end{cases}$$

where $\overline{x}_i^{(K)}$ is the K-th nearest neighbor of $\overline{x}_i$ in the sphered space. Note that $0 \le \overline{A}_{i,j} \le 1$ and $\overline{A}_{i,j}$ is non-increasing as $\|\overline{x}_i - \overline{x}_j\|$ increases. Thus it has the same meaning as our affinity matrix. K is suggested to be determined by

$$K = \max(n/5, 50).$$

Let $\overline{\mu}_\ell^{[i]}$ be the local weighted mean of the sphered samples in class ℓ around $\overline{x}_i$, and let $\overline{\mu}^{[i]}$ be the local weighted mean of the sphered samples around $\overline{x}_i$:

$$\overline{\mu}_\ell^{[i]} \equiv \frac{1}{\overline{n}_\ell^{[i]}} \sum_{j:y_j=\ell} \overline{A}_{i,j} \overline{x}_j,$$

$$\overline{\mu}^{[i]} \equiv \frac{1}{\overline{n}^{[i]}} \sum_{j=1}^{n} \overline{A}_{i,j} \overline{x}_j = \frac{1}{\overline{n}^{[i]}} \sum_{\ell=1}^{c} \overline{n}_\ell^{[i]} \overline{\mu}_\ell^{[i]},$$

where

$$\overline{n}_\ell^{[i]} \equiv \sum_{j:y_j=\ell} \overline{A}_{i,j},$$

$$\overline{n}^{[i]} \equiv \sum_{j=1}^{n} \overline{A}_{i,j}.$$

Let $\overline{S}^{(b)}$ be the *average between sum-of-squares matrix* defined as

$$\overline{S}^{(b)} \equiv \sum_{i=1}^{n} \frac{1}{\overline{n}^{[i]}} \sum_{\ell=1}^{c} \overline{n}_\ell^{[i]} (\overline{\mu}_\ell^{[i]} - \overline{\mu}^{[i]})(\overline{\mu}_\ell^{[i]} - \overline{\mu}^{[i]})^\top.$$

The LDI transformation matrix $\overline{T}_{LDI}$ is defined as

$$\overline{T}_{LDI} \equiv \underset{\overline{T} \in \mathbb{R}^{d \times r}}{\operatorname{argmax}} \left[\overline{T}^\top \overline{S}^{(b)} \overline{T}\right]$$

$$\text{subject to } \overline{T}^\top \overline{T} = I_r.$$

$\overline{T}_{LDI}$ is a transformation matrix for sphered samples; the LDI transformation matrix T_{LDI} for non-sphered samples is given by

$$T_{LDI} = (S^{(w)})^{-\frac{1}{2}} \overline{T}_{LDI}.$$

Similar to FDA (and LFDA), $\overline{T}_{LDI}$ can be efficiently computed by solving a generalized eigenvalue problem.

The average between sum-of-squares matrix $\overline{S}^{(b)}$ is conceptually very similar to the local between-class scatter matrix $\widetilde{S}^{(b)}$ in LFDA. Indeed, as proved in Appendix E, we can express $\overline{S}^{(b)}$ in a pairwise manner as

$$\overline{S}^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} \overline{W}^{(b)}_{i,j} (\overline{x}_i - \overline{x}_j)(\overline{x}_i - \overline{x}_j)^\top, \tag{19}$$

where

$$\overline{W}^{(b)}_{i,j} \equiv \begin{cases} \displaystyle\sum_{k=1}^{n} \frac{1}{\overline{n}^{[k]}} \left(\frac{1}{\overline{n}^{[k]}} - \frac{1}{\overline{n}^{[k]}_{\ell}} \right) \overline{A}_{i,k} \overline{A}_{j,k} & \text{if } y_i = y_j = \ell, \\ \displaystyle\sum_{k=1}^{n} \frac{1}{(\overline{n}^{[k]})^2} \overline{A}_{i,k} \overline{A}_{j,k} & \text{if } y_i \neq y_j. \end{cases} \tag{20}$$

However, there exist critical differences between LDI and LFDA. A significant difference is that the values for the sample pairs in different classes are also localized in LDI (see Eq. 20), while they are kept unlocalized in LFDA (see Eq. 11). This implies that far apart sample pairs in different classes could be made close in LDI, which is not desirable in supervised dimensionality reduction. Furthermore, the computation of $\overline{S}^{(b)}$ is slightly less efficient than $\widetilde{S}^{(b)}$ since $\overline{W}^{(b)}$ includes the summation over k.

Another important difference between LDI and LFDA is that the within-class scatter matrix $S^{(w)}$ is *not* localized in LDI. However, as we showed in Section 3.1, the within-class scatter matrix $S^{(w)}$ also accounts for collapsing the within-class multimodal structure (i.e., far apart sample pairs in the same class are made close). This phenomenon is experimentally confirmed in Section 5.2.

4.2 Mixture Discriminant Analysis

FDA can be interpreted as maximum likelihood estimation of Gaussian distributions with common covariance and different means for each class. Based on this view, Hastie and Tibshirani (1996b) proposed *mixture discriminant analysis* (MDA), which extends FDA to maximum likelihood estimation of Gaussian *mixture* distributions.

A maximum likelihood solution is obtained by an EM-type algorithm (cf. Dempster et al., 1977). However, this is an iterative algorithm and gives only a local optimal solution. Therefore, the computation of MDA is rather slow and there is no guarantee that the global solution can be obtained. Furthermore, the number of mixture components (clusters) in each class as well as the initial location of cluster centers should be determined by users. For cluster centers, using standard techniques such as k-means clustering (MacQueen, 1967; Everitt et al., 2001) or learning vector quantization (Kohonen, 1989) are recommended. However, they are also iterative algorithms and have no guarantee that the global solution can be obtained. Furthermore, there seems to be no systematic method for determining the number of clusters.

On the other hand, the proposed LFDA contains no tuning parameters (given that the affinity matrix is determined by the local scaling method, see Appendix D.4) and the global solution can be obtained analytically. However, it still lacks a probabilistic interpretation, which remains open currently.

4.3 Neighborhood Component Analysis

Goldberger et al. (2005) proposed a supervised dimensionality reduction method called *neighborhood component analysis* (NCA). The NCA transformation matrix T_{NCA} is defined as follows.

$$T_{NCA} \equiv \underset{T \in \mathbb{R}^{d \times r}}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \sum_{j:y_j=y_i} p_{i,j}(TT^\top) \right), \tag{21}$$

where

$$p_{i,j}(U) \equiv \begin{cases} \dfrac{\exp\left\{-(x_i - x_j)^\top U (x_i - x_j)\right\}}{\sum_{k \neq i} \exp\left\{-(x_i - x_k)^\top U (x_i - x_k)\right\}} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{22}$$

The above definition corresponds to maximizing the expected number of correctly classified samples by a stochastic variant of nearest neighbor classifiers. Therefore, NCA seeks a transformation matrix T such that the between-class separability is maximized.

Eqs. (21) and (22) imply that nearby data pairs in the same class are made close, which is similar to the proposed LFDA. Indeed, the simulation results in Section 5.2 show that NCA tends to preserve the multimodal structure of the data very well. However, a crucial weakness of NCA is optimization: the optimization problem (21) is *non-convex*. Therefore, there is no guarantee that the globally optimal solution can be obtained. Goldberger et al. (2005) proposed using a gradient ascent method for optimization:

$$T \longleftarrow T + \varepsilon \nabla J_{NCA}(T), \tag{23}$$

where ε (> 0) is the step size and the gradient $\nabla J_{NCA}(T)$ is given by

$$\nabla J_{NCA}(T) = 2T \sum_{i=1}^{n} \left(\left\{ \sum_{j:y_j=y_i} p_{i,j}(TT^\top) \right\} \left\{ \sum_{j=1}^{n} p_{i,j}(TT^\top)(x_i - x_j)(x_i - x_j)^\top \right\} \right.$$
$$\left. - \sum_{j:y_j=y_i} p_{i,j}(TT^\top)(x_i - x_j)(x_i - x_j)^\top \right).$$

The gradient ascent iteration (23) is computationally rather inefficient. Also, the choice of the step size ε is troublesome. If the step size is small enough, the convergence to one of the local optima is guaranteed but such a choice makes the convergence very slow; on the other hand, if the step size is too large, gradient flows oscillate and proper convergence properties may not be guaranteed anymore. Furthermore, the choice of the termination condition in the iterative algorithm is often cumbersome in practice.

Because of the non-convexity of the optimization problem, the quality of the obtained solution depends on the *initialization* of the matrix T. A useful heuristic to alleviate the local optimum problem is to employ the FDA (or LFDA) result as an initial matrix for optimization (Goldberger et al., 2005). In the experiments in Section 5, using the LFDA result as an initial matrix appears to be better than the random initialization. However, the local optima problem still remains even with the above heuristic.

