Fisher's Linear Discriminant Analysis

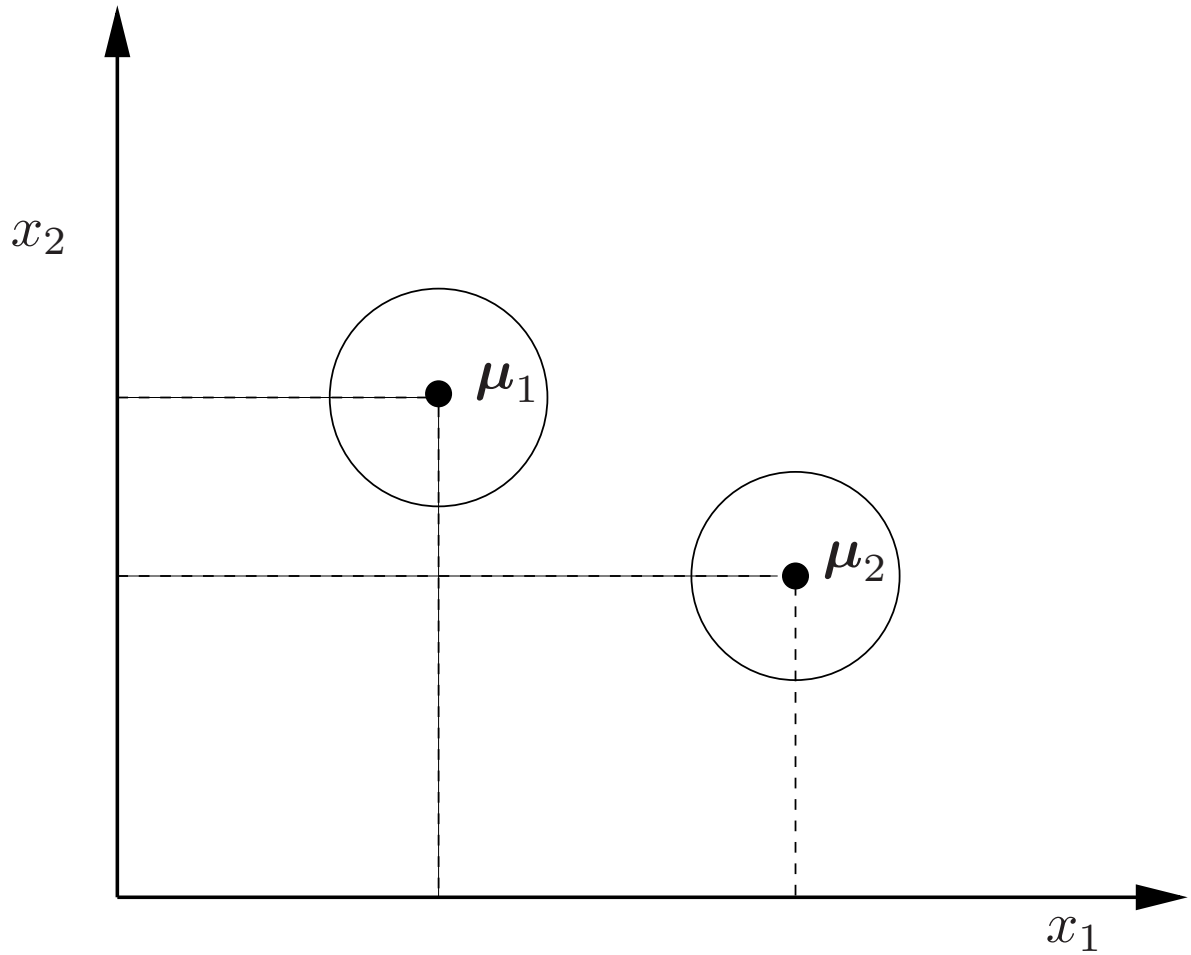
Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

