# Geometric interpretation of Fisher's linear discriminant analysis through communication theory

Jun FUJIKI and Masaru TANAKA
Fukuoka University
8-19-1, Nanakuma, Jonan-ku, Fukuoka 814-0180
fujiki@fukuoka-u.ac.jp, sieger@math.sci.fukuoka-u.ac.jp

Hitoshi SAKANO
NTT Data
3-3-9, Toyosu, Koutou-ku, Tokyo 135-8671
sakanoh@nttdata.co.jp

Akisato KIMURA
NTT Communication Science Laboratories
3-1, Morinosato Wakamiya Atsugi-shi, Kanagawa 243-0198
akisato@ieee.org

## Abstract

*This paper provides a geometrical aspect of Fisher's linear discriminant analysis (FLDA), which has been widely used owing to its simple formulation and low computational costs. Our approach is based on a new framework of pattern recognition that can be modelded by a communication of class information. This model is quite different from a commonly used framework of pattern recognition as a mapping from the set of patterns to the set of classes. In the new framework, patterns can be regarded as class information with redundant encoding. We show that the geometry of two class FLDA can be described via communication theory of noisy channel.*

## 1 Introduction

Fisher's linear discriminant analysis (FLDA) [7] has been widely used as a discriminative feature extractor in the fields of pattern recognition, computer vision and machine learning [3, 10] for a long time owing to its simple formulation and low computational costs. Lots of extensions and modifications of FLDA have ever been proposed [2, 5, 8, 9, 11, 12, 13, 14, 15], and these FLDA-based methods suggest that FLDA is a fundamental and important method in pattern recognition. Therefore, to understand geometries of pattern recognition, understanding a geometry of FLDA is very important. Then, this paper provides a geometrical aspect of two class FLDA by a new framework of pattern recognition.

Our new framework of pattern recognition is inspired by the communication theory [1, 4]: We regard pattern recognition as a communication of class information on a noisy channel, in contrast to a common framework of pattern recognition as a mapping from a pattern space consisting of spatial/temporal data such as images and sounds, into class space. In the framework, pattern recognition is regarded as a mapping from a pattern into what a pattern represents that belongs to some discrete class. Usually, the dimension of pattern space is very large, mean while the dimension of class space is small. Therefore, pattern recognition as a compressive mapping from redundant pattern space into concise class space can be viewed as dimensional reduction and/or information compression. In the context of pattern recognition, the mapping from pattern space to class space should be carefully constructed to achieve high recognition rate. This mapping is generally decomposed of two steps: one is from a pattern space into a feature space, and the other from a feature space into a class space, which is called "feature extraction" in general.

This paper gives a new framework of pattern recognition from the point of view of communication theory. In a common, framework of pattern recognition, a pattern space is the basis of theoretical analysis, meanwhile in our framework, a class space plays a central role in the analysis. Every class as an element of a class space is given a priori, and a pattern can be regarded as class information with redundant encoding. Pattern recognition is a process of decoding observed patterns with channel noises into class labels by projecting it to small dimensional space (Fig. 1). The image of a pattern should have discrete values, however, noise contamination makes it continuously. Then, each label is embedded to a continuous space and the value of continuous label is regarded as a label likelihood. This leads to a new framework of pattern recognition, such that pattern recognition can be regarded as communication of the class label, then, pattern recognition can be understood in the framework of the communication theory.

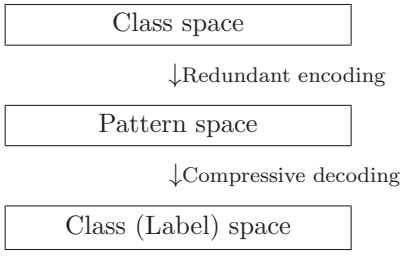


Figure 1. Proposed pattern recognition process.

Making use of the above idea, we can understand two class FLDA from the point of the view of information geometry.

## 2 Fisher's linear discriminant analysis

In this section, we explain FLDA, which is a linear method of supervised learning. FLDA projects data into a small dimensional space by emphasizing the separability of classes, that is, the data belonging to the same class would be located near and the data belonging to the different classes would be located far.