When a dimensionality reduction technique is applied to classification tasks, we often want to embed the data samples into spaces with several different dimensions—the best dimensionality is later chosen by, for example, cross-validation (Stone, 1974; Wahba, 1990). In such a scenario, NCA requires to optimize the transformation matrix *individually* for each dimensionality r of the embedding space. On the other hand, LFDA needs to compute the transformation matrix *only once* for the largest r; its sub-matrices become the optimal solutions for smaller dimensions. Therefore, LFDA is computationally more efficient than NCA in this scenario.

A simple MATLAB implementation of NCA is available.[4] We use this software in Section 5.

4.4 Maximally Collapsing Metric Learning

In order to overcome the computational problem of NCA, Globerson and Roweis (2006) proposed an alternative method called *maximally collapsing metric learning* (MCML).

Let $p^*_{i,j}$ be the 'ideal' value of $p_{i,j}(U)$ defined by Eq. (22):

$$p^*_{i,j} \propto \begin{cases} 1 & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j, \end{cases}$$

where $p^*_{i,j}$ is normalized so that

$$\sum_{j \neq i} p^*_{i,j} = 1.$$

$p^*_{i,j}$ can be attained if all samples in the same class collapse into a *single* point while samples in other classes are mapped to other locations. In reality, however, any U may not be able to attain $p_{i,j}(U) = p^*_{i,j}$ exactly; instead the optimal approximation to $p^*_{i,j}$ under the *Kullback-Leibler divergence* (Kullback and Leibler, 1951) is obtained. This is formally defined as

$$U_{MCML} \equiv \underset{U \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left(\sum_{i,j=1}^{n} p^*_{i,j} \log \frac{p^*_{i,j}}{p_{i,j}(U)} \right) \quad \text{subject to } U \in PSD(r), \tag{24}$$

where $PSD(r)$ is the set of all positive semidefinite matrices of rank r (i.e., r eigenvalues are positive and others are zero). Once U_{MCML} is obtained, the MCML transformation matrix T_{MCML} is computed by

$$T_{MCML} = (\phi_1 | \phi_2 | \cdots | \phi_r), \tag{25}$$

where $\{\phi_k\}_{k=1}^{r}$ are the eigenvectors associated with the *positive* eigenvalues $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_r > 0$ of the following eigenvalue problem:

$$U_{MCML}\phi = \eta\phi.$$

One of the motivations of MCML is to alleviate the difficulty of optimization in NCA. However, MCML still has a weakness in optimization: the optimization problem (24) is convex only when $r = d$, that is, the dimensionality is not reduced but only the *distance metric* of the original space is changed. This means that if $r < d$ (which is our primal focus in this paper), we may not be able to

[4] Implementation available at '`http://www.cs.berkeley.edu/~fowlkes/software/nca/`'.

obtain the globally optimal solution. Globerson and Roweis (2006) proposed the following heuristic algorithm to *approximate* T_{MCML}.

First, the optimization problem (24) with $r = d$ is solved:

$$\widehat{U}_{MCML} \equiv \operatorname*{argmin}_{U \in \mathbb{R}^{d\times d}} \left(\sum_{i,j=1}^{n} p^*_{i,j} \log \frac{p^*_{i,j}}{p_{i,j}(U)} \right)$$
$$\text{subject to } U \in PSD(d). \tag{26}$$

Although Eq. (26) is convex, an analytic form of the unique optimal solution $\widehat{U}_{MCML}$ is not known yet. Globerson and Roweis (2006) proposed using the following alternate iterative procedure for obtaining $\widehat{U}_{MCML}$.

$$U \longleftarrow U - \varepsilon \nabla J_{MCML}(U), \tag{27}$$

$$U \longleftarrow \sum_{k=1}^{d} \max(0, \widehat{\eta}_k) \widehat{\phi}_k \widehat{\phi}_k^\top, \tag{28}$$

where ε (> 0) is the step size, $\widehat{\eta}_k$ and $\widehat{\phi}_k$ are eigenvalues and eigenvectors of U, and the gradient $\nabla J_{MCML}(U)$ is given by

$$\nabla J_{MCML}(U) = \sum_{i,j=1}^{n} (p^*_{i,j} - p_{i,j}(U))(x_i - x_j)(x_i - x_j)^\top.$$

Then the eigenvalue decomposition of $\widehat{U}_{MCML}$ is carried out and eigenvalues $\widehat{\eta}_1 \geq \widehat{\eta}_2 \geq \cdots \geq \widehat{\eta}_d$ and associated eigenvectors $\{\widehat{\phi}_k\}_{k=1}^{d}$ are obtained:

$$\widehat{U}_{MCML}\widehat{\phi} = \widehat{\eta}\widehat{\phi}.$$

Finally, $\{\phi_k\}_{k=1}^{r}$ in Eq. (25) are replaced by $\{\widehat{\phi}_k\}_{k=1}^{r}$, which yields

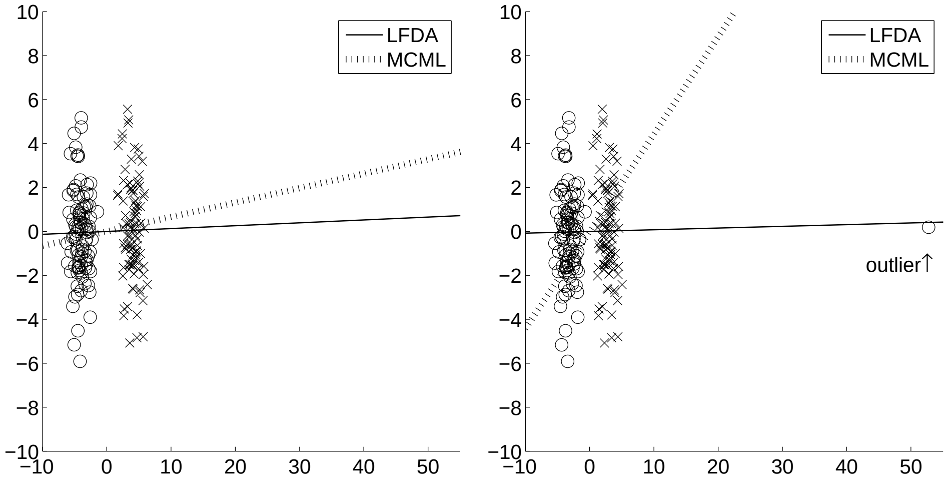
$$T_{MCML} \approx (\widehat{\phi}_1 | \widehat{\phi}_2 | \cdots | \widehat{\phi}_r). \tag{29}$$

This approximation is shown to be practically useful (Globerson and Roweis, 2006), although there seems to be no theoretical analysis for this approximation.

MCML may have an advantage over NCA in computation: there exists the analytic approximation (29) that can be computed efficiently using the solution of another convex optimization problem (26). However, MCML still relies on the gradient-based alternate iterative algorithm (27)–(28) to solve the convex optimization problem (26), which is computationally very expensive since the eigenvalue decomposition of a d-dimensional matrix should be carried out in each iteration (see Eq. 28). Furthermore, the difficulty of appropriately choosing the step size and the termination condition in the iterative procedure still remains.

Since MCML requires all the samples in the same class to collapse into a single point, it is not necessarily useful in dimensionality reduction of multimodal data samples. Furthermore, the MCML results can be significantly influenced by outliers since the outliers are also required to collapse into the same single point together with other samples. This phenomenon is illustrated in Figure 3, where a single outlier significantly changes the MCML result.

Globerson and Roweis (2006) showed that the sufficient statistics of the MCML algorithm are pointwise scatter matrices (cf. Section 3.3). Since LFDA also has an interpretation in terms of pointwise scatter matrices, there may be a link between LFDA and MCML and this needs to be investigated in the future work.



(a) Toy data set 1

(b) Toy data set 1’

Figure 3: Toy examples of dimensionality reduction. The toy data set 1 is equivalent to the one used in Figure 1(a). The data set 1’ includes a single outlier.

4.5 Remark on Rank Constraint

The optimization problem of MCML (see Eq. 24) is not generally convex since the rank constraint is non-convex (Boyd and Vandenberghe, 2004). The non-convexity induced by the rank constraint seems to be a universal problem in dimensionality reduction. NCA eliminates the rank constraint by decomposing U into $TT^\top$ (see Eqs. 21 and 22). However, even with this decomposition, the optimization problem is still non-convex. On the other hand, FDA, LDI, and LFDA cast the optimization problem in the form of the *Rayleigh quotient*. This is computationally very advantageous since it allows us to analytically determine the range of the embedding space. However, we cannot determine the distance metric in the embedding space since the Rayleigh quotient is invariant under linear transformations. For this reason, an additional criterion is needed to determine the distance metric (see also Section 3.3).

