Fisher Discriminant Analysis

Statistical Machine Learning (ENGG*6600*02)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh
Summer 2023

One-dimensional Subspace

Scatters in Two-Class Case

- Assume we have two classes, $\{\boldsymbol{x}_i^{(1)}\}_{i=1}^{n_1}$ and $\{\boldsymbol{x}_i^{(2)}\}_{i=1}^{n_2}$, where n_1 and n_2 denote the sample size of the first and second class, respectively, and $\boldsymbol{x}_i^{(j)}$ denotes the i-th instance of the j-th class.
- If the data instances of the j-th class are projected onto a one-dimensional subspace (vector $\boldsymbol{u}$) by $\boldsymbol{u}^\top \boldsymbol{x}_i^{(j)}$, the mean and the variance of the projected data are $\boldsymbol{u}^\top \boldsymbol{\mu}_j$ and $\boldsymbol{u}^\top \boldsymbol{S}_j \boldsymbol{u}$, respectively, where $\boldsymbol{\mu}_j$ and $\boldsymbol{S}_j$ are the mean and covariance matrix (scatter) of the j-th class.
- The mean of the j-th class is:

$$x_i^{(j)} \mapsto u^\top x_i^{(j)}$$

$$\mathbb{R}^d \ni \boldsymbol{\mu}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{x}_i^{(j)}. \qquad \mu_j \mapsto u^\top \mu_j \tag{1}$$

$$S_j \mapsto u^\top S_j u$$

Scatters in Two-Class Case

$\|a-b\|_2^2 = (a-b)^\top(a-b)$

- After projection onto the one-dimensional subspace, the distance between the means of classes is:

$1\times d \quad d\times d \quad d\times 1$

$$
\begin{aligned}
\mathbb{R} \ni d_B &:= (\boldsymbol{u}^\top \boldsymbol{\mu}_1 - \boldsymbol{u}^\top \boldsymbol{\mu}_2)^\top (\boldsymbol{u}^\top \boldsymbol{\mu}_1 - \boldsymbol{u}^\top \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u}\boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&\overset{(a)}{=} \mathbf{tr}((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u}\boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \overset{(b)}{=} \mathbf{tr}(\boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u}) \\
&\overset{(c)}{=} \boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u} \overset{(d)}{=} \boldsymbol{u}^\top \boldsymbol{S}_B\, \boldsymbol{u},
\end{aligned}
\tag{2}
$$

where (a) is because $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u}\boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is a scalar, (b) is because of the cyclic property of trace, (c) is because $\boldsymbol{u}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{u}$ is a scalar, and (d) is because we define:

$$
\mathbb{R}^{d\times d} \ni \boldsymbol{S}_B := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top,
\tag{3}
$$

as the **between-scatter** of classes.

- The Eq. (2) can also be interpreted in this way: the d_B is the variance of projection of the class means or the squared length of reconstruction of the class means.

Scatters in Two-Class Case

- We saw that the variance of projection is $\boldsymbol{u}^\top \boldsymbol{S}_j \boldsymbol{u}$ for the j-th class. If we add up the variances of projections of the two classes, we have:

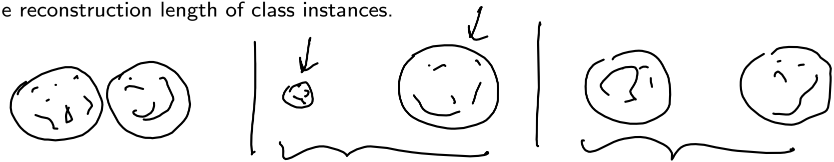
$$\mathbb{R} \ni d_W := \boldsymbol{u}^\top \boldsymbol{S}_1 \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{S}_2 \boldsymbol{u} = \boldsymbol{u}^\top (\boldsymbol{S}_1 + \boldsymbol{S}_2)\, \boldsymbol{u}$$
$$\overset{(a)}{=} \boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}, \tag{4}$$

where:

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S}_W := \boldsymbol{S}_1 + \boldsymbol{S}_2, \tag{5}$$

is the **within-scatter** of classes.

- The d_W is the summation of projection variance of class instances or the summation of the reconstruction length of class instances.



Scatters in Multi-Class Case: Variant 1

- Assume $\{\boldsymbol{x}_i^{(j)}\}_{i=1}^{n_j}$ are the instances of the j-th class where we have multiple classes. In this case, the **between-scatter** is defined as:

$$\underbrace{(\mu_1-\mu_2)(\mu_1-\mu_2)^\top}$$

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S}_B := \underbrace{\sum_{j=1}^{c}(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^\top}, \tag{6}$$

where c is the number of classes and:

$$\mathbb{R}^{d} \ni \boldsymbol{\mu} := \underbrace{\frac{1}{\sum_{k=1}^{c} n_k}\sum_{j=1}^{c} n_j\,\boldsymbol{\mu}_j} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i}, \tag{7}$$

is the weighted mean of means of classes or the total mean of data.