Assuming that $N$ data are classified to $C$ classes. Let $n_c$ be the number of data classified to the class $c$, and they are named $\boldsymbol{x}_1^{(c)}, \ldots, \boldsymbol{x}_{n_c}^{(c)}$ ($N = \sum_{c=1}^{C} n_c$). Then, the mean and the variance-covariance matrix of the class $c$ be

$$\boldsymbol{\mu}^{(c)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \boldsymbol{x}_i^{(c)} ,$$
$$\Sigma_{\mathrm{W}}^{(c)} = \frac{1}{n_c} \sum_{i=1}^{n_c} \left( \boldsymbol{x}_i^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left( \boldsymbol{x}_i^{(c)} - \boldsymbol{\mu}^{(c)} \right)^{\top} ,$$

respectively. Here, $\boldsymbol{\mu}^{(c)}$ and $\Sigma_{\mathrm{W}}^{(c)}$ are called a *class mean* and a *within class variance*, respectively.

Let the mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\Sigma_{\mathrm{T}}$ of all data be

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{n_c} \boldsymbol{x}_i^{(c)} = \frac{1}{N} \sum_{c=1}^{C} n_c \boldsymbol{\mu}^{(c)} ,$$
$$\Sigma_{\mathrm{T}} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{n_c} \left( \boldsymbol{x}_i^{(c)} - \boldsymbol{\mu} \right) \left( \boldsymbol{x}_i^{(c)} - \boldsymbol{\mu} \right)^{\top} ,$$

respectively. Let $\Sigma_{\mathrm{W}}$, called a *within-class variance*, be the average of the variance-covariance matrix of every classes, and let $\Sigma_{\mathrm{B}}$, called a *between-class variance*, be the variance-covariance matrix when all data is assumed to concentrate to the mean of a corresponding class. These are represented by

$$\Sigma_{\mathrm{W}} = \frac{1}{N} \sum_{c=1}^{C} n_c \Sigma_{\mathrm{W}}^{(c)}, \tag{1}$$
$$\Sigma_{\mathrm{B}} = \frac{1}{N} \sum_{c=1}^{C} n_c \left( \boldsymbol{\mu}^{(c)} - \boldsymbol{\mu} \right) \left( \boldsymbol{\mu}^{(c)} - \boldsymbol{\mu} \right)^{\top} ,$$

respectively. Then, there holds $\Sigma_{\mathrm{T}} = \Sigma_{\mathrm{W}} + \Sigma_{\mathrm{B}}$.

We consider two class FLDA. In this case, data are projected (compressed) to a 1-dimensional linear space spanned by $\boldsymbol{w} \neq \boldsymbol{0}$. For projected data, there holds

$$\boldsymbol{w}^{\top} \Sigma_{\mathrm{T}} \boldsymbol{w} = \boldsymbol{w}^{\top} \Sigma_{\mathrm{W}} \boldsymbol{w} + \boldsymbol{w}^{\top} \Sigma_{\mathrm{B}} \boldsymbol{w} .$$

In the discrimination, a between-class variance is more important than a within-class variance. Then, $\boldsymbol{w}$ is chosen to maximize the ratio $\frac{\boldsymbol{w}^{\top} \Sigma_{\mathrm{B}} \boldsymbol{w}}{\boldsymbol{w}^{\top} \Sigma_{\mathrm{W}} \boldsymbol{w}}$, and $\boldsymbol{w}$ is explicitly given in the form

$$\boldsymbol{w} = \Sigma_{\mathrm{W}}^{-1} \left( \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} \right) , \tag{2}$$

by solving a general eigenvalue problem as the eigenvector corresponding to the largest eigenvalue of the matrix $\Sigma_{\mathrm{W}}^{-1} \Sigma_{\mathrm{B}}$.

## 3 Overview of communication theory

In this section, we overview the communication theory [1, 4].