5. Numerical Examples

In this section, we numerically evaluate the performance of LFDA and existing methods.

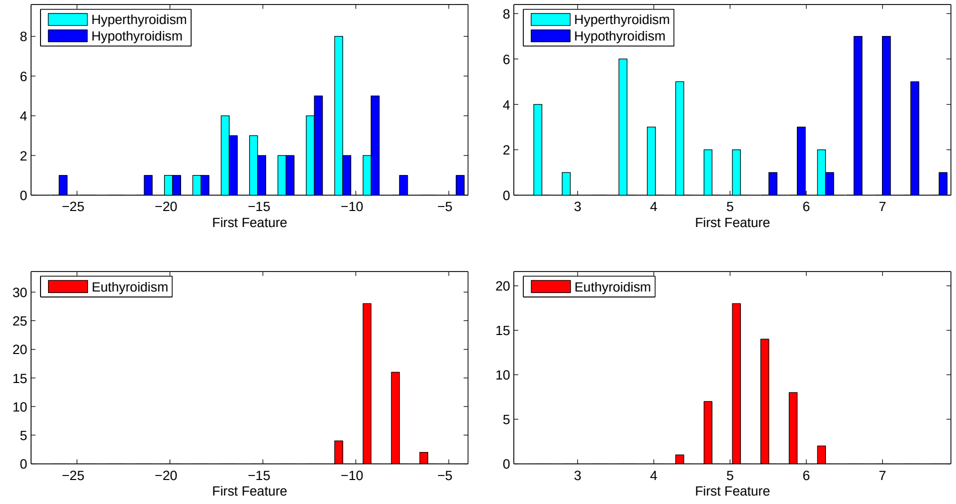
5.1 Exploratory Data Analysis

Here we use the *Thyroid disease* data set available from the UCI machine learning repository (Blake and Merz, 1998) and illustrate how LFDA can be used for exploratory data analysis.

The original data consists of 5-dimensional input vector x of the following laboratory tests.

1. T3-resin uptake test.

2. Total Serum thyroxin as measured by the isotopic displacement method.



(a) FDA

(b) LFDA

Figure 4: Histograms of the first feature values obtained by FDA and LFDA for the *Thyroid disease* data set. The top row corresponds to the sick patients and the bottom row corresponds to the healthy patients.

3. Total Serum triiodothyronine as measured by radioimmuno assay.

4. Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay.

5. Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

The task is to predict whether patients' thyroids are *euthyroidism*, *hypothyroidism*, or *hyperthyroidism* (Coomans et al., 1983), that is, whether patients' thyroids are normal, hypo-functioning, or hyper-functioning (Blake and Merz, 1998). The diagnosis (the class label) is based on a complete medical record, including anamnesis, scan etc. Here we merge the hypothyroidism class and the hyperthyroidism class into a single class and create binary labeled data (whether thyroids are normal or not). Our goal is to predict whether patients' thyroids are normal, hypo-functioning, or hyper-functioning from the binary labeled data samples.

Figure 4 depicts the histograms of the first feature values obtained by FDA and LFDA—the top row corresponds to the sick patients and the bottom row corresponds to the healthy patients. This shows that both FDA and LFDA separate the patients with normal thyroids from sick patients reasonably well. In addition to between-class separability, LFDA clearly preserves the multimodal structure among sick patients (i.e., hypo-functioning and hyper-functioning), which is lost by ordinary FDA. Another interesting finding from the figure is that the first feature values obtained by LFDA has a strong negative correlation to the functioning level of thyroids—this could be used for predicting the functioning level of thyroids.

Data Set	d	'○'-and-'●' class	'×' class
Letter recognition	16	'A' & 'C'	'B'
Iris	4	'Setosa' & 'Virginica'	'Versicolour'

Table 1: Two-class data sets used for visualization experiments ($r = 2$).

5.2 Data Visualization

Here we apply the proposed and existing dimensionality reduction methods to benchmark data sets and investigate how they behave in data visualization tasks.

We use the *Letter recognition* data set and the *Iris* data set available from the UCI machine learning repository (Blake and Merz, 1998). Table 1 describes the specifications of the data sets. Each data set contains three types of samples specified by '○', '●', and '×'. We merged '○' and '●' into a single class and created two-class problems. We test LFDA, FDA, LPP, LDI, NCA, and MCML and evaluate the between-class separability (i.e., '○' and '●' are well separated from '×') and the within-class multimodality preservation capability (i.e., '○' and '●' are well grouped). For LPP and LFDA, we determined the affinity matrix by the local scaling method (see Appendix D.4). For NCA, we used the LFDA result as an initial matrix since this initialization scheme appears to work better than the random initialization. FDA allows us to extract only one meaningful feature in two-class classification problems (see Section 2.2), so we choose the second feature randomly here. Figures 5 and 6 depict the samples embedded in the two-dimensional space found by each method. The horizontal axis is the first feature found by each method, while the vertical axis is the second feature.

First, we compare the embedding results of LFDA with those of FDA and LPP. For the *Letter recognition* data set (see the top row of Figure 5), LFDA nicely separates samples in different classes from each other, and at the same time, it clearly preserves within-class multimodality. FDA separates '○' and '●' from '×' well, but within-class multimodality is lost, that is, '○' and '●' are mixed. LPP gives two separate clusters of samples, but samples in different classes are mixed in one of the clusters. For the *Iris* data set (see the top row of Figure 6), LFDA simultaneously achieves between-class separation and within-class multimodality preservation. On the other hand, FDA tends to mix samples in different classes, which would be caused by within-class multimodality. LPP also works well for this data set because three clusters are well separated from each other in the original high-dimensional space. Overall, LFDA is found to be more appropriate for embedding labeled multimodal data samples than FDA and LPP, implying that our primal goal has been successfully achieved.

Next, we compare the results of LFDA with those of LDI, NCA, and MCML. For the *Letter recognition* data set (see Figure 5), LFDA, LDI, NCA, and MCML separate the samples in different classes from each other very well. However, LDI and MCML collapse '○' and '●' into a single cluster, while LFDA and NCA preserve the multimodal structure clearly. The NCA result is almost identical to the LFDA result (i.e., the initial value of the NCA iteration), but the result may vary if the initial value for the gradient ascent algorithm is changed. For the *Iris* data set (see Figure 6), LFDA, LDI, and NCA work excellently in both between-class separation and within-class multimodality preservation. On the other hand, MCML mixes the samples in different classes. Overall, LDI works fairly well, but the within-class multimodal structure is sometimes lost since LDI only partially takes within-class multimodality into account (see Section 4.1). NCA also works very well, which

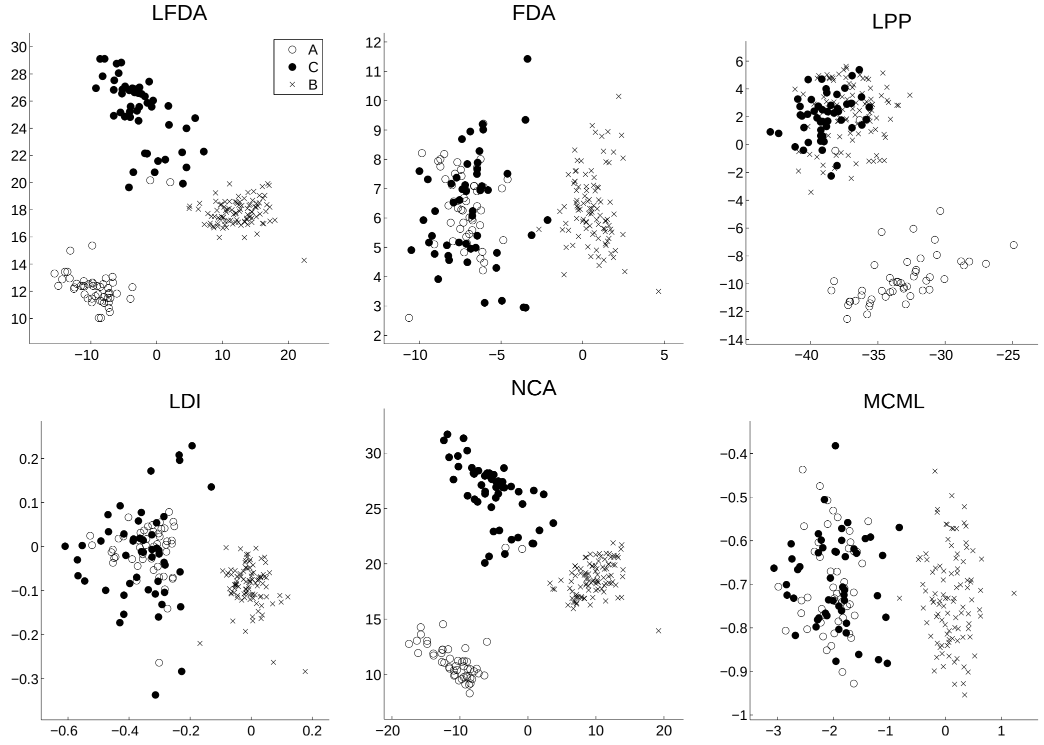


Figure 5: Visualization of the *Letter recognition* data set.