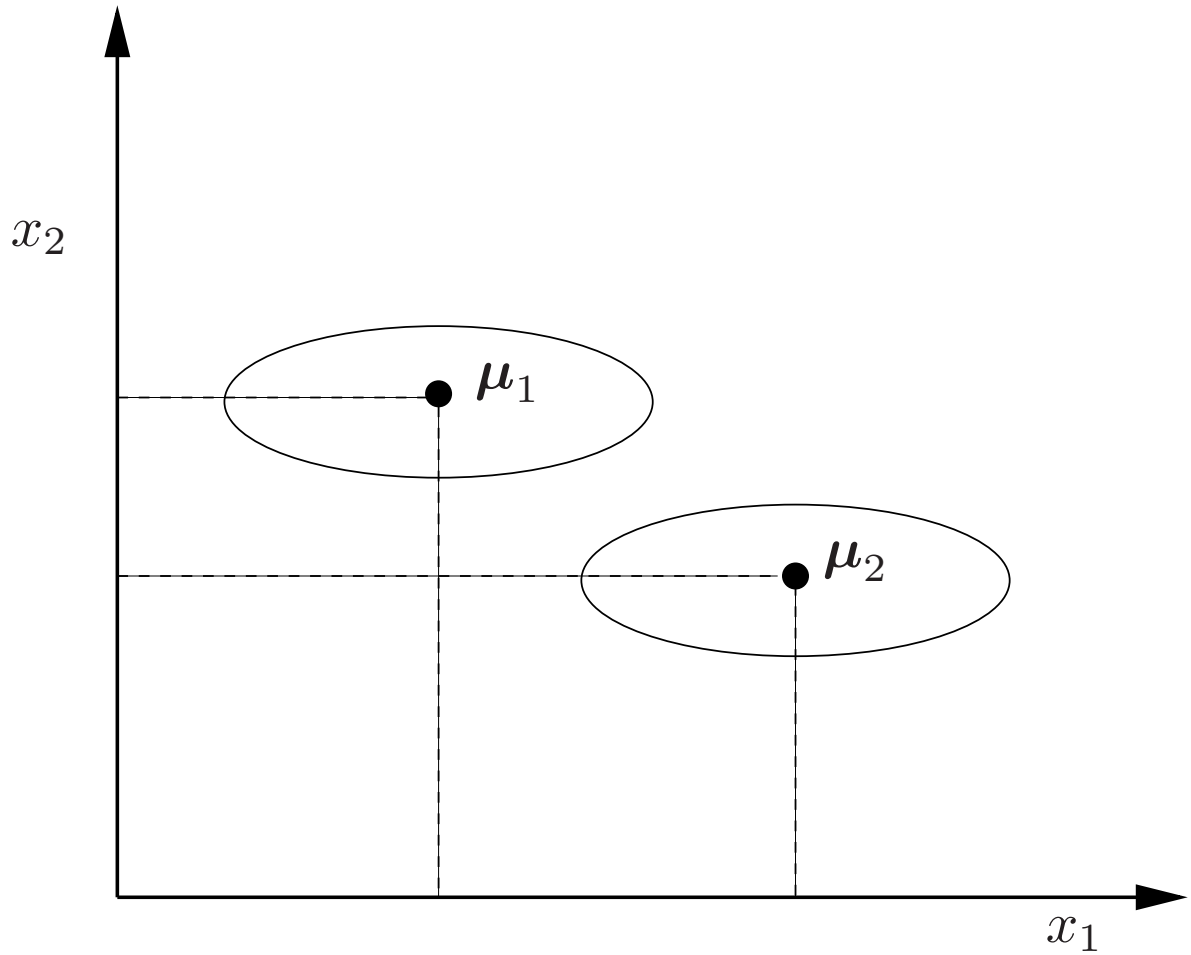
FLD or LDA

- Introduced by Fisher (1936)
- One of widely-used linear discriminant analysis (LDA) methods
- Curse of dimensionality
  - Linear dimensionality reduction: PCA, ICA, FLD, MDS
  - Nonlinear dimensionality reduction: Isomap, LLE, Laplacian eigenmap
- FLD aims at achieving an optimal linear dimensionality reduction for classification