### 3.1 Entropy and transinformation

Let $\boldsymbol{X}$ be a random variable of a signal $\boldsymbol{x}$ generated by a probability density function (pdf) $p_{\boldsymbol{X}}(\boldsymbol{x})$. The ambiguity of signal is measured by the entropy of $\boldsymbol{X}$ as $H(\boldsymbol{X}) = -\int p_{\boldsymbol{X}}(\boldsymbol{x}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}$.

For two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ corresponding to signals $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, the joint entoropy of the pair $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the conditional entropy of $\boldsymbol{Y}$ given $\boldsymbol{X}$ are defined by
$H(\boldsymbol{X}, \boldsymbol{Y}) = -\int p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) \log p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$, and
$H(\boldsymbol{Y}|\boldsymbol{X}) = -\int p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) \log p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y}$, respectively, where the pair $\boldsymbol{x}$ and $\boldsymbol{y}$ are generated by the pdf $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$, and $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})}$ be the conditional pdf of $\boldsymbol{y}$ given $\boldsymbol{x}$. Let $I(\boldsymbol{X}, \boldsymbol{Y})$, called a *mutual information* of $\boldsymbol{X}$ and $\boldsymbol{Y}$, be defined by the difference of entropies (signal ambiguity) as $H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y})$. The mutual information gives the common information between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and it equals 0 when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent one another.

Let a signal $\boldsymbol{y}$ be a transmitted signal of $\boldsymbol{x}$ contaminated by a noise $\boldsymbol{n}$ through continuous channel (the channel transmits continuous signals), that is, there holds $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}$. Because of the noise, $\boldsymbol{x}$ cannot be exactly restored by $\boldsymbol{y}$. Let $\boldsymbol{X}$ and $\boldsymbol{N}$ be random variables of a signal $\boldsymbol{x}$ and a noise $\boldsymbol{n}$, respectively. We assume $\boldsymbol{X}$ and $\boldsymbol{N}$ are independent. The loss of information due to noise is $H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{N})$, then, transmitted information of $\boldsymbol{X}$ equals to the mutual information $J = I(\boldsymbol{X}, \boldsymbol{Y})$. Then, $J$ is also called *transinformation* from the point of view of data transmission.

### 3.2 Riemannian signal space

Let a noise vector $\boldsymbol{n}$ be an additive Gaussian noise of mean $\boldsymbol{0}$ depends on the signal $\boldsymbol{x}$, that is, the conditional pdf of $\boldsymbol{n}$ given $\boldsymbol{x}$ be

$$p_{\boldsymbol{N}|\boldsymbol{X}}(\boldsymbol{n}|\boldsymbol{x}) = A(\boldsymbol{x}) \exp\left\{ -\frac{1}{2} \boldsymbol{n}^{\top} V^{-1}(\boldsymbol{x}) \boldsymbol{n} \right\} \tag{3}$$

be holds, where $V(\boldsymbol{x})$ be the variance-covariance matrix of $\boldsymbol{n}$ given $\boldsymbol{x}$, and $A(\boldsymbol{x}) = (2\pi)^{-n/2} (\mathtt{det}\, V(\boldsymbol{x}))^{-1/2}$ be a normalization factor.

By the noise, signals are shifted, and the way of shift of signals can be represented by the ellipsoid $\boldsymbol{n}^{\top} V^{-1}(\boldsymbol{x}) \boldsymbol{n} = \text{const.}$, which corresponds to Gaussian noise as large (small) noise is incident to the direction along long (short) axis. The metric in the signal space is defined by the characteristic of the signal shift as the square of an infinitesimal distance between two signals $\boldsymbol{x}$ and $\boldsymbol{x}' = \boldsymbol{x} + d\boldsymbol{x}$ is represented by $ds^2 = \{d(\boldsymbol{x}, \boldsymbol{x}')\}^2 = d\boldsymbol{x}^{\top} G(\boldsymbol{x}) d\boldsymbol{x}$, where $G(\boldsymbol{x}) = \frac{1}{n} V^{-1}(\boldsymbol{x})$ and $n$ is the dimension of the signal space. The matrix $G(\boldsymbol{x})$ is called a *Gramian matrix*. It is called a Riemannian space that the space where the Gramian matrix is defined at all points. When the signal space is a Riemannian space, the space is called a *Riemannian signal space*.