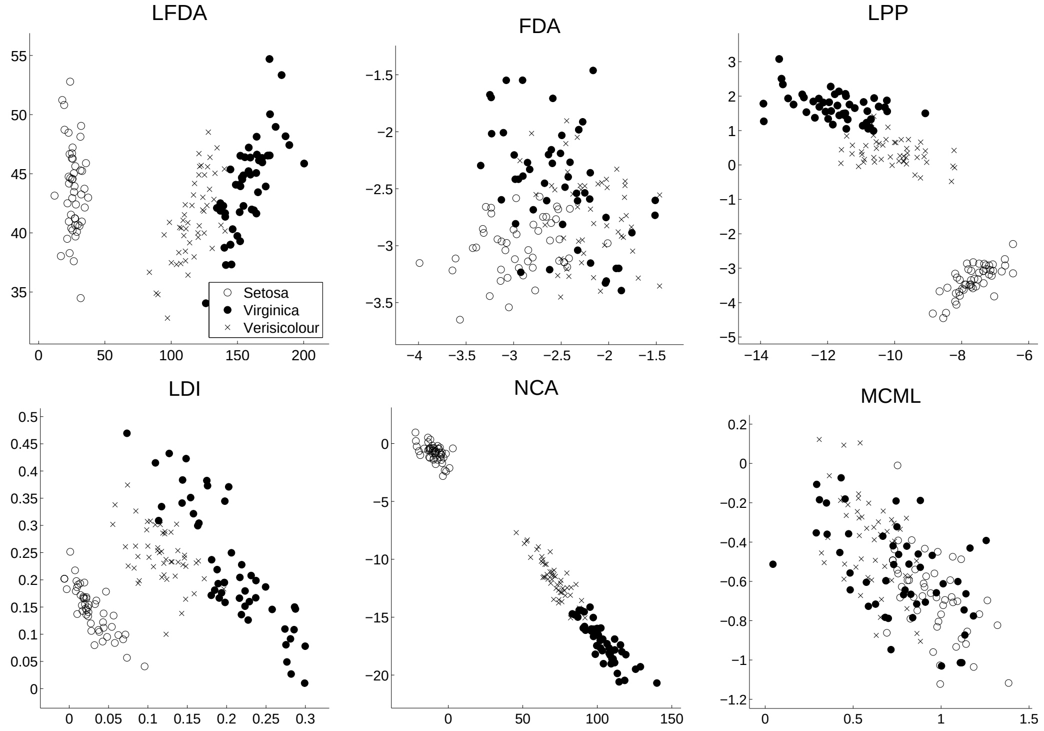


Figure 6: Visualization of the *Iris* data set.

Data name	Input dimensionality	# of training samples	# of test samples	# of realizations
*banana	2	400	4900	100
breast-cancer	9	200	77	100
diabetes	8	468	300	100
flare-solar	9	666	400	100
german	20	700	300	100
heart	13	170	100	100
image	18	1300	1010	20
ringnorm	20	400	7000	100
splice	60	1000	2175	20
*thyroid	5	140	75	100
titanic	3	150	2051	100
twonorm	20	400	7000	100
*waveform	21	400	4600	100
*USPS-eo	256	1000	1000	20
*USPS-sl	256	1000	1000	20

Table 2: List of binary classification data sets. Data sets indicated by '$*$' contain intrinsic within-class multimodal structures.

implies that the heuristic to use the LFDA result as an initial value is useful. However, NCA does not provide significant performance improvement over LFDA in the above simulations. The MCML results have similar tendencies to FDA.

Based on the above simulation results, we conclude that LFDA is a promising method in the visualization of multimodal labeled data.

5.3 Classification

Here we apply the proposed and existing dimensionality reduction techniques to classification tasks, and objectively evaluate the effectiveness of LFDA.

There are several measures for quantitatively evaluating separability of data samples in different classes (e.g., Fukunaga, 1990; Globerson et al., 2005). Here we use a simple one: *misclassification rate by a one-nearest-neighbor classifier*. As explained in Section 3.3, the LFDA criterion is invariant under linear transformations, while the misclassification rate by a one-nearest-neighbor classifier depends on the distance metric. This means that the following simulation results are highly dependent on the normalization scheme (15).

We employ the *IDA* data sets,[5] which are standard binary classification data sets originally used in Rätsch et al. (2001). In addition, we use two binary classification data sets created from the *USPS handwritten digit* data set. The first task (USPS-eo) is to separate even numbers from odd numbers and the second task (USPS-sl) is to separate small numbers ('0' to '4') from large numbers ('5' to '9'). For training and testing, 100 samples are randomly chosen for each digit. Table 2 summarizes

[5] Data sets available at `http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm`.

Data set	LFDA	LDI	NCA	MCML	LPP	PCA
*banana	$^\circ 13.7 \pm 0.8$	$^\circ 13.6 \pm 0.8$	14.3 ± 2.0	39.4 ± 6.7	$^\circ 13.6 \pm 0.8$	$^\circ 13.6 \pm 0.8$
breast-cancer	$^\circ 34.7 \pm 4.3$	36.4 ± 4.9	34.9 ± 5.0	$^\circ 34.0 \pm 5.8$	$^\circ 33.5 \pm 5.4$	$^\circ 34.5 \pm 5.0$
diabetes	32.0 ± 2.5	$^\circ 30.8 \pm 1.9$	—	$^\circ 31.2 \pm 2.1$	31.5 ± 2.5	$^\circ 31.2 \pm 3.0$
flare-solar	$^\circ 39.2 \pm 5.0$	$^\circ 39.3 \pm 4.8$	—	—	$^\circ 39.2 \pm 4.9$	$^\circ 39.1 \pm 5.1$
german	$^\circ 29.9 \pm 2.8$	30.7 ± 2.4	$^\circ 29.8 \pm 2.6$	31.3 ± 2.4	30.7 ± 2.4	$^\circ 30.2 \pm 2.4$
heart	$^\circ 21.9 \pm 3.7$	23.9 ± 3.1	23.0 ± 4.3	23.3 ± 3.8	23.3 ± 3.8	24.3 ± 3.5
image	$^\circ 3.2 \pm 0.8$	$^\circ 3.0 \pm 0.6$	—	4.7 ± 0.8	3.6 ± 0.7	$^\circ 3.4 \pm 0.5$
ringnorm	21.1 ± 1.3	$^\circ 17.5 \pm 1.0$	21.8 ± 1.3	22.0 ± 1.2	20.6 ± 1.1	21.6 ± 1.4
splice	$^\circ 16.9 \pm 0.9$	17.9 ± 0.8	—	$^\circ 17.3 \pm 0.9$	23.2 ± 1.2	22.6 ± 1.3
*thyroid	$^\circ 4.6 \pm 2.6$	8.0 ± 2.9	$^\circ 4.5 \pm 2.2$	18.5 ± 3.8	$^\circ 4.2 \pm 2.9$	$^\circ 4.9 \pm 2.6$
titanic	$^\circ 33.1 \pm 11.9$	$^\circ 33.1 \pm 11.9$	$^\circ 33.0 \pm 11.9$	$^\circ 33.1 \pm 11.9$	$^\circ 33.0 \pm 11.9$	$^\circ 33.0 \pm 12.0$
twonorm	$^\circ 3.5 \pm 0.4$	4.1 ± 0.6	3.7 ± 0.6	$^\circ 3.5 \pm 0.4$	3.7 ± 0.7	3.6 ± 0.6
*waveform	$^\circ 12.5 \pm 1.0$	20.7 ± 2.5	$^\circ 12.6 \pm 0.8$	17.9 ± 1.5	$^\circ 12.4 \pm 1.0$	12.7 ± 1.2
*USPS1	9.0 ± 0.8	12.5 ± 0.9	—	—	7.2 ± 1.0	$^\circ 3.5 \pm 0.7$
*USPS2	12.9 ± 1.2	25.9 ± 1.7	—	11.7 ± 1.3	7.5 ± 0.8	$^\circ 3.9 \pm 0.8$
Computation time (ratio)	1.00	1.11	97.23	70.61	1.04	0.91

Table 3: Means and standard deviations of the misclassification rate when the embedding dimensionality is chosen by cross validation. For each data set, the best method and comparable ones based on the t-test at the significance level 5% are marked by '$\circ$'. Data sets indicated by '$*$' contain the intrinsic within-class multimodal structure.

the specifications of the data sets. The *ringnorm*, *twonorm*, and *waveform* data sets contain features with only noise. The *thyroid*, *waveform*, *USPS-eo*, and *USPS-sl* data sets contain intrinsic within-class multimodal structures since they are converted from multi-class problems by merging some of the classes. The *banana* data set is also multimodal.

