# Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction

**Masashi Sugiyama** SUGI@CS.TITECH.AC.JP
Department of Computer Science, Tokyo Institute of Technology, 2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

## Abstract

Dimensionality reduction is one of the important preprocessing steps in high-dimensional data analysis. In this paper, we consider the supervised dimensionality reduction problem where samples are accompanied with class labels. Traditional Fisher discriminant analysis is a popular and powerful method for this purpose. However, it tends to give undesired results if samples in some class form several separate clusters, i.e., multimodal. In this paper, we propose a new dimensionality reduction method called local Fisher discriminant analysis (LFDA), which is a localized variant of Fisher discriminant analysis. LFDA takes local structure of the data into account so the multimodal data can be embedded appropriately. We also show that LFDA can be extended to non-linear dimensionality reduction scenarios by the kernel trick.

## 1. Introduction

The goal of dimensionality reduction is to embed high-dimensional data samples in a low-dimensional space while most of 'intrinsic information' contained in the data is preserved. Once dimensionality reduction is carried out appropriately, we can utilize the compact representation of the data for various succeeding tasks such as visualization, classification, etc. In this paper, we consider the supervised dimensionality reduction problem where samples are accompanied with class labels.

*Fisher discriminant analysis* (FDA) (Fisher, 1936; Fukunaga, 1990) is a popular method for linear dimensionality reduction[1], which maximizes between-class scatter and minimizes within-class scatter. This traditional FDA is known to work well and is practically useful even now. However, it tends to give undesired results if samples in some class form several separate clusters (i.e., *multimodal*) (Fukunaga, 1990).

Multimodality is often 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 particular disease. Even in a traditional task of hand-written digit recognition, multimodality appears if digits are classified into, e.g., even and odd numbers. More generally, multimodality can naturally appear if multi-class classification problems are solved by a set of two-class 'one-versus-rest' problems.

To embed multimodal data well, it is important to preserve local structure of the data. *Locality-preserving projection* (LPP) (He & Niyogi, 2004) meets this requirement. LPP keeps nearby data pairs in the original space close in the embedding space, by which multimodal data can be embedded without losing its local structure. However, LPP is an unsupervised dimensionality reduction method which 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*

Appearing in *Proceedings of the 23$^{rd}$ International Conference on Machine Learning*, Pittsburgh, PA, 2006. 

[1] Usually, FDA may refer to a classification method which first projects the data samples onto a one-dimensional space and then classifies the samples by thresholding. The one-dimensional embedding space used above is given as the maximizer of the so-called *Fisher criterion*. This Fisher criterion is often used for dimensionality reduction to a subspace with dimension more than one (Fukunaga, 1990). With some abuse, we refer to the dimensionality reduction method based on the Fisher criterion as FDA in the following (see Section 2.2 for detail).

(LFDA). LFDA combines the ideas of FDA and LPP: between-class separability is maximized while *within-class local structure* is preserved.

Because of the local structure preservation property, multimodal labeled data can be effectively embedded by LFDA. Furthermore, LFDA can give more separate embedding than FDA. To explain the reason intuitively, we first note that FDA can be regarded as maximizing between-class scatter under the constraint of keeping within-class scatter to a certain level (Fukunaga, 1990). When samples in some class are multimodal, keeping within-class scatter to a certain level appears to be quite hard since multimodal samples should be typically merged into a single cluster. Due to this strong constraint, less degree of freedom is left for increasing separability, and thus FDA results in less separate embedding. On the other hand, LFDA does not require multimodal samples to fall into a single cluster. As a result, more degree of freedom is left for increasing separability and thus highly separate embedding can be obtained.

By the so-called *kernel trick* (Schölkopf & Smola, 2002), FDA and LPP can be extended to non-linear dimensionality reduction scenarios (Mika et al., 2003; Belkin & Niyogi, 2003). We show that LFDA can also be non-linearized by the kernel trick.

## 2. Linear Dimensionality Reduction

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

### 2.1. Formulation

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

$$\textstyle\sum_{i=1}^{\ell} n_i = n. \tag{1}$$

Let $\boldsymbol{X}$ be the matrix of all samples:

$$\boldsymbol{X} = (\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n). \tag{2}$$

Let $\boldsymbol{z}_i \in \mathbb{R}^m$ $(1 \leq m \leq d)$ be embedded samples, where $m$ is the dimension of the embedding space. Effectively we consider $d$ to be large and $m$ to be small, but not limited to such cases.

For the moment, we focus on linear dimensionality reduction, i.e., using a $d \times m$ transformation matrix $\boldsymbol{T}$, $\boldsymbol{z}_i$ is given by

$$\boldsymbol{z}_i = \boldsymbol{T}^\top \boldsymbol{x}_i. \tag{3}$$

Later in Section 3.5, we discuss non-linear dimensionality reduction scenarios.

### 2.2. Fisher Linear Discriminant

Here we briefly review the definition of Fisher criterion for dimensionality reduction (Fisher, 1936; Fukunaga, 1990). With some abuse, we refer to the dimensionality reduction method based on the Fisher criterion as Fisher discriminant analysis (FDA), where the original FDA embeds the data samples only in one-dimensional space.