By the Gramian matrix, the square-length of the noise vector $\boldsymbol{n}$ is $\boldsymbol{n}^\top G(\boldsymbol{x})\boldsymbol{n}$. Therefore, the average of the square-length is $\overline{\boldsymbol{n}^\top G(\boldsymbol{x})\boldsymbol{n}} = \mathtt{tr}[G(\boldsymbol{x})V(\boldsymbol{x})] = \frac{1}{n}\mathtt{tr}[\mathtt{I}] = 1$, where $\mathtt{I}$ is the identity matrix. Then, the metric is defined so that the average of the noise-length equals to 1. When the dimension of the signal space is sufficiently large, the noise-length is 1 for almost all noise because of low of large number. That is, almost all signals are shifted to be on the unit hypersphere of center $\boldsymbol{x}$. The unit hypersphere is called a *noise hypersphere*. Note that the shape of a noise hypersphere is depend on the signal $\boldsymbol{x}$.

When Gaussian noise vector $\boldsymbol{n}$ is sufficiently small, by neglecting higher order term, $H(\boldsymbol{X}|\boldsymbol{Y})$ is approximated by $H(\boldsymbol{N}|\boldsymbol{X})$ [1], transfinformation is approximated by $J = H(\boldsymbol{X}) - H(\boldsymbol{N}|\boldsymbol{X})$.

Let $g(\boldsymbol{x})$ be $\mathtt{det}\, G(\boldsymbol{x})$, there holds

$$p_{\boldsymbol{N}|\boldsymbol{X}}(\boldsymbol{n}|\boldsymbol{x}) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \sqrt{g(\boldsymbol{x})} \exp\left\{-\frac{n}{2}\boldsymbol{n}^\top G(\boldsymbol{x})\boldsymbol{n}\right\},$$

and therefore the entropy of noise at $\boldsymbol{x}$ in signal space is computed by using Eq. (3) as

$$H(\boldsymbol{N}|\boldsymbol{x}) = -\log\left(b_n\sqrt{g(\boldsymbol{x})}\right)$$

where $b_n = \left(\frac{n}{2\pi e}\right)^{\frac{n}{2}}$. Then, the average transinformation of the signal space is

$$J = H(\boldsymbol{X}) - H(\boldsymbol{N}|\boldsymbol{X}) = \int p_{\boldsymbol{X}}(\boldsymbol{x}) \log \frac{b_n\sqrt{g(\boldsymbol{x})}}{p_{\boldsymbol{X}}(\boldsymbol{x})} d\boldsymbol{x}\,.$$

Because the average transinformation depends on $p_{\boldsymbol{X}}(\boldsymbol{x})$, we control $p_{\boldsymbol{X}}(\boldsymbol{x})$ to maximize the average transinformation in order to realize the best signal transmission. The maximized average transinformation satisfies the condition $I = \max_{p_{\boldsymbol{X}}(\boldsymbol{x})} J$, and the value $I$ is called a *channel capacity* of the signal space. By Lagrange's multiplier method, the channel capacity can be exactly obtained as

$$I = \log(b_n U)$$

where $U = \int \sqrt{g(\boldsymbol{x})}d\boldsymbol{x}$ at $p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{U}\sqrt{g(\boldsymbol{x})}$. Because the volume of the noise hypersphere is $\frac{\pi}{\sqrt{g(\boldsymbol{x})}}$, $\sqrt{g(\boldsymbol{x})}$ is inversely proportional to the volume of the noise hypersphere, that is, proportional to the number of the noise hypersphere packed in the neighborhood of $\boldsymbol{x}$. $U$ is the average of $\sqrt{g(\boldsymbol{x})}$ over the signal space. Therefore, $b_n U$ represents the number of the noise hypersphere packed in the signal space, and then, the same number of signals can be correctly discriminated. Consequently, the channel capacity is equivalent to the logarithm of the number of the noise hypersphere.