An Example: Isotropic Case



FLD: A Graphical Illustration



Two Classes

Given a set of data points, $\{\boldsymbol{x} \in \mathbb{R}^D\}$, one wished to find a linear projection of the data onto a 1-dimensional space, $y = \boldsymbol{w}^\top \boldsymbol{x}$.
Sample means for $\boldsymbol{x}$:

$$\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}.$$

Sample means for the projected points:

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{N_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{w}^\top \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{\mu}_i.$$

We know that the difference between sample means is not always a good measure of the separation between projected points:

$$|\tilde{\mu}_1 - \tilde{\mu}_2| = \left| \boldsymbol{w}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right|.$$

Scale $\|\boldsymbol{w}\| \uparrow \quad \Rightarrow \quad |\tilde{\mu}_1 - \tilde{\mu}_2| \uparrow$ (not desirable!).

FLD: Two Classes

Define the within-class scatter for projected samples by $\tilde{s}_1^2 + \tilde{s}_2^2$, where

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{\mu}_i)^2 = \boldsymbol{w}^\top \underbrace{\left[\sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \right]}_{\boldsymbol{S}_i} \boldsymbol{w}.$$

FLD finds

$$\boldsymbol{w} = \arg\max_{\boldsymbol{w}} \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}},$$

where $\boldsymbol{S}_W = \boldsymbol{S}_1 + \boldsymbol{S}_2$ (within-class scatter matrix) and $\boldsymbol{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$ (between-class scatter matrix).

$$\boxed{\arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}} \Rightarrow \boxed{\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}}$$

(generalized eigenvalue problem).

Multiple Discriminant Functions

For the case of K classes, FLD involves $K-1$ discriminant functions, i.e., the projection is from $\mathbb{R}^D$ to $\mathbb{R}^{K-1}$.
Given a set of data $\{\boldsymbol{x} \in \mathbb{R}^D\}$, one wishes to find a linear lower-dimensional embedding $\boldsymbol{W}^\top \in \mathbb{R}^{(K-1)\times D}$ such that $\{\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}\}$ are classified as well as possible in the lower-dimensional space.

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_{K-1} \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_{K-1}^\top \end{bmatrix}}_{\boldsymbol{W}^\top} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}}_{\boldsymbol{x}}.$$

Scatter Matrices

Within-class scatter matrix

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i) (\boldsymbol{x} - \boldsymbol{\mu}_i)^\top .$$

Between-class scatter matrix

$$\boldsymbol{S}_B = \sum_{i=1}^{K} \sum_{C_i} (\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top = \sum_{i=1}^{K} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top .$$

Total scatter matrix: $\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$

$$\boldsymbol{S}_T = \sum_{\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{\mu}) (\boldsymbol{x} - \boldsymbol{\mu})^\top .$$

$$\text{Rank}(\boldsymbol{S}_B) \leq K - 1, \quad \text{Rank}(\boldsymbol{S}_W) \leq N - K, \quad \text{Rank}(\boldsymbol{S}_T) \leq N - 1.$$

Total Scatter Matrix

Define $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K]$ where $\boldsymbol{X}_i$ is a matrix whose columns are associated with data vectors belonging to C_i.
Define

$$\begin{aligned} \boldsymbol{H}_W &= \left[\boldsymbol{X}_1 - \boldsymbol{\mu}_1 \boldsymbol{e}_1^\top, \ldots, \boldsymbol{X}_K - \boldsymbol{\mu}_K \boldsymbol{e}_K^\top\right], \\ \boldsymbol{H}_B &= \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu})\boldsymbol{e}_1^\top, \ldots, (\boldsymbol{\mu}_K - \boldsymbol{\mu})\boldsymbol{e}_K^\top\right], \\ \boldsymbol{H}_T &= \left[\boldsymbol{x}_1 - \boldsymbol{\mu}_1, \ldots, \boldsymbol{x}_N - \boldsymbol{\mu}\right]. \end{aligned}$$

One can easily see that $\boldsymbol{H}_T = \boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{e}^\top = \boldsymbol{H}_W + \boldsymbol{H}_B$.
We also have $\boldsymbol{S}_W = \boldsymbol{H}_W \boldsymbol{H}_W^\top, \quad \boldsymbol{S}_B = \boldsymbol{H}_B \boldsymbol{H}_B^\top, \quad \boldsymbol{S}_T = \boldsymbol{H}_T \boldsymbol{H}_T^\top$.
Since $\boldsymbol{H}_W \boldsymbol{H}_B^\top = 0$, we have

$$\boldsymbol{S}_T = (\boldsymbol{H}_W + \boldsymbol{H}_B)(\boldsymbol{H}_W + \boldsymbol{H}_B)^\top = \boldsymbol{S}_W + \boldsymbol{S}_B.$$

The column vectors of $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ are linear combinations of centered data samples.