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

$$\boldsymbol{S}^{(w)} = \sum_{i=1}^{\ell} \sum_{j:y_j=i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top, \tag{4}$$

$$\boldsymbol{S}^{(b)} = \sum_{i=1}^{\ell} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top, \tag{5}$$

where $^\top$ denotes the transpose, $\boldsymbol{\mu}_i$ is the mean of samples in the class $i$ and $\boldsymbol{\mu}$ is the mean of all samples:

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{j:y_j=i} \boldsymbol{x}_j, \tag{6}$$

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i. \tag{7}$$

Using $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$, the FDA transformation matrix $\boldsymbol{T}_{FDA}$ is defined as follows:

$$\boldsymbol{T}_{FDA} = \operatorname*{argmax}_{\boldsymbol{T} \in \mathbb{R}^{d \times m}} \operatorname{tr}\left( (\boldsymbol{T}^\top \boldsymbol{S}^{(w)} \boldsymbol{T})^{-1} \boldsymbol{T}^\top \boldsymbol{S}^{(b)} \boldsymbol{T} \right). \tag{8}$$

That is, $\boldsymbol{T}$ is determined so that between-class scatter is maximized while within-class scatter is minimized.

It is known that $\boldsymbol{T}_{FDA}$ is given by

$$\boldsymbol{T}_{FDA} = (\boldsymbol{\varphi}_1 | \boldsymbol{\varphi}_2 | \cdots | \boldsymbol{\varphi}_m), \tag{9}$$

where $\{\boldsymbol{\varphi}_i\}_{i=1}^d$ are the generalized eigenvectors associated to the generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of the following generalized eigenvalue problem:

$$\boldsymbol{S}^{(b)} \boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(w)} \boldsymbol{\varphi}. \tag{10}$$

### 2.3. Locality-Preserving Projection

Here we briefly review the definition of locality-preserving projection (LPP) (He & Niyogi, 2004). Let $\boldsymbol{A}$ be the *affinity matrix*, i.e., the $n$-dimensional matrix with the $(i, j)$-th element being the affinity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We suppose that elements of $\boldsymbol{A}$ are in $[0, 1]$ and take larger values if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are 'close'.

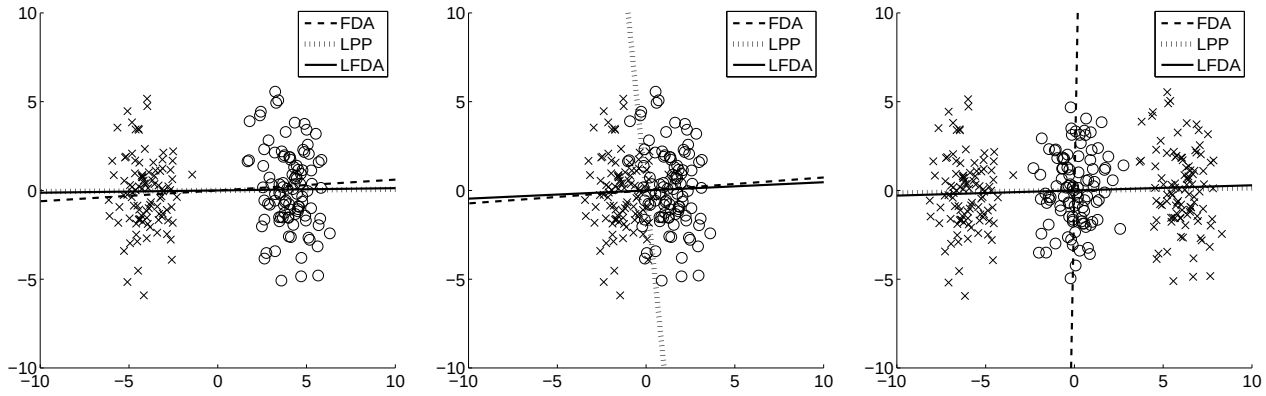


*Figure 1.* Examples of dimensionality reduction by FDA, LPP and LFDA.

There are several different manners in defining $\boldsymbol{A}$. A simple one is to define $\boldsymbol{A}_{i,j} = 1$ if $\boldsymbol{x}_j$ is the $k$-nearest neighbor of $\boldsymbol{x}_i$ or vise versa; otherwise $\boldsymbol{A}_{i,j} = 0$. More elaborate methods can be found, e.g., in (He & Niyogi, 2004; Zelnik-Manor & Perona, 2005). The latter may be useful for those who do not have any subjective/prior preference.