We test LFDA, LDI, NCA, MCML, LPP, and principal component analysis (PCA). Note that LPP and PCA are unsupervised dimensionality reduction methods, while others are supervised methods. NCA is not tested for the *diabetes*, *flare-solar*, *image*, *splice*, *USPS-eo*, and *USPS-sl* data sets and MCML is not tested for the *flare-solar* and *USPS-eo* data sets since the execution time is too long.

Figure 7 depicts the mean misclassification rate by a one-nearest-neighbor classifier as functions of the dimensionality r of the reduced space. The error bars are omitted for clear visibility. Instead, we plotted the results of the following significance test: for each dimensionality r, the mean misclassification rate by the best method and comparable ones based on the *t-test* (Henkel, 1979) at the significance level 5% are marked by '$\circ$'. The results show that LFDA works quite well, but overall there is no single best method that consistently outperforms the others.

Table 3 describes the mean and standard deviation of the misclassification rate by each method when the embedding dimensionality r is chosen by 5-*fold cross validation* (Stone, 1974; Wahba, 1990); for the *USPS-eo* and *USPS-sl* data sets, we used 20-*fold cross validation* since this was more accurate. For each data set, the best method and comparable ones based on the t-test at the significance level 5% are indicated by '$\circ$'. The table shows that overall LFDA has excellent

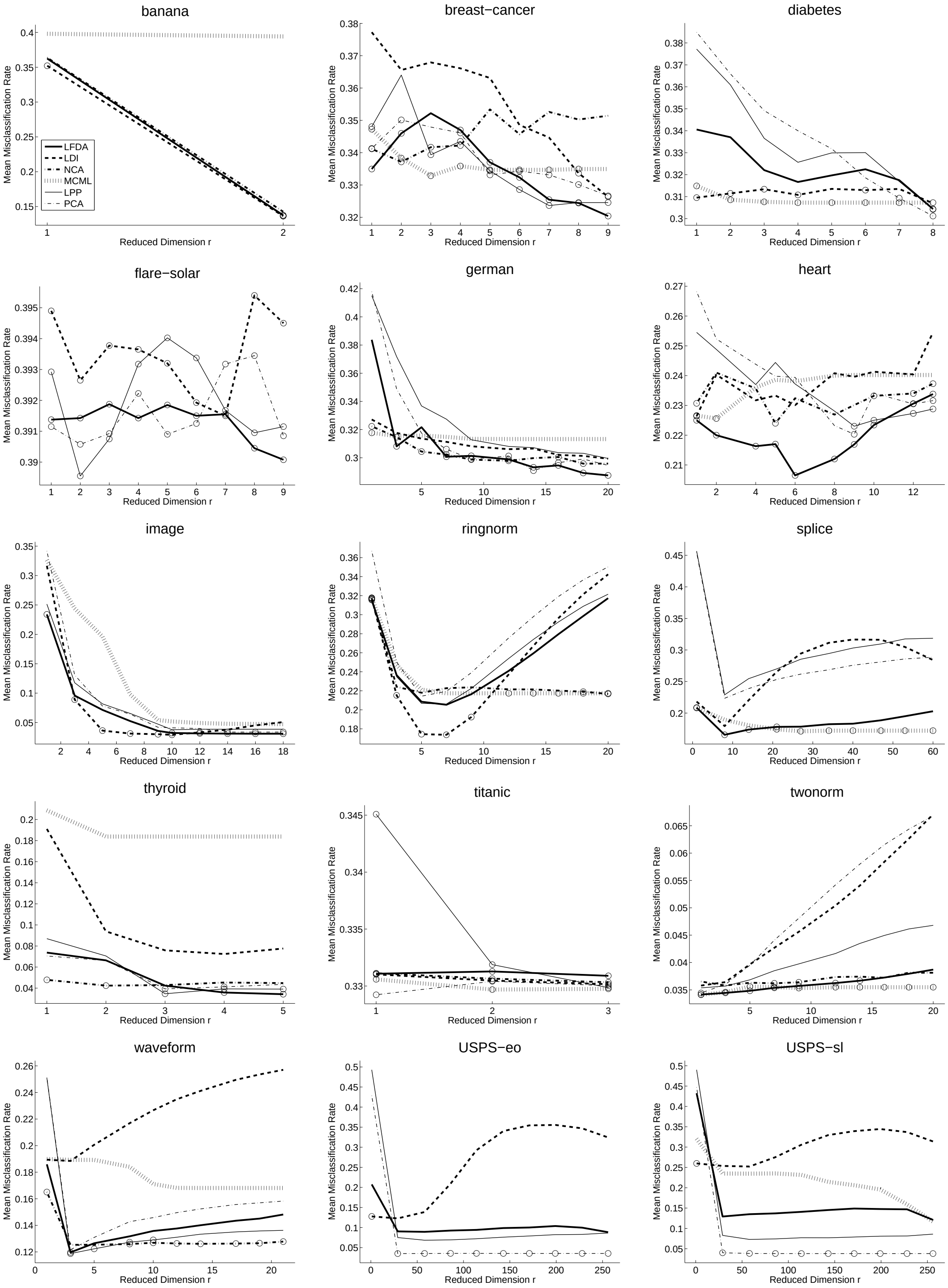


Figure 7: Mean misclassification rates by a one-nearest-neighbor method as functions of the dimensionality of the embedding space. For each dimension, the best method and comparable ones based on the t-test at the significance level 5% are marked by ‘○’.

Data set	LFDA	EUCLID	FDA
*banana	$^{\circ}13.7 \pm 0.8$	$^{\circ}13.6 \pm 0.8$	$^{\circ}13.6 \pm 0.8$
breast-cancer	34.7 ± 4.3	$^{\circ}32.7 \pm 4.8$	$^{\circ}32.9 \pm 4.5$
diabetes	32.0 ± 2.5	$^{\circ}30.1 \pm 2.1$	$^{\circ}30.6 \pm 2.2$
flare-solar	$^{\circ}39.2 \pm 5.0$	$^{\circ}39.2 \pm 5.1$	$^{\circ}39.0 \pm 4.9$
german	$^{\circ}29.9 \pm 2.8$	$^{\circ}29.5 \pm 2.5$	30.5 ± 2.8
heart	$^{\circ}21.9 \pm 3.7$	23.2 ± 3.7	24.0 ± 3.7
image	$^{\circ}3.2 \pm 0.8$	$^{\circ}3.4 \pm 0.5$	6.5 ± 1.7
ringnorm	$^{\circ}21.1 \pm 1.3$	35.0 ± 1.4	31.2 ± 1.6
splice	$^{\circ}16.9 \pm 0.9$	28.9 ± 1.5	33.7 ± 1.5
*thyroid	$^{\circ}4.6 \pm 2.6$	$^{\circ}4.4 \pm 2.2$	5.3 ± 2.5
titanic	$^{\circ}33.1 \pm 11.9$	$^{\circ}33.0 \pm 11.9$	$^{\circ}33.1 \pm 11.9$
twonorm	$^{\circ}3.5 \pm 0.4$	6.7 ± 0.7	5.0 ± 0.7
*waveform	$^{\circ}12.5 \pm 1.0$	15.8 ± 0.7	17.6 ± 1.4
*USPS-eo	9.0 ± 0.8	$^{\circ}3.6 \pm 0.7$	15.1 ± 2.4
*USPS-sl	12.9 ± 1.2	$^{\circ}3.8 \pm 0.8$	13.3 ± 2.9

Table 4: Means and standard deviations of the misclassification rate. The LFDA results are copied from Table 3. 'EUCLID' denotes a naive one-nearest-neighbor classification without dimensionality reduction. 'FDA' denotes a naive one-nearest-neighbor classification after the samples are projected onto a one-dimensional FDA subspace.

performance. LDI and MCML also work quite well, but they tend to perform rather poorly for the multimodal data sets specified by '$*$'. NCA also works well, but it does not compare favorably with LFDA. Note that NCA with random initialization was slightly worse; therefore our heuristic to use the LFDA results for initialization would be reasonable. LPP and PCA perform well, despite the fact that they are unsupervised dimensionality reduction methods. In particular, PCA has excellent performance for the USPS data sets since the projection onto the two-dimensional PCA subspace already gives reasonably separate embedding (He and Niyogi, 2004).