FLD: Multiple Classes

Define

$$\begin{aligned}\widetilde{\boldsymbol{S}}_W &= \sum_{i=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{Y}_i} (\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_i)^\top \\ \widetilde{\boldsymbol{S}}_B &= \sum_{i=1}^{K} N_i (\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}})^\top .\end{aligned}$$

One can easily show that

$$\begin{aligned}\widetilde{\boldsymbol{S}}_W &= \boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W}, \\ \widetilde{\boldsymbol{S}}_B &= \boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W}.\end{aligned}$$

FLD seeks $K-1$ discriminant functions $\boldsymbol{W}$ such that $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$:

$$\begin{aligned}
\boldsymbol{W} &= \arg\max_{\boldsymbol{W}} \mathcal{J}_{FLD} \\
&= \arg\max_{\boldsymbol{W}} \operatorname{tr}\left\{\widetilde{\boldsymbol{S}}_W^{-1}\widetilde{\boldsymbol{S}}_B\right\} \\
&= \arg\max_{\boldsymbol{W}} \operatorname{tr}\left\{\left(\boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W}\right)\right\},
\end{aligned}$$

leading to

$$\boxed{\arg\max_{\boldsymbol{W}} \mathcal{J}_{FLD}} \Rightarrow \boxed{\boldsymbol{S}_B \boldsymbol{w}_i = \lambda_i \boldsymbol{S}_W \boldsymbol{w}_i}.$$

generalized eigenvalue problem

Rayleigh Quotient

Definition

Let $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ be symmetric. The Rayleigh quotient $R(\boldsymbol{x}, \boldsymbol{A})$ is defined by

$$R(\boldsymbol{x}, \boldsymbol{A}) = \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}.$$

Theorem

Let $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ be symmetric with its eigenvalues being $\{\lambda_1 \geq \cdots \geq \lambda_m\}$. For $\forall \boldsymbol{x} \neq \boldsymbol{0} \in \mathbb{R}^m$, we have

$$\lambda_m \leq \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} \leq \lambda_1,$$

and in particular,

$$\lambda_m = \min_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}, \qquad \lambda_1 = \max_{\boldsymbol{x} \neq 0} \frac{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}.$$

An Extremal Property of Generalized Eigenvalues

Theorem

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $m \times m$ matrices, with $\boldsymbol{A}$ being nonnegative definite and $\boldsymbol{B}$ positive definite. For $h = 1, \ldots, m$, define

$$\boldsymbol{X}_h = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_h], \quad \boldsymbol{Y}_h = [\boldsymbol{x}_h, \ldots, \boldsymbol{x}_m],$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are linear independent eigenvectors of $\boldsymbol{B}^{-1}\boldsymbol{A}$ corresponding to the eigenvalues

$$\lambda_1\left(\boldsymbol{B}^{-1}\boldsymbol{A}\right) \geq \cdots \geq \lambda_m\left(\boldsymbol{B}^{-1}\boldsymbol{A}\right).$$

Then

$$\begin{aligned} \lambda_m\left(\boldsymbol{B}^{-1}\boldsymbol{A}\right) &= \min_{\boldsymbol{Y}_{h+1}^{\top}\boldsymbol{B}\boldsymbol{x}=0} \frac{\boldsymbol{x}^{\top}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\top}\boldsymbol{B}\boldsymbol{x}}, \\ \lambda_1\left(\boldsymbol{B}^{-1}\boldsymbol{A}\right) &= \max_{\boldsymbol{X}_{h-1}^{\top}\boldsymbol{B}\boldsymbol{x}=0} \frac{\boldsymbol{x}^{\top}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\top}\boldsymbol{B}\boldsymbol{x}}, \end{aligned}$$

where $\boldsymbol{x} = 0$ is excluded.