Using $\boldsymbol{A}$, the LPP transformation matrix $\boldsymbol{T}_{LPP}$ is defined as follows.

$$\boldsymbol{T}_{LPP} = \operatorname*{argmin}_{\boldsymbol{T} \in \mathbb{R}^{d \times m}} \frac{1}{2} \sum_{i,j=1}^{n} \boldsymbol{A}_{i,j} \|\boldsymbol{T}^\top \boldsymbol{x}_i - \boldsymbol{T}^\top \boldsymbol{x}_j\|^2$$
$$\text{subject to } \boldsymbol{T}^\top \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^\top \boldsymbol{T} = \boldsymbol{I}, \quad (11)$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{D}$ is the $n$-dimensional diagonal matrix with $i$-th diagonal element being

$$\boldsymbol{D}_{i,i} = \sum_{j=1}^{n} \boldsymbol{A}_{i,j}. \quad (12)$$

Eq.(11) implies that in LPP, $\boldsymbol{T}$ is determined so that *nearby* data pairs in the original space are kept close in the embedding space. Note that $\boldsymbol{T}^\top \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^\top \boldsymbol{T} = \boldsymbol{I}$ is a constraint to avoid a trivial solution $\boldsymbol{T} = \boldsymbol{O}$.

It is known that the LPP transformation matrix $\boldsymbol{T}_{LPP}$ is given by

$$\boldsymbol{T}_{LPP} = (\boldsymbol{\psi}_{d-m+1} | \boldsymbol{\psi}_{d-m+2} | \cdots | \boldsymbol{\psi}_d), \quad (13)$$

where $\{\boldsymbol{\psi}_i\}_{i=1}^d$ are the eigenvectors associated to the eigenvalues $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d$ of the following eigenvalue problem:

$$\boldsymbol{X} \boldsymbol{L} \boldsymbol{X}^\top \boldsymbol{\psi} = \gamma \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^\top \boldsymbol{\psi}, \quad (14)$$

where

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}. \quad (15)$$

### 2.4. Typical Behavior

In Figure 1, dimensionality reduction results obtained by FDA and LPP are shown, where two-dimensional two-class data samples are embedded in a one-dimensional subspace. In LPP, the affinity matrix $\boldsymbol{A}$ is determined by the *local scaling method* (Zelnik-Manor & Perona, 2005).

For the simplest data set depicted in Figure 1(a), both FDA and LPP give reasonable results where samples of different classes are nicely separated. For the data set depicted in Figure 1(b), FDA still works well. However, LPP mixes samples of different classes into one cluster, which 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 gives an undesired result. The reason is that the levels of between-class scatter and within-class scatter are not evaluated in an intuitively natural way because of the two separate clusters of the same class. See also (Fukunaga, 1990).

## 3. Local Fisher Discriminant Analysis

As illustrated above, FDA can perform poorly if samples in some class form several separate clusters (i.e., multimodal). In other words, the undesired behavior is caused by the *globality* when evaluating within-class scatter and between-class scatter. On the other hand, LPP can make samples of different classes overlapped if they are close in the original high-dimensional space.

To overcome this problem, we propose combining the idea of FDA and LPP: we evaluate the levels of within-class scatter and between-class scatter in a *local* manner, by which class separability and local structure

preservation could be attained at the same time. We call our new method *local Fisher discriminant analysis* (LFDA).

### 3.1. Reformulating FDA

To introduce our new method, let us first reformulate FDA in a *pairwise* manner.

**Lemma 1** *$\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$ are expressed as*

$$\boldsymbol{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^{(w)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \tag{16}$$

$$\boldsymbol{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^{(b)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \tag{17}$$

*where*

$$\boldsymbol{A}_{i,j}^{(w)} = \begin{cases} 1/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{18}$$

$$\boldsymbol{A}_{i,j}^{(b)} = \begin{cases} 1/n - 1/n_c & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \tag{19}$$

**(Proof)** It follows from Eq.(4) that

$$\begin{aligned}
\boldsymbol{S}^{(w)} &= \sum_{i=1}^{\ell}\sum_{j:y_j=i}\left(\boldsymbol{x}_j - \frac{1}{n_i}\sum_{p:y_p=i}\boldsymbol{x}_p\right)\left(\boldsymbol{x}_j - \frac{1}{n_i}\sum_{q:y_q=i}\boldsymbol{x}_q\right)^\top \\
&= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sum_{i=1}^{\ell}\frac{1}{n_i}\sum_{p,q:y_p=y_q=i}\boldsymbol{x}_p\boldsymbol{x}_q^\top \\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\boldsymbol{A}_{i,j}^{(w)}\right)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sum_{i,j=1}^{n}\boldsymbol{A}_{i,j}^{(w)}\boldsymbol{x}_i\boldsymbol{x}_j^\top \\
&= \frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{A}_{i,j}^{(w)}(\boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{x}_j\boldsymbol{x}_j^\top - \boldsymbol{x}_i\boldsymbol{x}_j^\top - \boldsymbol{x}_j\boldsymbol{x}_i^\top),
\end{aligned} \tag{20}$$

which yields Eq.(16). Let $\boldsymbol{S}^{(m)}$ be the *mixture scatter matrix* (Fukunaga, 1990):

$$\begin{aligned}
\boldsymbol{S}^{(m)} &= \boldsymbol{S}^{(w)} + \boldsymbol{S}^{(b)} \\
&= \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top.
\end{aligned} \tag{21}$$