### 3.3 Imbedding/reduction mapping

In this subsection, we explain a continuous channel as a sequential mapping in the signal space (Fig. 2). At First, signals in a Riemannian signal space $S^n$ is redundantly encoded as elements in a high dimensional channel signal space $S^m$. The elements in $S^m$ are contaminated by noises. After that, noise-contaminated signals in $S^m$ are decoded as elements in $S^n$. For the sake of clearity, we note $\overline{S^n}$ as the Riemannian signal space consists of decoded elements.

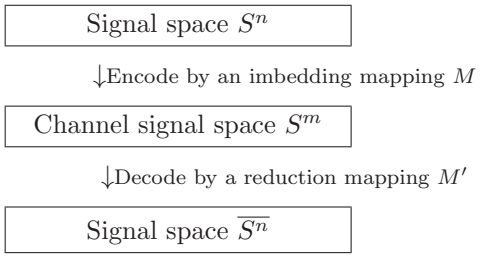


Figure 2. Channel model.

To utilize the channel efficiently, the best imbedded mapping $M : S^n \to S^m$, and the best reduction mapping $M' : S^m \to \overline{S^n}$, should be chosen carefully.

Let $\boldsymbol{f}$ be a mapping from $\boldsymbol{x} \in S^n$, into $\boldsymbol{y} \in S^m$. The mapping $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ is called an *imbedding mapping*. The mapping $\boldsymbol{f}(\boldsymbol{x})$ is locally linearized by $d\boldsymbol{y} = T(\boldsymbol{x})d\boldsymbol{x}$, where $T(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$. Here, $T$ is called a *tangent matrix* because the set of $T$'s columns is the basis of the tangent space of $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$.

On the other hand, the signal $\boldsymbol{y}$ is inversely transformed to signal $\boldsymbol{x} \in \overline{S^n}$. Let $\boldsymbol{h}$ be a mapping from $\boldsymbol{y}$ into $\boldsymbol{x}$. The mapping $\boldsymbol{x} = \boldsymbol{h}(\boldsymbol{y})$ is called a *reduction mapping*. The mapping $\boldsymbol{h}(\boldsymbol{x})$ is locally linearized by $d\boldsymbol{x} = R(\boldsymbol{y})d\boldsymbol{y}$, where $R(\boldsymbol{y}) = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}}$. Here, $R$ is called a *reduction matrix*. Note that $R(\boldsymbol{x})T(\boldsymbol{x}) = \mathtt{I}$ holds because the decoded signals should be the same as the original.

Let a noise $\boldsymbol{m}$ in $S^m$ be sufficiently small, then $\boldsymbol{h}(\boldsymbol{y}+\boldsymbol{m}) = \boldsymbol{x}+R(\boldsymbol{x})\boldsymbol{m}$ holds. Then, the noise in $\overline{S^n}$ is represented by $\boldsymbol{n} = R(\boldsymbol{x})\boldsymbol{m}$. Then, if the noise satisfies $R(\boldsymbol{x})\boldsymbol{m} = \boldsymbol{0}$, $\boldsymbol{x}$ is correctly decoded by the reduction mapping. Note that the set of $\boldsymbol{m}$ constructs the null space of $R(\boldsymbol{x})$, and the reduction mapping $\boldsymbol{h}$ locally projects $\boldsymbol{y}$ into $\overline{S^n}$ along the null space of $R(\boldsymbol{x})$.

### 3.4 Optimal reduction mapping

Let a noise $\boldsymbol{m}$ in $S^m$ be an additive Gaussian noise of a mean $\boldsymbol{0}$ and a variance-covariance matrix $V(\boldsymbol{y})$.

By introducing a Gramian matrix at $\boldsymbol{y}$ in $S^m$ as $F(\boldsymbol{y}) = \frac{1}{m}V^{-1}(\boldsymbol{y})$, $S^m$ can be regarded as a Riemannian signal space. The Gramian matrix in $S^m$ provides the metric in $\overline{S^n}$ as $G = \frac{m}{n}(RF^{-1}R^\top)^{-1}$, since the mapping from $S^m$ into $\overline{S^n}$ is the reduction mapping $R$. To realize an efficient transmission, the channel capacity $I = \log\{b_n \int \sqrt{g(\boldsymbol{x})}d\boldsymbol{x}\}$ should be maximized, that is, $g(\boldsymbol{x})$ should be maximized.