Relation to Least Squares Regression: Binary Class

Given a training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \{1, -1\}$, consider a linear discriminant function:

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^\top \boldsymbol{x}_i + b.$$

Partition the data matrix into two groups, each group of which contains examples in class 1 or class 2, i.e., $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, where $\boldsymbol{X}_1 \in \mathbb{R}^{D \times N_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{D \times N_2}$.
Define binary label vector $\boldsymbol{y} \in \mathbb{R}^N$, then LS regression is formulated as

$$\arg\min_{\boldsymbol{w}, b} \|\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w} - b\mathbf{1}_N\|^2,$$

where $\mathbf{1}_N$ is the N-dimensional vector of all ones, which can be re-written as

$$\arg\min_{\boldsymbol{w}, b} \left\| \begin{bmatrix} \boldsymbol{X}_1^\top & \mathbf{1}_{N_1} \\ \boldsymbol{X}_2^\top & \mathbf{1}_{N_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} - \begin{bmatrix} \mathbf{1}_{N_1} \\ -\mathbf{1}_{N_2} \end{bmatrix} \right\|^2.$$

The solution to this LS problem satisfies the normal equation:

$$\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \\ \mathbf{1}_{N_1}^\top & \mathbf{1}_{N_2}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1^\top & \mathbf{1}_{N_1} \\ \boldsymbol{X}_2^\top & \mathbf{1}_{N_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \\ \mathbf{1}_{N_1}^\top & \mathbf{1}_{N_2}^\top \end{bmatrix} \begin{bmatrix} \mathbf{1}_{N_1} \\ -\mathbf{1}_{N_2} \end{bmatrix},$$

which is written as

$$\begin{bmatrix} \boldsymbol{X}_1\boldsymbol{X}_1^\top + \boldsymbol{X}_2\boldsymbol{X}_2^\top & \boldsymbol{X}_1\mathbf{1}_{N_1} + \boldsymbol{X}_2\mathbf{1}_{N_2} \\ \mathbf{1}_{N_1}^\top\boldsymbol{X}_1^\top + \mathbf{1}_{N_2}^\top\boldsymbol{X}_2^\top & \mathbf{1}_{N_1}^\top\mathbf{1}_{N_1} + \mathbf{1}_{N_2}^\top\mathbf{1}_{N_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1\mathbf{1}_{N_1} - \boldsymbol{X}_2\mathbf{1}_{N_2} \\ \mathbf{1}_{N_1}^\top\mathbf{1}_{N_1} - \mathbf{1}_{N_2}^\top\mathbf{1}_{N_2} \end{bmatrix}.$$

Recall

$$\begin{aligned} \boldsymbol{S}_B &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \\ \boldsymbol{S}_W &= \left(\boldsymbol{X}_1 - \boldsymbol{\mu}_1\mathbf{1}_{N_1}^\top\right)\left(\boldsymbol{X}_1 - \boldsymbol{\mu}_1\mathbf{1}_{N_1}^\top\right)^\top + \left(\boldsymbol{X}_2 - \boldsymbol{\mu}_2\mathbf{1}_{N_2}^\top\right)\left(\boldsymbol{X}_2 - \boldsymbol{\mu}_2\mathbf{1}_{N_2}^\top\right)^\top \\ &= \boldsymbol{X}_1\boldsymbol{X}_1^\top - N_1\boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top + \boldsymbol{X}_2\boldsymbol{X}_2^\top - N_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top. \end{aligned}$$

With $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$, the normal equation is written as

$$\begin{bmatrix} \boldsymbol{S}_W + N_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\top + N_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top & N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2 \\ (N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2)^\top & N_1 + N_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} = \begin{bmatrix} N_1\boldsymbol{\mu}_1 - N_2\boldsymbol{\mu}_2 \\ N_1 - N_2 \end{bmatrix}.$$

Solve the 2nd equation for b to obtain

$$b = \frac{(N_1 - N_2) - (N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2)^\top \boldsymbol{w}}{N_1 + N_2}.$$

Substitute this into the 1st equation to obtain

$$\left[\boldsymbol{S}_W + \frac{N_1 N_2}{N_1 + N_2}\boldsymbol{S}_B\right] \boldsymbol{w} = 2N_1N_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Note that the vector $\boldsymbol{S}_B\boldsymbol{w}$ is in the direction of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ for $\forall \boldsymbol{w}$, since $\boldsymbol{S}_B\boldsymbol{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{w}$.