Then we have

$$\begin{aligned}
\boldsymbol{S}^{(b)} &= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{n}\sum_{i,j=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_j^\top - \boldsymbol{S}^{(w)} \\
&= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{1}{n}\right)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sum_{i,j=1}^{n}\frac{1}{n}\boldsymbol{x}_i\boldsymbol{x}_j^\top - \boldsymbol{S}^{(w)} \\
&= \frac{1}{2}\sum_{i,j=1}^{n}\left(\frac{1}{n} - \boldsymbol{A}_{i,j}^{(w)}\right) \\
&\qquad \times (\boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{x}_j\boldsymbol{x}_j^\top - \boldsymbol{x}_i\boldsymbol{x}_j^\top - \boldsymbol{x}_j\boldsymbol{x}_i^\top),
\end{aligned} \tag{22}$$

which yields Eq.(17). ■

Note that $1/n - 1/n_c$ is negative while $1/n_c$ and $1/n$ are positive.

The above pairwise representation gives us a new interpretation of FDA: it tries to keep in-class data pairs close (since $\boldsymbol{A}_{i,j}^{(w)}$ is positive and $\boldsymbol{A}_{i,j}^{(b)}$ is negative if $y_i = y_j$) and between-class data pairs apart (since $\boldsymbol{A}_{i,j}^{(b)}$ is positive if $y_i \neq y_j$).

### 3.2. Definition of LFDA

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

$$\overline{\boldsymbol{S}}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(w)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \tag{23}$$

$$\overline{\boldsymbol{S}}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(b)}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \tag{24}$$

where

$$\overline{\boldsymbol{A}}_{i,j}^{(w)} = \begin{cases} \boldsymbol{A}_{i,j}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{25}$$

$$\overline{\boldsymbol{A}}_{i,j}^{(b)} = \begin{cases} \boldsymbol{A}_{i,j}(1/n - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \tag{26}$$

Compared with the global counterparts $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$, the values for in-class pairs are weighted by the affinity in $\overline{\boldsymbol{S}}^{(w)}$ and $\overline{\boldsymbol{S}}^{(b)}$. This means that far-apart in-class pairs have less influence in $\overline{\boldsymbol{S}}^{(w)}$ and $\overline{\boldsymbol{S}}^{(b)}$.

Using $\overline{\boldsymbol{S}}^{(w)}$ and $\overline{\boldsymbol{S}}^{(b)}$, we define the LFDA transformation matrix $\boldsymbol{T}_{LFDA}$ as

$$\boldsymbol{T}_{LFDA} = \operatorname*{argmax}_{\boldsymbol{T}\in\mathbb{R}^{d\times m}} \operatorname{tr}\left((\boldsymbol{T}^\top\overline{\boldsymbol{S}}^{(w)}\boldsymbol{T})^{-1}\boldsymbol{T}^\top\overline{\boldsymbol{S}}^{(b)}\boldsymbol{T}\right). \tag{27}$$

That is, we determine $\boldsymbol{T}$ so that *nearby* data pairs of the *same class* are close and data pairs of different

classes are apart. Data pairs of the same class but far apart are not imposed to be close.

If the affinity matrix $\boldsymbol{A}$ is taken to be one for all in-class pairs (i.e., all in-class pairs are 'equally close' each other), $\overline{\boldsymbol{S}}^{(w)}$ and $\overline{\boldsymbol{S}}^{(b)}$ are reduced to $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$ so LFDA is reduced to the ordinary FDA. Therefore, LFDA may be regarded as a natural localized variant of FDA.

Since Eq.(27) is the same form as Eq.(8), we can easily obtain $\boldsymbol{T}_{LFDA}$ by solving a generalized eigenvalue problem. In Section 3.4, we show an efficient method for computing $\boldsymbol{T}_{LFDA}$.

Examples of dimensionality reduction by LFDA are illustrated in Figure 1, where the affinity matrix $\boldsymbol{A}$ is determined by the same way as being done for LPP. The figures show that LFDA gives desirable results for all three data sets, i.e., LFDA can compensate the drawbacks of FDA and LPP.

### 3.3. Why is LFDA good?

LFDA does not impose far-apart data pairs of the same class to be close, by which local structure of the data tends to be preserved. This itself is a useful property, e.g., in data visualization tasks because 'interesting' structure such as multimodality is not lost by dimensionality reduction.

In addition to that, LFDA can even provide more separate embedding than FDA. To explain the reason, let us recall that the FDA transformation matrix $\boldsymbol{T}_{FDA}$ can also be obtained as follows (Fukunaga, 1990).

$$\boldsymbol{T}_{FDA} = \mathop{\mathrm{argmax}}_{\boldsymbol{T}\in\mathbb{R}^{d\times m}} \mathrm{tr}\left(\boldsymbol{T}^\top \boldsymbol{S}^{(b)} \boldsymbol{T}\right) \quad \text{subject to } \boldsymbol{T}^\top \boldsymbol{S}^{(w)} \boldsymbol{T} = \boldsymbol{I}. \tag{28}$$

This representation implies that FDA maximizes between-class scatter under the constraint of keeping within-class scatter to a certain level.