The computation time of each method summed over 9 data sets for which NCA is tested is described in the bottom of Table 3. For better comparison, we normalized the values by the computation time of LFDA. This shows that LFDA is much faster than NCA and MCML,[6] and is comparable to LDI, LPP, and PCA.

The misclassification rates by a naive one-nearest-neighbor classification without dimensionality reduction ('EUCLID') are described in Table 4. The table shows that, on the whole, the performance of LFDA is comparable to EUCLID. This implies that the use of LFDA is advantageous when the dimensionality of the original data is very high since the computation time in the test phase can be reduced. Table 4 also includes the misclassification rates by a naive one-nearest-neighbor classification after the samples are projected onto a one-dimensional FDA subspace, showing that LFDA tends to outperform FDA.

[6] In our implementation of MCML, we used a constant step size for the gradient descent. The computation time could be improved if, for example, an Armijo like step size rule (Bertsekas, 1976) is employed.

Based on the above simulation results, we conclude that the proposed LFDA is a promising dimensionality reduction technique also in classification scenarios.

6. Conclusions

We discussed the problem of supervised dimensionality reduction. FDA (Fisher, 1936; Fukunaga, 1990; Duda et al., 2001) works well for this purpose, given that data samples in each class are *unimodal* Gaussian. However, samples in a class are often multimodal in practice, for example, when multi-class classification problems are solved by a set of two-class 'one-versus-rest' problems. LPP (He and Niyogi, 2004) can work well in dimensionality reduction of multimodal data. However, it is an unsupervised method and does not necessarily useful in supervised dimensionality reduction scenarios. In this paper, we proposed a new method called LFDA, which effectively combines the ideas of FDA and LPP. LFDA allows us to reduce the dimensionality of multimodal labeled data appropriately by maximizing between-class separability and preserving the within-class *local* structure at the same time. The derivation of LFDA is based on a novel *pairwise* interpretation of FDA (see Section 3.1). The original FDA provides a meaningful result only when the dimensionality of the embedding space is smaller than the number of classes because of the rank deficiency of the between-class scatter matrix. On the other hand, LFDA does not share this limitation and can be employed for dimensionality reduction into *any* dimensional spaces (see Section 3.3). This is a significant improvement over the original FDA.

As discussed in Section 3.3, the LFDA criterion is invariant under linear transformations. This means that the *range* of the transformation matrix can be uniquely determined, but the *distance metric* in the embedding space cannot be determined. In this paper, we determined the distance metric in a heuristic manner. Although this normalization scheme is shown to be reasonable in experiments, there is still room for further improvement. An important future direction is to develop a computationally efficient method of determining the distance metric of the embedding space, for example, following the lines of Goldberger et al. (2005), Globerson and Roweis (2006), and Weinberger et al. (2006).

We showed in Section 3.4 that a non-linear variant of LFDA can be obtained by employing the kernel trick. FDA, LPP, and MCML can also be kernelized similarly (Baudat and Anouar, 2000; Mika et al., 2003; Belkin and Niyogi, 2003; He and Niyogi, 2004; Globerson and Roweis, 2006). As shown in these papers, the performance of the kernelized methods heavily depend on the choice of the family and parameters of kernel functions. Therefore, how to optimally determine the kernel function for supervised dimensionality reduction needs to be explored.

The performance of LFDA depends on the choice of the affinity matrix. In this paper, we simply employed a standard definition as it is (see Appendix D.4). Although this standard choice appeared to be reasonable in experiments, it is important to find the optimal way to define the affinity matrix in the context of supervised dimensionality reduction.

MDA (Hastie and Tibshirani, 1996b) provides a solid probabilistic framework for supervised dimensionality reduction with multimodality (see Section 4.2). On the other hand, LFDA still lacks a probabilistic interpretation. An interesting future direction is to analyze the behavior of LFDA in terms of density models.

Acknowledgments

The author would like to thank Klaus-Robert Müller, Hideki Asoh, Motoaki Kawanabe, and Stefan Harmeling for fruitful discussions. His gratitude also goes to anonymous reviewers for their valuable comments; particularly, the weighting scheme of the eigenvectors in Eq. (15) was suggested by one of the reviewers. He also acknowledges financial support from MEXT (Grant-in-Aid for Young Scientists 17700142 and Grant-in-Aid for Scientific Research (B) 18300057). A part of this work has been done when the author was staying at University of Edinburgh, which is supported by EU Erasmus Mundus Scholarship. He would like to thank Sethu Vijayakumar for the warm hospitality.

Appendix A. Proof of Lemma 1

It follows from Eq. (1) that

$$\begin{aligned}
S^{(w)} &= \sum_{\ell=1}^{c}\sum_{i:y_i=\ell}\left(x_i - \frac{1}{n_\ell}\sum_{j:y_j=\ell} x_j\right)\left(x_i - \frac{1}{n_\ell}\sum_{j:y_j=\ell} x_j\right)^\top \\
&= \sum_{i=1}^{n} x_i x_i^\top - \sum_{\ell=1}^{c}\frac{1}{n_\ell}\sum_{i,j:y_i=y_j=\ell} x_i x_j^\top \\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} W_{i,j}^{(w)}\right) x_i x_i^\top - \sum_{i,j=1}^{n} W_{i,j}^{(w)} x_i x_j^\top \\
&= \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}^{(w)}(x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top),
\end{aligned}$$

which yields Eq. (5). Let $S^{(m)}$ be the *mixture scatter matrix* (Fukunaga, 1990):

$$\begin{aligned}
S^{(m)} &\equiv S^{(w)} + S^{(b)} \\
&= \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top.
\end{aligned}$$

Then we have

$$\begin{aligned}
S^{(b)} &= \sum_{i=1}^{n} x_i x_i^\top - \frac{1}{n}\sum_{i,j=1}^{n} x_i x_j^\top - S^{(w)} \\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{1}{n}\right) x_i x_i^\top - \sum_{i,j=1}^{n}\frac{1}{n} x_i x_j^\top - S^{(w)} \\
&= \frac{1}{2}\sum_{i,j=1}^{n}\left(\frac{1}{n} - W_{i,j}^{(w)}\right)(x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top),
\end{aligned}$$

which yields Eq. (6). ■

Appendix B. Interpretation of FDA

In Section 3.1, we claimed that FDA tries to keep data pairs in the same class ‘close’ and data pairs in different classes ‘apart’. Here we show this claim more formally.

For

$$v_{i,j} \equiv T^\top (x_i - x_j),$$

let us investigate the change in the Fisher criterion (3) when $v_{i,j}$ yields $\alpha v_{i,j}$ with $\alpha > 0$. Note that there does not generally exist a transformation T' that keeps all $v_{i,j}$ and only changes a particular pair. For this reason, the following analysis may be regarded as comparing the values of the Fisher criterion for two *different* data sets. This analysis will give an insight into what kind of transformation matrices the Fisher criterion favors.

Let

$$\begin{aligned}
W &\equiv T^\top S^{(w)} T,\\
B &\equiv T^\top S^{(b)} T,\\
W_\alpha &\equiv W - \beta W_{i,j}^{(w)} v_{i,j} v_{i,j}^\top,\\
B_\alpha &\equiv B - \beta W_{i,j}^{(b)} v_{i,j} v_{i,j}^\top,\\
\beta &\equiv \frac{1-\alpha^2}{2}.
\end{aligned}$$

Note that W_α and B_α correspond to the within-class and between-class scatter matrices for $\alpha v_{i,j}$, respectively. We assume that W and W_α are positive definite and B and B_α are positive semi-definite. Then the values of the Fisher criterion (3) for $v_{i,j}$ and $\alpha v_{i,j}$ are expressed as $\mathrm{tr}\left(W^{-1}B\right)$ and $\mathrm{tr}\left(W_\alpha^{-1}B_\alpha\right)$, respectively.