The reduction mapping which maximize $g(\boldsymbol{x})$ is called an *optimal reduction mapping* and the decoding by the optimal reduction mapping is called an *optimal signal extraction method*. By Lagrange's multiplier method, The optimal reduction mapping is derived as

$$\widetilde{R} = \frac{m}{n}\widetilde{G}^{-1}T^\top F\,.$$

where $\widetilde{G} = \frac{m}{n}T^\top FT\left(= \frac{m}{n}\left(\widetilde{R}F^{-1}\widetilde{R}^\top\right)^{-1}\right)$, and this is the pdf of noise in $\overline{S^n}$ under the optimal reduction mapping. The $\widetilde{G}$ is called an *optimal Gramian matrix*.

It is known that the optimal Gramian matrix is proportional to Fisher information matrix for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$.

## 4 Geometry of two class FLDA

In this section, we explain the geometry of two class FLDA, which is equivalent to the optimal reduction mapping in the context of communication theory.

Let patterns be probabilistically generated from a pdf $p(\boldsymbol{x}; s)$, which is a mixture of two pdfs $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ at a rate of $\frac{1-s}{2} : \frac{1+s}{2}$ as
$p(\boldsymbol{x}; s) = \frac{1-s}{2} p_1(\boldsymbol{x}) + \frac{1+s}{2} p_2(\boldsymbol{x})$.

Let $p_i(\boldsymbol{x})$'s are assumed to be a Gaussian distribution of mean $\boldsymbol{\mu}^{(i)}$, the mixture $p(\boldsymbol{x}; s)$ is also a Gaussian distribution of mean $\boldsymbol{\mu} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} + s\frac{\boldsymbol{\mu}_2-\boldsymbol{\mu}_1}{2}$. Then, the mixture rate $s$ can be computed by $s = \frac{\langle 2\boldsymbol{\mu}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)|\boldsymbol{\mu}\rangle}{\langle \boldsymbol{\mu}_2-\boldsymbol{\mu}_1|\boldsymbol{\mu}\rangle}$ from $\boldsymbol{\mu}$, where $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. The value $s$ is firstly defined as the mixture rate of two pdfs. However, for a given sample $\boldsymbol{x}$, we can give another meanings for the value $s(\boldsymbol{x}) = \frac{\langle 2\boldsymbol{x}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)|\boldsymbol{x}\rangle}{\langle \boldsymbol{\mu}_2-\boldsymbol{\mu}_1|\boldsymbol{x}\rangle}$ as a "label likelihood" of the $\boldsymbol{x}$, which gives the suitable pdf for the $\boldsymbol{x}$.

In our new framework, the imbedding mapping and the reduction mapping is set to $\boldsymbol{x} = \boldsymbol{f}(s)$ and $s = h(\boldsymbol{x})$, respectively, as follows:

$$\boldsymbol{x} = \boldsymbol{f}(s) = \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} + s\frac{\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1}{2} \in S^2, \quad (4)$$

$$s = h(\boldsymbol{x}) = \frac{\langle 2\boldsymbol{\mu} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)|\boldsymbol{x}\rangle}{\langle \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)|\boldsymbol{x}\rangle} \in S^1 . \quad (5)$$

As mentioned before, there holds $s = h(\boldsymbol{f}(s))$ when there are no noise. However, the channel is contaminated by noise, and the observed signal $\boldsymbol{x}$ equals to the sum of the truth $\boldsymbol{x}_0$ and the noise $\boldsymbol{m}$ as $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{m}$. Therefore, an observed (transmitted) label likelihood is the sum of the true label $s$ and the noise $n$ as $s = s_0 + n$. Then, discrete label $s \in \{-1, 1\}$ is transmitted to a continuous label likelihood $s \in S^1 = \mathbb{R}$.