When one of the classes is multimodal, the above constraint is actually quite restrictive since the multimodal data should be typically merged into a single cluster. Therefore, the remaining degree of freedom which can be used for maximizing between-class scatter may not be that much. As a result, FDA can result in less separate embedding.

On the other hand, it is straightforward to show that $\boldsymbol{T}_{LFDA}$ can also be expressed as follows.

$$\boldsymbol{T}_{LFDA} = \mathop{\mathrm{argmax}}_{\boldsymbol{T}\in\mathbb{R}^{d\times m}} \mathrm{tr}\left(\boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(b)} \boldsymbol{T}\right) \quad \text{subject to } \boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(w)} \boldsymbol{T} = \boldsymbol{I}. \tag{29}$$

The constraint in Eq.(29) is less restrictive than that in Eq.(28) since faraway pairs of in-class samples are not imposed to be close. Due to this weaker constraint, the degree of freedom which can be used for maximizing separability between different classes remains more than FDA. Thus, LFDA can result in more separate embedding.

### 3.4. Efficiently Computing LFDA Transformation Matrix

Here, we provide an efficient method to compute the LFDA transformation matrix $\boldsymbol{T}_{LFDA}$.

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

$$\begin{aligned}\overline{\boldsymbol{S}}^{(m)} &= \overline{\boldsymbol{S}}^{(w)} + \overline{\boldsymbol{S}}^{(b)} \\ &= \frac{1}{2}\sum_{i,j=1}^{n} \overline{\boldsymbol{A}}_{i,j}^{(m)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top,\end{aligned} \tag{30}$$

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

$$\overline{\boldsymbol{A}}_{i,j}^{(m)} = \overline{\boldsymbol{A}}_{i,j}^{(w)} + \overline{\boldsymbol{A}}_{i,j}^{(b)} = \begin{cases} \boldsymbol{A}_{i,j}/n & \text{if } y_i = y_j, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \tag{31}$$

Then the same solution $\boldsymbol{T}_{LFDA}$ can be still obtained if $\overline{\boldsymbol{S}}^{(b)}$ in Eq.(27) is replaced by $\overline{\boldsymbol{S}}^{(m)}$ because of the following identity (cf. (Fukunaga, 1990)):

$$\begin{aligned}&\mathrm{tr}\left((\boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(w)} \boldsymbol{T})^{-1} \boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(m)} \boldsymbol{T}\right) \\ &\quad = \mathrm{tr}\left((\boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(w)} \boldsymbol{T})^{-1} \boldsymbol{T}^\top \overline{\boldsymbol{S}}^{(b)} \boldsymbol{T}\right) + m.\end{aligned} \tag{32}$$

In the following, we employ the criterion with $\overline{\boldsymbol{S}}^{(m)}$ since it is simpler.

The left-hand side of Eq.(32) is the same form as FDA. Therefore, the solution $\boldsymbol{T}_{LFDA}$ is analytically given as follows.

$$\boldsymbol{T}_{LFDA} = (\overline{\boldsymbol{\varphi}}_1 | \overline{\boldsymbol{\varphi}}_2 | \cdots | \overline{\boldsymbol{\varphi}}_m), \tag{33}$$

where $\{\overline{\boldsymbol{\varphi}}_i\}_{i=1}^d$ are the generalized eigenvectors associated to the generalized eigenvalues $\overline{\lambda}_1 \geq \overline{\lambda}_2 \geq \cdots \geq \overline{\lambda}_d$ of the following generalized eigenvalue problem:

$$\overline{\boldsymbol{S}}^{(m)} \overline{\boldsymbol{\varphi}} = \overline{\lambda}\, \overline{\boldsymbol{S}}^{(w)} \overline{\boldsymbol{\varphi}}. \tag{34}$$

Below, we show how $\overline{\boldsymbol{S}}^{(m)}$ and $\overline{\boldsymbol{S}}^{(w)}$ can be computed efficiently.

Since

$$\begin{aligned}\overline{\boldsymbol{S}}^{(m)} &= \frac{1}{2}\sum_{i,j=1}^{n} \overline{\boldsymbol{A}}_{i,j}^{(m)}(\boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{x}_j\boldsymbol{x}_j^\top - \boldsymbol{x}_i\boldsymbol{x}_j^\top - \boldsymbol{x}_j\boldsymbol{x}_i^\top) \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(m)}\right)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sum_{i,j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(m)}\boldsymbol{x}_i\boldsymbol{x}_j^\top,\end{aligned} \tag{35}$$