The standard *matrix inversion lemma* (e.g., Albert, 1972) yields

$$\begin{aligned}
W^{-1} &= (W_\alpha + \beta W_{i,j}^{(w)} v_{i,j} v_{i,j}^\top)^{-1}\\
&= W_\alpha^{-1} - \frac{W_\alpha^{-1} v_{i,j} (W_\alpha^{-1} v_{i,j})^\top}{(\beta W_{i,j}^{(w)})^{-1} + \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle}.
\end{aligned}$$

If $y_i = y_j$, we have $W_{i,j}^{(w)} > 0$ and $W_{i,j}^{(b)} < 0$. Then we have

$$\begin{aligned}
\mathrm{tr}\left(W^{-1}B\right) &= \mathrm{tr}\left((W_\alpha + \beta W_{i,j}^{(w)} v_{i,j} v_{i,j}^\top)^{-1}(B_\alpha + \beta W_{i,j}^{(b)} v_{i,j} v_{i,j}^\top)\right)\\
&= \mathrm{tr}\left(W_\alpha^{-1} B_\alpha\right) + \beta W_{i,j}^{(b)} \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle\\
&\quad - \frac{\langle B_\alpha W_\alpha^{-1} v_{i,j}, W_\alpha^{-1} v_{i,j}\rangle + \beta W_{i,j}^{(b)} \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle^2}{(\beta W_{i,j}^{(w)})^{-1} + \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle}\\
&= \mathrm{tr}\left(W_\alpha^{-1} B_\alpha\right) - \frac{\langle B_\alpha W_\alpha^{-1} v_{i,j}, W_\alpha^{-1} v_{i,j}\rangle - W_{i,j}^{(b)} W_{i,j}^{(w)^{-1}} \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle}{(\beta W_{i,j}^{(w)})^{-1} + \langle W_\alpha^{-1} v_{i,j}, v_{i,j}\rangle}.
\end{aligned} \qquad (30)$$

If $\alpha < 1$, we have $\beta > 0$ since $\alpha > 0$ by definition. Therefore, Eq. (30) yields

$$\mathrm{tr}\left(W_\alpha^{-1} B_\alpha\right) > \mathrm{tr}\left(W^{-1} B\right),$$

where we used the facts that W_α is positive definite and B_α is positive semi-definite. This implies that the value of the Fisher criterion increases if a data pair in the same class is made close.

Similarly, if $y_i \neq y_j$, we have $W_{i,j}^{(w)} = 0$ and $W_{i,j}^{(b)} > 0$. Then we have

$$\begin{aligned}\operatorname{tr}\left(W_\alpha^{-1} B_\alpha\right) &= \operatorname{tr}\left((W - \beta W_{i,j}^{(w)} v_{i,j} v_{i,j}^\top)^{-1} (B - \beta W_{i,j}^{(b)} v_{i,j} v_{i,j})\right) \\ &= \operatorname{tr}\left(W^{-1} B\right) - \beta W_{i,j}^{(b)} \langle W^{-1} v_{i,j}, v_{i,j} \rangle. \end{aligned} \tag{31}$$

If $\alpha > 1$, we have $\beta < 0$ and hence Eq. (31) yields

$$\operatorname{tr}\left(W_\alpha^{-1} B_\alpha\right) > \operatorname{tr}\left(W^{-1} B\right).$$

This implies that the value of the Fisher criterion increases if a pair of samples in different classes are separated from each other.

Appendix C. Efficient Computation of T_{LFDA}

As shown in Eq. (15), the LFDA transformation matrix T_{LFDA} can be computed analytically using the generalized eigenvectors and generalized eigenvalues of the following generalized eigenvalue problem.

$$\widetilde{S}^{(b)} \widetilde{\varphi} = \widetilde{\lambda} \widetilde{S}^{(w)} \widetilde{\varphi}.$$

Given $\widetilde{S}^{(b)}$ and $\widetilde{S}^{(w)}$, the computational complexity of calculating T_{LFDA} is $O(rd^2)$. Here, we provide an efficient computing method of $\widetilde{S}^{(b)}$ and $\widetilde{S}^{(w)}$.

Let $\widetilde{S}^{(m)}$ be the *local mixture scatter matrix* defined by

$$\widetilde{S}^{(m)} \equiv \widetilde{S}^{(b)} + \widetilde{S}^{(w)}.$$

From Eqs. (9)–(11), we can immediately show that $\widetilde{S}^{(m)}$ is expressed in the following pairwise form.

$$\widetilde{S}^{(m)} = \frac{1}{2} \sum_{i,j=1}^{n} \widetilde{W}_{i,j}^{(m)} (x_i - x_j)(x_i - x_j)^\top,$$

where $\widetilde{W}^{(m)}$ is the n-dimensional matrix with (i, j)-th element being

$$\widetilde{W}_{i,j}^{(m)} \equiv \begin{cases} A_{i,j}/n & \text{if } y_i = y_j, \\ 1/n & \text{if } y_i \neq y_j. \end{cases}$$

Since

$$\begin{aligned}\widetilde{S}^{(m)} &= \frac{1}{2} \sum_{i,j=1}^{n} \widetilde{W}_{i,j}^{(m)} (x_i x_i^\top + x_j x_j^\top - x_i x_j^\top - x_j x_i^\top) \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \widetilde{W}_{i,j}^{(m)} \right) x_i x_i^\top - \sum_{i,j=1}^{n} \widetilde{W}_{i,j}^{(m)} x_i x_j^\top, \end{aligned}$$

$\widetilde{S}^{(m)}$ can be expressed in a matrix form as

$$\widetilde{S}^{(m)} = X \widetilde{L}^{(m)} X^\top, \tag{32}$$

where

$$\widetilde{L}^{(m)} \equiv \widetilde{D}^{(m)} - \widetilde{W}^{(m)}, \tag{33}$$

and $\widetilde{D}^{(m)}$ is the n-dimensional diagonal matrix with i-th diagonal element being

$$\widetilde{D}^{(m)}_{i,i} \equiv \sum_{j=1}^{n} \widetilde{W}^{(m)}_{i,j}.$$

Similarly, $\widetilde{S}^{(w)}$ can be expressed in a matrix form as

$$\widetilde{S}^{(w)} = X\widetilde{L}^{(w)}X^\top, \tag{34}$$

where

$$\widetilde{L}^{(w)} \equiv \widetilde{D}^{(w)} - \widetilde{W}^{(w)}, \tag{35}$$

and $\widetilde{D}^{(w)}$ is the n-dimensional diagonal matrix with i-th diagonal element being

$$\widetilde{D}^{(w)}_{i,i} \equiv \sum_{j=1}^{n} \widetilde{W}^{(w)}_{i,j}.$$

$\widetilde{L}^{(m)}$ and $\widetilde{L}^{(w)}$ are n-dimensional matrices and could be very high dimensional. However, $\widetilde{L}^{(w)}$ can be made block-diagonal if the samples $\{x_i\}_{i=1}^n$ are sorted according to the labels $\{y_i\}_{i=1}^n$. Furthermore, diagonal sub-matrices of $\widetilde{L}^{(w)}$ can be sparse if the affinity matrix A is sparsely defined (see Appendix D for detail). Therefore, directly calculating $\widetilde{S}^{(w)}$ by Eq. (34) may be already computationally efficient.

On the other hand, computing $\widetilde{S}^{(m)}$ directly by Eq. (32) is not so efficient since $\widetilde{W}^{(m)}$ is *dense*. This problem can be alleviated as follows. $\widetilde{W}^{(m)}$ can be decomposed as

$$\widetilde{W}^{(m)} = \frac{1}{n}1_n 1_n^\top + \widetilde{W}^{(m_1)} + \widetilde{W}^{(m_2)},$$

where 1_n is the n-dimensional vector with all ones and $\widetilde{W}^{(m_1)}$ and $\widetilde{W}^{(m_2)}$ are the n-dimensional matrices with (i,j)-th element being

$$\widetilde{W}^{(m_1)}_{i,j} \equiv \begin{cases} A_{i,j}/n & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j, \end{cases}$$

$$\widetilde{W}^{(m_2)}_{i,j} \equiv \begin{cases} -1/n & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j. \end{cases}$$

Then $\widetilde{S}^{(m)}$ can be expressed as

$$\widetilde{S}^{(m)} = X\widetilde{D}^{(m)}X^\top - \frac{1}{n}X1_n(X1_n)^\top - X\widetilde{W}^{(m_1)}X^\top - X\widetilde{W}^{(m_2)}X^\top, \tag{36}$$

where the diagonal matrix $\widetilde{D}^{(m)}$ is expressed in terms of $\widetilde{W}^{(m_1)}$ as

$$\widetilde{D}^{(m)}_{i,i} = 1 - \frac{n_{y_i}}{n} + \sum_{j=1}^{n} \widetilde{W}^{(m_1)}_{i,j}.$$