Thus we write

$$\frac{N_1 N_2}{N_1 + N_2}\boldsymbol{S}_B\boldsymbol{w} = (2N_1 N_2 - \alpha)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Then we have

$$\boldsymbol{w} = \alpha \boldsymbol{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

which is identical to FLD solutions except for scaling factor.

Simultaneous Diagonalization

The goal: Given two symmetric matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, find a linear transformation $\boldsymbol{W}$ such that

$$\begin{aligned} \boldsymbol{W}^\top \boldsymbol{\Sigma}_1 \boldsymbol{W} &= \boldsymbol{I}, \\ \boldsymbol{W}^\top \boldsymbol{\Sigma}_2 \boldsymbol{W} &= \boldsymbol{\Lambda}. \quad \text{(diagonal)} \end{aligned}$$

Methods: It turns out that simultaneous diagonalization involves the generalized eigen-decomposition.

- Two-stage method
  1. whitening
  2. unitary transformation
- Single-stage method: generalized eigenvalue decomposition

Simultaneous Diagonalization: Algorithm Outline

1. First, whiten $\boldsymbol{\Sigma}_1$, i.e.,

$$\begin{aligned} \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}_1^\top \boldsymbol{\Sigma}_1 \boldsymbol{U}_1 \boldsymbol{D}^{-\frac{1}{2}} &= \boldsymbol{I}, \\ \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}_1^\top \boldsymbol{\Sigma}_2 \boldsymbol{U}_1 \boldsymbol{D}^{-\frac{1}{2}} &= \boldsymbol{K}, \quad \text{(not diagonal)}, \end{aligned}$$

   where $\boldsymbol{\Sigma}_1 = \boldsymbol{U}_1 \boldsymbol{D} \boldsymbol{U}_1^\top$.

2. Second, apply an unitary transformation to diagonalize $\boldsymbol{K}$, i.e.,

$$\begin{aligned} \boldsymbol{U}_2^\top \boldsymbol{I} \boldsymbol{U}_2 &= \boldsymbol{I}, \\ \boldsymbol{U}_2^\top \boldsymbol{K} \boldsymbol{U}_2 &= \boldsymbol{\Lambda}, \end{aligned}$$

   where $\boldsymbol{K} = \boldsymbol{U}_2 \boldsymbol{\Lambda} \boldsymbol{U}_2^\top$.

Then, the transformation $\boldsymbol{W}$ which simultaneously diagonalizes $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, is given by, $\boldsymbol{W} = \boldsymbol{U}_1 \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{U}_2$, such that $\boldsymbol{W}^\top \boldsymbol{\Sigma}_1 \boldsymbol{W} = \boldsymbol{I}$ and $\boldsymbol{W}^\top \boldsymbol{\Sigma}_2 \boldsymbol{W} = \boldsymbol{\Lambda}$

Simultaneous Diagonalization: Generalized Eigen-Decomposition

Alternatively we can diagonalize two symmetric matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ as