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

$$\overline{\boldsymbol{S}}^{(m)} = \boldsymbol{X}\overline{\boldsymbol{L}}^{(m)}\boldsymbol{X}^\top, \tag{36}$$

where $\overline{\boldsymbol{L}}^{(m)} = \overline{\boldsymbol{D}}^{(m)} - \overline{\boldsymbol{A}}^{(m)}$ and $\overline{\boldsymbol{D}}^{(m)}$ is the $n$-dimensional diagonal matrix with $i$-th diagonal element being $\overline{\boldsymbol{D}}_{i,i}^{(m)} = \sum_{j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(m)}$. Similarly, $\overline{\boldsymbol{S}}^{(w)}$ can be expressed in a matrix form as

$$\overline{\boldsymbol{S}}^{(w)} = \boldsymbol{X}\overline{\boldsymbol{L}}^{(w)}\boldsymbol{X}^\top, \tag{37}$$

where $\overline{\boldsymbol{L}}^{(w)} = \overline{\boldsymbol{D}}^{(w)} - \overline{\boldsymbol{A}}^{(w)}$ and $\overline{\boldsymbol{D}}^{(w)}$ is the $n$-dimensional diagonal matrix with $i$-th diagonal element being $\overline{\boldsymbol{D}}_{i,i}^{(w)} = \sum_{j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(w)}$.

$\overline{\boldsymbol{L}}^{(m)}$ and $\overline{\boldsymbol{L}}^{(w)}$ are $n$-dimensional matrices so could be very high dimensional. However, $\overline{\boldsymbol{L}}^{(w)}$ could be *sparse* so computing $\overline{\boldsymbol{S}}^{(w)}$ directly by Eq.(37) may be efficient. On the other hand, computing $\overline{\boldsymbol{S}}^{(m)}$ directly by Eq.(36) is not so efficient since $\overline{\boldsymbol{A}}^{(m)}$ is *dense*.

This problem can be alleviated as follows. $\overline{\boldsymbol{A}}^{(m)}$ can be decomposed as

$$\overline{\boldsymbol{A}}^{(m)} = \mathbf{1}\mathbf{1}^\top/n + \overline{\boldsymbol{A}}^{(m')}, \tag{38}$$

where $\mathbf{1}$ is the $n$-dimensional vector with all ones and $\overline{\boldsymbol{A}}^{(m')}$ is the $n$-dimensional matrix with $(i,j)$-th element being

$$\overline{\boldsymbol{A}}_{i,j}^{(m')} = \begin{cases} (\boldsymbol{A}_{i,j} - 1)/n & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j. \end{cases} \tag{39}$$

Then $\overline{\boldsymbol{S}}^{(m)}$ can be expressed as follows:

$$\overline{\boldsymbol{S}}^{(m)} = \boldsymbol{X}\overline{\boldsymbol{D}}^{(m)}\boldsymbol{X}^\top - (\boldsymbol{X}\mathbf{1})(\boldsymbol{X}\mathbf{1})^\top/n - \boldsymbol{X}\overline{\boldsymbol{A}}^{(m')}\boldsymbol{X}^\top, \tag{40}$$

where $\overline{\boldsymbol{D}}^{(m)}$ is expressed by using $\overline{\boldsymbol{A}}^{(m')}$ as

$$\overline{\boldsymbol{D}}_{i,i}^{(m)} = 1 + \sum_{j=1}^{n}\overline{\boldsymbol{A}}_{i,j}^{(m')}. \tag{41}$$

$\overline{\boldsymbol{A}}^{(m')}$ becomes a block-diagonal matrix if $\{\boldsymbol{x}_i\}_{i=1}^n$ are sorted by the labels, which implies that the third term in the right-hand side of Eq.(40) may be computed efficiently. Since the first two terms in the right-hand side of Eq.(40) can also be computed efficiently, computing $\overline{\boldsymbol{S}}^{(m)}$ by Eq.(40) may be more efficient than directly by Eq.(36).

To further improve computational efficiency, the affinity matrix $\boldsymbol{A}$ may be computed in a classwise manner, i.e., the between-class affinity is set to be zero.

### 3.5. Kernel LFDA for Non-Linear Dimensionality Reduction

So far, we focused on linear dimensionality reduction. Here we extend our discussion to non-linear dimensionality reduction scenarios.

From Eqs.(36) and (37), the generalized eigenvalue problem (34) can be expressed as

$$\boldsymbol{X}\overline{\boldsymbol{L}}^{(m)}\boldsymbol{X}^\top\overline{\boldsymbol{\varphi}} = \overline{\lambda}\boldsymbol{X}\overline{\boldsymbol{L}}^{(w)}\boldsymbol{X}^\top\overline{\boldsymbol{\varphi}}. \tag{42}$$