We assume the distribution of $\boldsymbol{m}$ does not depend on the points in $S^2$. This assumption corresponds the homoscedasticity assumption in FLDA (All variance-covariance matrix of classes could be replaced to the same variance-covariance matrix by Eq. (1)). We also assume that noise $\boldsymbol{m}$ follows a Gaussian distribution with a mean $\mathbf{0}$ and a variance-covariance matrix $V$. Under these assumptions, the Gramian matrix in $S^2$ is computed as $F = \frac{1}{2}V^{-1}$.

From Eq. (4)-(5), the tangent matrix $T$ and the reduction matrix $R$ are obtained as

$$T = \frac{\partial \boldsymbol{f}}{\partial s} = \frac{\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1}{2} ,$$

$$R = \frac{\partial h}{\partial \boldsymbol{x}}^\top = \frac{1}{\langle \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{x}\rangle}(4\boldsymbol{x} - \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top - \frac{\langle 2\boldsymbol{x} - \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|\boldsymbol{x}\rangle}{\langle \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|\boldsymbol{x}\rangle^2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top .$$

Here, $R$ depends on $s$ and $s$ should be determined so as to maximize the channel capacity of the label transmission. Consequently, the optimal reduction matrix $\widetilde{R}$ is given as

$$\widetilde{R} = 2\widetilde{G}^{-1}T^\top F = \frac{1}{2\mathtt{det}\,\widetilde{G}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top V^{-1} .$$

By comparing $(2\mathtt{det}\,\widetilde{G})\widetilde{R}^\top = V^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ with Eq. (2), the maximization of channel capacity in transmitting class labels results in FLDA since $V = \Sigma_{\mathrm{W}}$ and $\boldsymbol{\mu}_c = \boldsymbol{\mu}^{(c)} (c = 1, 2)$. Then it is geometrically proved that two class FLDA decodes communicating classes to maximize the channel capacity.

## 5 Conclusion

This paper gives the geometry of two class Fisher's linear discriminant analysis from the point of view of communication theory. As a result, Fisher's linear discriminant analysis gives the maximum channel capacity with transmitting class labels.

## References


[1] S. Amari, Theory of Information Spaces — "A Geometrical Foundation of the Analysis of Communication Systems," RAAG Memoirs, **4**:373-418, 1968.

[2] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," Neural Computation, **12**(10):2385-2404, 2006.

[3] P. N. Belhumeur et al., "Eigenfaces vs Fisherfaces: Recognition using class specific linear projection," IEEE TPAMI, **19**:711-720, 1997.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, 2006.

[5] H. P. Decell and S. M. Mayekar, "Feature Combinations and the Divergence Criterion," Computers and Math. with Applications, **3**:71-76, 1977.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, 1973.

[7] R. A. Fishier, "The use of multiple measurements in taxonomic problems," Ann. of Eigenics, **7**:179-188, 1936.

[8] J. Fujiki, "Finding discriminant axes from multiple viewpoints," In Proc. the 13th IAPR in Int'l Conf. on Mach. Vision Applications, 73-76, 2013.

[9] N. Gkalelis, V. Mezaris and I. Kompatsiaris, "Mixture subclass discriminant analysis," IEEE Signal Processing Lett. **18**(5):319-322, 2011.

[10] T. Hastie, A. Buja and R. Tibshirani, "Penalized Discriminant Analysis," The Ann. of Statist., **23**(1):73-102, 1995.

[11] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixture," J. R. Soc. of Statist. Soc. B., **58**:155-176, 1996.

[12] M. Loog, and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," IEEE TPAMI, **26**(6):732-739, 2004.

[13] H. Sakano et al. "Extended Fisher Criterion Based on Auto-correlation Matrix Information," In Proc. of Structural, Syntactic, and Statist. Patt. Recognit. - Joint IAPR Int'l Workshop,
SSPR& SPR 2012, 409-416, 2012.

[14] A. Sierra, "High-order Fishers discriminant analysis," Patt. Recognit., **35**(6):1291-1302, 2002.

[15] M. Zhu, A. M. Martinez, "Subclass discriminant analysis," IEEE TPAMI, **28**(8):1274-1286, 2006.