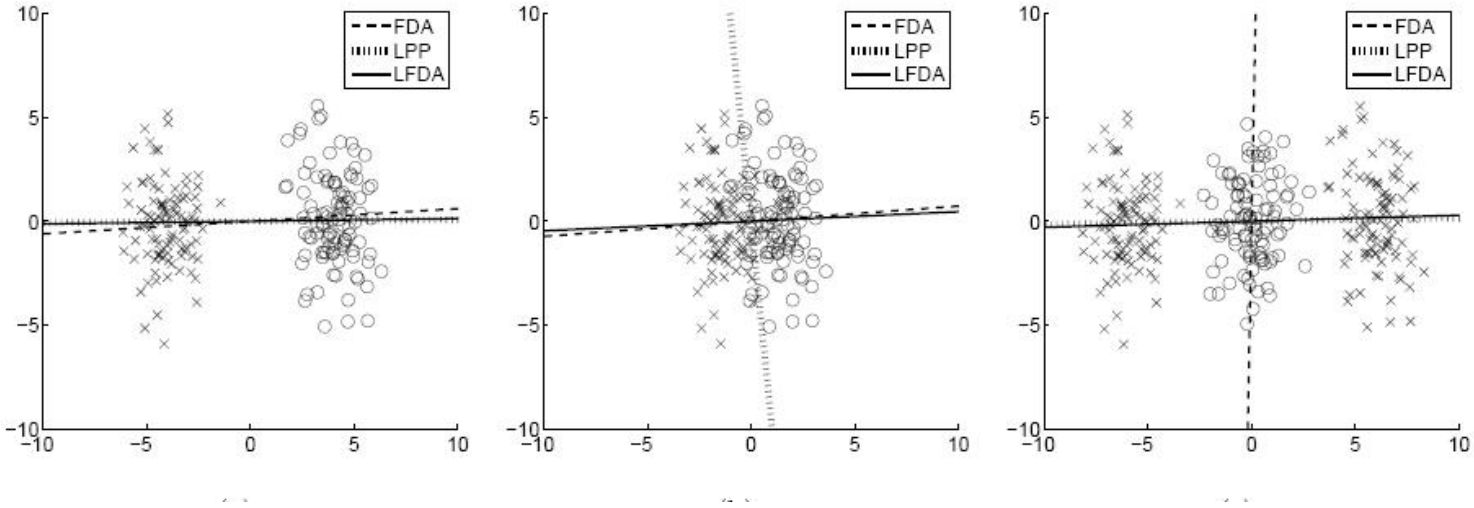
$$\begin{aligned} \boldsymbol{W}^\top \mathbf{\Sigma}_1 \boldsymbol{W} &= \boldsymbol{I}, \\ \boldsymbol{W}^\top \mathbf{\Sigma}_2 \boldsymbol{W} &= \mathbf{\Lambda}, \quad \text{(diagonal)} \end{aligned}$$

where $\mathbf{\Lambda}$, $\boldsymbol{W}$ are eigenvalues and eigenvectors of $\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2$, i.e.,

$$\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 \boldsymbol{W} = \boldsymbol{W}\mathbf{\Lambda}.$$

Prove it!

Example: Multi-Modal Data



Alternative Expressions of $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$

Alternatively, $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ are expressed as

$$\begin{aligned}
\boldsymbol{S}_W &= \frac{1}{2}\sum_i \sum_j A_{ij}^W (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \\
\boldsymbol{S}_B &= \frac{1}{2}\sum_i \sum_j A_{ij}^B (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top,
\end{aligned}$$