When $d \leq n$, any vector $\overline{\boldsymbol{\varphi}} \in \mathbb{R}^d$ can be expressed by using some vector $\overline{\boldsymbol{\alpha}} \in \mathbb{R}^n$ as $\overline{\boldsymbol{\varphi}} = \boldsymbol{X}\overline{\boldsymbol{\alpha}}$. Then, multiplying Eq.(42) by $\boldsymbol{X}^\top$ from the left-hand side, we have

$$\boldsymbol{K}\overline{\boldsymbol{L}}^{(m)}\boldsymbol{K}\overline{\boldsymbol{\alpha}} = \overline{\lambda}\boldsymbol{K}\overline{\boldsymbol{L}}^{(w)}\boldsymbol{K}\overline{\boldsymbol{\alpha}}, \tag{43}$$

where $\boldsymbol{K}$ is the $n$-dimensional matrix with the $(i,j)$-th element being

$$\boldsymbol{K}_{i,j} = \boldsymbol{x}_i^\top\boldsymbol{x}_j. \tag{44}$$

This implies that $\{\boldsymbol{x}_i\}_{i=1}^n$ appear only in terms of their inner products, so we can non-linearize the algorithm by the *kernel trick* (Schölkopf & Smola, 2002), which is briefly explained below.

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

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/2\sigma^2\right). \tag{45}$$

For other choices, see, e.g., (Schölkopf & Smola, 2002). Because of the reproducing property of $K(\boldsymbol{x},\boldsymbol{x}')$, $\boldsymbol{K}$ is now the *kernel matrix*, i.e., the $(i,j)$-th element is defined as

$$\boldsymbol{K}_{i,j} = \langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle = K(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{46}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathcal{H}$. Then the embedded image of $\phi(\boldsymbol{x}')$ by LFDA in $\mathcal{H}$ is given by

$$(\overline{\boldsymbol{\alpha}}_1|\overline{\boldsymbol{\alpha}}_2|\cdots|\overline{\boldsymbol{\alpha}}_m)^\top\begin{pmatrix}K(\boldsymbol{x}_1,\boldsymbol{x}')\\K(\boldsymbol{x}_2,\boldsymbol{x}')\\\vdots\\K(\boldsymbol{x}_n,\boldsymbol{x}')\end{pmatrix}, \tag{47}$$

where $\{\overline{\boldsymbol{\alpha}}_i\}_{i=1}^d$ is the generalized eigenvectors associated to the generalized eigenvalues $\overline{\lambda}_1 \geq \overline{\lambda}_2 \geq \cdots \geq \overline{\lambda}_d$ of the generalized eigenvalue problem (43). Note that $\boldsymbol{K}\overline{\boldsymbol{L}}^{(m)}\boldsymbol{K}$ can be efficiently calculated by

$$\boldsymbol{K}\overline{\boldsymbol{L}}^{(m)}\boldsymbol{K} = \boldsymbol{K}\overline{\boldsymbol{D}}^{(m)}\boldsymbol{K} - (\boldsymbol{K}\boldsymbol{1})(\boldsymbol{K}\boldsymbol{1})^\top/n - \boldsymbol{K}\overline{\boldsymbol{A}}^{(m')}\boldsymbol{K}. \quad (48)$$

## 4. Numerical Results

In this section, we apply the existing and proposed dimensionality reduction methods to benchmark data sets for visualization. In LPP and LFDA, the affinity matrix is determined by the same method used in Section 2.4.

We use *Iris*, *Letter recognition*, *Segment*, and *Thyroid disease* data sets available from the UCI machine learning repository. They are all multi-class classification data sets. We created two-class problems in the standard 'one-versus-rest' manner. Figure 2 shows the embedded samples in two-dimensional space by FDA, LFDA, and LPP.

For Iris data set, FDA tends to mix samples of different classes, while LFDA separates them very well. Multimodality of the '×'-class can be clearly observed in the results of LFDA, which implies that FDA did not work well because of multimodality. This experimental result supports the qualitative justification of LFDA given in Section 3.3. LPP also works well for this data set since three clusters (two '×'-class clusters and one '∘'-class cluster) are well separated from each other in the original space.

For the other three data sets, FDA can separate samples of different classes quite well. LFDA also separates them equally well, but moreover it preserves multimodality of the '×'-class more prominently than FDA. This would be a desirable property in data visualization tasks. LPP also preserves multimodality of the data well, but it tends to mix samples of different classes.

Overall, LFDA is found to be useful in data visualization tasks.

## 5. Conclusions

Dimensionality reduction based on the Fisher criterion (FDA) works well given data samples of each class form a single cluster (i.e., unimodal). On the other hand, samples in some class can be multimodal, e.g., when multi-class classification problems are solved by a set of two-class 'one-versus-rest' problems. In this paper, we first showed that FDA can be regarded as keeping in-class data pairs close and between-class data pairs apart. Based on this novel interpretation, we proposed a localized variant of FDA called local Fisher discriminant analysis (LFDA). We experimentally showed that LFDA can preserve multimodal structure of the data better than FDA, and LFDA can even provide more separate embedding than FDA.

Our future work includes the comparison with recently proposed methods, e.g, (Goldberger et al., 2005; Globerson & Roweis, 2006).

## Acknowledgments

The author would like to thank Yuki Shinada, Klaus-Robert Müller, Hideki Asoh, Stefan Harmeling and anonymous reviewers for their comments. He also acknowledges financial support from MEXT (Grant-in-Aid for Young Scientists 17700142).