Note that n_{y_i} in the above equation is the number of samples in the class which the sample x_i belongs to. $\widetilde{W}^{(m_2)}$ is a constant block-diagonal matrix if the samples $\{x_i\}_{i=1}^n$ are sorted according to the labels $\{y_i\}_{i=1}^n$. Therefore, $X\widetilde{W}^{(m_2)}X^\top$ in the right-hand side of Eq. (36) can be computed efficiently. Similarly, $\widetilde{W}^{(m_1)}$ can also be made block-diagonal, so $X\widetilde{W}^{(m_1)}X^\top$ in the right-hand side of Eq. (36) may also be computed efficiently; if the affinity matrix A is sparse, the computational efficiency can be further improved. The first two terms in the right-hand side of Eq. (36) can also be computed efficiently. Therefore, computing $\widetilde{S}^{(m)}$ by Eq. (36) may be more efficient than directly by Eq. (32). Finally, we can compute $\widetilde{S}^{(b)}$ efficiently by using $\widetilde{S}^{(m)}$ as

$$\widetilde{S}^{(b)} = \widetilde{S}^{(m)} - \widetilde{S}^{(w)}.$$

To further improve computational efficiency, the affinity matrix A may be computed in a class-wise manner since we do not need the affinity values for sample pairs in different classes. This speeds up the nearest neighbor search which is often carried out when defining A (see Appendix D). The nearest neighbor search itself could also be a bottleneck, but this may be eased by incorporating the prior knowledge of the data structure or by approximation (see Saul and Roweis, 2003, and references therein).

The above efficient implementation of LFDA is summarized as a pseudo code in Figure 2.

Appendix D. Definitions of Affinity Matrix

Here, we briefly review typical choices of the affinity matrix A.

D.1 Heat Kernel

A standard choice of the affinity matrix A is

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right), \tag{37}$$

where $\sigma\ (> 0)$ is a tuning parameter which controls the 'decay' of the affinity (e.g., Belkin and Niyogi, 2003).

D.2 Euclidean Neighbor

The heat kernel gives a non-sparse affinity matrix. It would be computationally advantageous if the affinity matrix is made sparse. A sparse affinity matrix can be obtained by assigning positive affinity values only to neighboring samples. More specifically, x_i and x_j are said to be *neighbors* if

$$\|x_i - x_j\| \leq \varepsilon,$$

where $\varepsilon\ (> 0)$ is a tuning parameter. Then $A_{i,j}$ is defined by Eq. (37) for two neighboring samples and $A_{i,j} = 0$ for non-neighbors (Tenenbaum et al., 2000).

This definition includes two tuning parameters (ε and σ), which are rather troublesome to determine in practice. To ease the problem, we may simply let $A_{i,j} = 1$ if x_i and x_j are neighbors and $A_{i,j} = 0$ otherwise. This corresponds to setting $\sigma = \infty$.

D.3 Nearest Neighbor

Tuning the distance threshold ε is practically rather cumbersome since the relation between the number of neighbors and the value of ε is not intuitively clear. Another option to determine the neighbors is to directly specify the number of neighbors (Roweis and Saul, 2000; Tenenbaum et al., 2000). Let $NN_i^{(K)}$ be the set of K nearest neighbor samples of x_i under the Euclidean distance, where K is a tuning parameter. If $x_j \in NN_i^{(K)}$ or $x_i \in NN_j^{(K)}$, x_i and x_j are regarded as neighbors; otherwise they are regarded as non-neighbors. Then the affinity matrix is defined by the heat kernel or in the simple zero-one manner.

D.4 Local Scaling

A drawback of the above definitions could be that the affinity is computed globally in the same way. The density of data samples may be different depending on regions. Therefore, it would be more appropriate to take the local scaling of the data into account. Following this idea, Zelnik-Manor and Perona (2005) proposed defining the affinity matrix as

$$A_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right).$$

σ_i represents the local scaling of the data samples around x_i, which is determined by

$$\sigma_i = \|x_i - x_i^{(K)}\|,$$

where $x_i^{(K)}$ is the K-th nearest neighbor of x_i. The parameter K is a tuning parameter, but Zelnik-Manor and Perona (2005) demonstrated that $K = 7$ works well on the whole. This would be a convenient heuristic for those who do not have any subjective/prior preferences. We employed the local scaling method with this heuristic all through the paper.

For computational efficiency, we may further sparsify the above affinity matrix based on, for example, the nearest neighbor idea, although this is not tested in this paper.

Appendix E. Pairwise Expression of $\overline{S}^{(b)}$

A pairwise expression of $\overline{S}^{(b)}$ can be derived as

$$
\begin{aligned}
\overline{S}^{(b)} &= \sum_{k=1}^{n} \frac{1}{\overline{n}^{[k]}} \sum_{\ell=1}^{c} \overline{n}_{\ell}^{[k]} (\overline{\mu}_{\ell}^{[k]} - \overline{\mu}^{[k]})(\overline{\mu}_{\ell}^{[k]} - \overline{\mu}^{[k]})^{\top} \\
&= \sum_{k=1}^{n} \frac{1}{\overline{n}^{[k]}} \left(\sum_{\ell=1}^{c} \overline{n}_{\ell}^{[k]} \overline{\mu}_{\ell}^{[k]} \overline{\mu}_{\ell}^{[k]\top} - \overline{n}^{[k]} \overline{\mu}^{[k]} \overline{\mu}^{[k]\top} \right) \\
&= \sum_{k=1}^{n} \frac{1}{\overline{n}^{[k]}} \left(\sum_{\ell=1}^{c} \overline{n}_{\ell}^{[k]} \overline{\mu}_{\ell}^{[k]} \overline{\mu}_{\ell}^{[k]\top} - \sum_{i=1}^{n} \overline{A}_{i,k} \overline{x}_i \overline{x}_i^{\top} + \sum_{i=1}^{n} \overline{A}_{i,k} \overline{x}_i \overline{x}_i^{\top} - \overline{n}^{[k]} \overline{\mu}^{[k]} \overline{\mu}^{[k]\top} \right) \\
&= \frac{1}{2} \sum_{k=1}^{n} \frac{1}{\overline{n}^{[k]}} \Bigg(- \sum_{\ell=1}^{c} \frac{1}{\overline{n}_{\ell}^{[k]}} \sum_{i,j: y_i = y_j = \ell} \overline{A}_{i,k} \overline{A}_{j,k} (\overline{x}_i - \overline{x}_j)(\overline{x}_i - \overline{x}_j)^{\top} \\
&\qquad\qquad\qquad\qquad + \frac{1}{\overline{n}^{[k]}} \sum_{i,j=1}^{n} \overline{A}_{i,k} \overline{A}_{j,k} (\overline{x}_i - \overline{x}_j)(\overline{x}_i - \overline{x}_j)^{\top} \Bigg) \\
&= \frac{1}{2} \sum_{i,j=1}^{n} \overline{W}_{i,j}^{(b)} (\overline{x}_i - \overline{x}_j)(\overline{x}_i - \overline{x}_j)^{\top},
\end{aligned}
$$

which yields Eqs. (19) and (20). ■

References

A. Albert. *Regression and the Moore-Penrose Pseudoinverse*. Academic Press, New York and London, 1972.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

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

D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2):174–184, 1976.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, R.I., 1997.

D. Coomans, M. Broeckaert, M. Jonckheer, and D. L. Massart. Comparison of multivariate discriminant techniques for clinical data—Application to the thyroid functional state. *Methods of Information in Medicine*, 22:93–101, 1983.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39(1):1–38, 1977.

R. O. Duda, P. E. Hart, and D. G. Stor. *Pattern Classification*. Wiley, New York, 2001.

N. Duffy and M. Collins. Convolution kernels for natural language. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold, London, 2001.

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

J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Inc., Boston, second edition, 1990.

T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations*, 5(1):S268–S275, 2003.

T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, 2003.

A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 497–504. MIT Press, Cambridge, MA, 2005.

A. Globerson and S. Roweis. Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 451–458, Cambridge, MA, 2006. MIT Press.

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, Cambridge, MA, 2005.

J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, New York, NY, 2004. ACM Press.

T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–615, 1996a.

T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, 58(1):155–176, 1996b.

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

R. E. Henkel. *Tests of Significance*. SAGE Publication, Beverly Hills, 1979.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

H. Kashima and T. Koyanagi. Kernels for semi-structured date. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 291–298, San Francisco, CA, 2002. Morgan Kaufmann.

H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the Twentieth International Conference on Machine Learning*, San Francisco, CA, 2003. Morgan Kaufmann.

T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 1989.

R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315–322, 2002.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA., USA, 1967. University of California Press.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–628, 2003.

G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3): 287–320, 2001.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4(Jun):119–155, 2003.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

G. Wahba. *Spline Model for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480, Cambridge, MA, 2006. MIT Press.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, Cambridge, MA, 2005.