$$\begin{aligned}
A_{ij}^W &= \begin{cases} \frac{1}{N_k} & \text{if } \boldsymbol{x}_i \in C_k \text{ and } \boldsymbol{x}_j \in C_k, \\ 0 & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_i \text{ are in different classes,} \end{cases} \\
A_{ij}^B &= \begin{cases} \frac{1}{N} - \frac{1}{N_k} & \text{if } \boldsymbol{x}_i \in C_k \text{ and } \boldsymbol{x}_j \in C_k, \\ \frac{1}{N} & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_i \text{ are in different classes.} \end{cases}
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{S}_W &= \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \\
&= \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in C_i} \left(\boldsymbol{x}_j - \frac{1}{N_i} \sum_{\boldsymbol{x}_u \in C_i} \boldsymbol{x}_u \right) \left(\boldsymbol{x}_j - \frac{1}{N_i} \sum_{\boldsymbol{x}_v \in C_i} \boldsymbol{x}_v \right)^\top \\
&= \sum_{i=1}^{K} \sum_{\boldsymbol{x}_j \in C_i} \left\{ \boldsymbol{x}_j \boldsymbol{x}_j^\top - \frac{1}{N_i} \sum_{\boldsymbol{x}_v \in C_i} \boldsymbol{x}_j \boldsymbol{x}_v^\top - \frac{1}{N_i} \sum_{\boldsymbol{x}_u \in C_i} \boldsymbol{x}_u \boldsymbol{x}_j^\top + \frac{1}{N_i^2} \sum_{\boldsymbol{x}_u \in C_i} \sum_{\boldsymbol{x}_v \in C_i} \boldsymbol{x}_u \boldsymbol{x}_v^\top \right\} \\
&= \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top - \sum_{i=1}^{K} \frac{1}{N_i} \sum_{\boldsymbol{x}_u \in C_i} \sum_{\boldsymbol{x}_v \in C_i} \boldsymbol{x}_u \boldsymbol{x}_v^\top \\
&= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} A_{ij}^W \right) \boldsymbol{x}_i \boldsymbol{x}_i^\top - \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^W \boldsymbol{x}_i \boldsymbol{x}_j^\top \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^W \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_j \boldsymbol{x}_j^\top - \boldsymbol{x}_i \boldsymbol{x}_j^\top - \boldsymbol{x}_j \boldsymbol{x}_i^\top \right) \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^W (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top .
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{S}_B &= \boldsymbol{S}_T - \boldsymbol{S}_W \\
&= \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top - \boldsymbol{S}_W \\
&= \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{1}{N} \right) \boldsymbol{x}_i \boldsymbol{x}_i^\top - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_j^\top \\
&\quad - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^W \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_j \boldsymbol{x}_j^\top - \boldsymbol{x}_i \boldsymbol{x}_j^\top - \boldsymbol{x}_j \boldsymbol{x}_i^\top \right) \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{N} - A_{ij}^W \right) \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top + \boldsymbol{x}_j \boldsymbol{x}_j^\top - \boldsymbol{x}_i \boldsymbol{x}_j^\top - \boldsymbol{x}_j \boldsymbol{x}_i^\top \right) \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{1}{N} - A_{ij}^W \right) (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top .
\end{aligned}
$$

Local Within-Class and Between-Class Scatter

Given weighted adjacency matrix $[A_{ij}]$, introduce local within-class scatter and local between-class scatter:

$$\begin{aligned}
\overline{\boldsymbol{S}}_W &= \frac{1}{2}\sum_i\sum_j \overline{A}_{ij}^W (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \\
\overline{\boldsymbol{S}}_B &= \frac{1}{2}\sum_i\sum_j \overline{A}_{ij}^B (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top,
\end{aligned}$$

$$\overline{A}_{ij}^W = \begin{cases} \frac{A_{ij}}{N_k} & \text{if } \boldsymbol{x}_i \in C_k \text{ and } \boldsymbol{x}_j \in C_k, \\ 0 & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_i \text{ are in different classes,} \end{cases}$$

$$\overline{A}_{ij}^B = \begin{cases} A_{ij}\left(\frac{1}{N} - \frac{1}{N_k}\right) & \text{if } \boldsymbol{x}_i \in C_k \text{ and } \boldsymbol{x}_j \in C_k, \\ \frac{1}{N} & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_i \text{ are in different classes.} \end{cases}$$

Local Fisher Discriminant Analysis (LFDA)

Proposed by M. Sugiyama (ICML-2006).

LFDA seeks $K-1$ discriminant functions $\boldsymbol{W}$ such that $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$:

$$\arg\max_{\boldsymbol{W}} \operatorname{tr}\left\{\left(\boldsymbol{W}^\top \overline{\boldsymbol{S}}_W \boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^\top \overline{\boldsymbol{S}}_B \boldsymbol{W}\right)\right\},$$

Local within-class scatter matrix

$$\overline{\boldsymbol{S}}_W = \frac{1}{2}\sum_i \sum_j \overline{A}_{ij}^W (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top,$$

Local between-class scatter matrix

$$\overline{\boldsymbol{S}}_B = \frac{1}{2}\sum_i \sum_j \overline{A}_{ij}^B (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top.$$