## References

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

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

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Boston: Academic Press, Inc. Second edition.

Globerson, A., & Roweis, S. (2006). Metric learning by collapsing classes. *Advances in Neural Information Processing Systems 18* (pp. 451–458). Cambridge, MA: MIT Press.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood components analysis. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 513–520. Cambridge, MA: MIT Press.

He, X., & Niyogi, P. (2004). Locality preserving projections. In S. Thrun, L. Saul and B. Schölkopf (Eds.), *Advances in neural information processing systems 16*. Cambridge, MA: MIT Press.

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

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 1601–1608. Cambridge, MA: MIT Press.

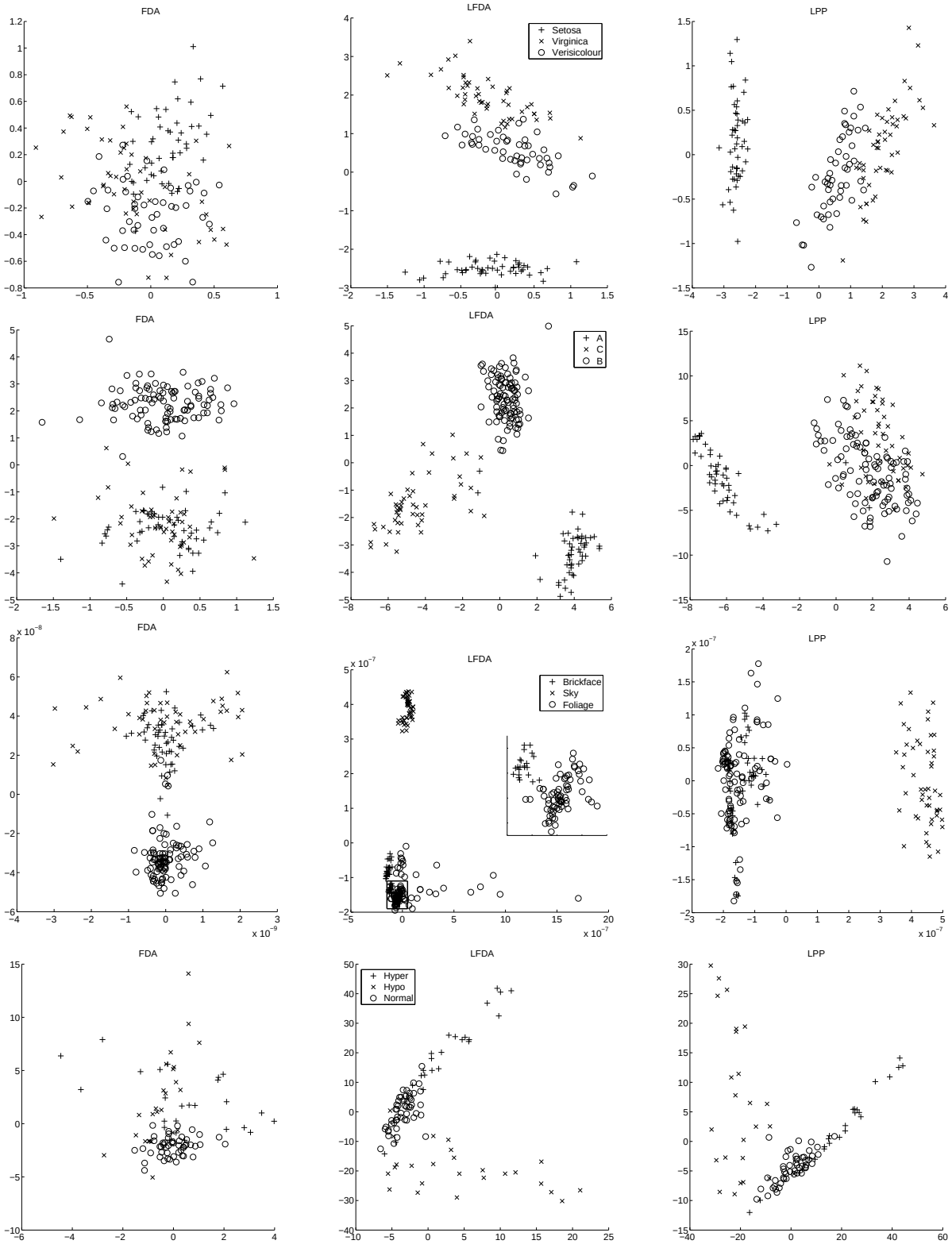


*Figure 2.* Results of data visualization ($m = 2$). From top to bottom, Iris data set ($d = 4$: 'Setosa' & 'Virginica' vs. 'Versicolour'), Letter recognition data set ($d = 16$: 'A' & 'C' vs. 'B'), Segment data set ($d = 18$: 'Brickface' & 'Sky' vs. 'Foliage'), and Thyroid disease data set ($d = 5$: 'Hyper' & 'Hypo' vs. 'Normal'). From left to right, FDA, LFDA, and LPP.