- It is noteworthy that some researches define the between-scatter in a weighted way:

$$\mathbb{R}^{d\times d} \ni \boldsymbol{S}_B := \sum_{j=1}^{c} n_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^\top. \tag{8}$$

Scatters in Multi-Class Case: Variant 1

- If we extend the Eq. (5) to c number of classes, the **within-scatter** is defined as:

$$\mathbb{R}^{d \times d} \ni \boldsymbol{S}_W := \sum_{j=1}^{c} \boldsymbol{S}_j \tag{9}$$

$$= \sum_{j=1}^{c} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_j)(\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_j)^\top, \tag{10}$$

where n_j is the sample size of the j-th class.

- In this case, the d_B and d_W are:

$$\mathbb{R} \ni d_B := \underbrace{\boldsymbol{u}^\top \boldsymbol{S}_B\, \boldsymbol{u},} \rightarrow \text{max} \tag{11}$$

$$\mathbb{R} \ni d_W := \underbrace{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u},} \rightarrow \text{min} \tag{12}$$

where $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are Eqs. (6) and (10).

Scatters in Multi-Class Case: Variant 2

- There is another variant for multi-class case in FDA. In this variant, the within-scatter is the same as Eq. (10). The between-scatter is, however, different.
- The **total-scatter** is defined as the covariance matrix of the whole data, regardless of classes [1]:

$$\underbrace{\mathbb{R}^{d \times d}} \ni \boldsymbol{S}_T := \underbrace{\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top}, \tag{13}$$

  where the total mean $\boldsymbol{\mu}$ is the Eq. (7). We can also use the scaled total-scatter by dropping the $1/n$ factor.
- On the other hand, the total scatter is equal to the summation of the within- and between-scatters:

$$\boldsymbol{S}_T = \underbrace{\boldsymbol{S}_W + \boldsymbol{S}_B}. \tag{14}$$

  Therefore, the between-scatter, in this variant, is obtained as:

$$\boldsymbol{S}_B := \boldsymbol{S}_T - \boldsymbol{S}_W. \tag{15}$$

Fisher Subspace: Variant 1

- In FDA, we want to maximize the projection variance (scatter) of means of classes and minimize the projection variance (scatter) of class instances. In other words, we want to maximize d_B and minimize d_W. The reason is that after projection, we want the within scatter of every class to be small and the between scatter of classes to be large; therefore, the instances of every class get close to one another and the classes get far from each other.
- The two mentioned optimization problems are:

$$\begin{cases} \underset{\boldsymbol{u}}{\text{maximize}} \quad d_B(\boldsymbol{u}), & (16) \\ \underset{\boldsymbol{u}}{\text{minimize}} \quad d_W(\boldsymbol{u}). \quad \Rightarrow \max_{u} \frac{1}{d_W} & (17) \end{cases}$$

- We can merge these two optimization problems as a regularized optimization problem:

$$\underset{\boldsymbol{u}}{\text{maximize}} \quad d_B(\boldsymbol{u}) - \alpha\, d_W(\boldsymbol{u}), \tag{18}$$

where $\alpha > 0$ is the regularization parameter.

- Another way of merging Eqs. (16) and (17) is:

$$\underset{\boldsymbol{u}}{\text{maximize}} \quad f(\boldsymbol{u}) := \frac{d_B(\boldsymbol{u})}{d_W(\boldsymbol{u})} = \boxed{\frac{\boldsymbol{u}^\top \boldsymbol{S}_B\, \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}},} \tag{19}$$

where $f(\boldsymbol{u}) \in \mathbb{R}$ is referred to as the **Fisher criterion** [2].

Fisher Subspace: Variant 1

- The Fisher criterion is a generalized Rayleigh-Ritz quotient (recall preliminaries):

$$f(\boldsymbol{u}) = R(\boldsymbol{S}_B, \boldsymbol{S}_W; \boldsymbol{u}). \tag{20}$$

According to the preliminaries slides, the optimization in Eq. (19) is equivalent to:

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{maximize}} \quad & \boldsymbol{u}^\top \boldsymbol{S}_B\, \boldsymbol{u} \\ \text{subject to} \quad & \boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u} = 1. \end{aligned} \tag{21}$$

- The Lagrangian [3] is:

$$\mathcal{L} = \boldsymbol{w}^\top \boldsymbol{S}_B\, \boldsymbol{w} - \lambda(\boldsymbol{w}^\top \boldsymbol{S}_W\, \boldsymbol{w} - 1),$$

where λ is the Lagrange multiplier. Equating the derivative of $\mathcal{L}$ to zero gives:

$$\begin{aligned} \mathbb{R}^d \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{u}} &= 2\, \boldsymbol{S}_B\, \boldsymbol{u} - 2\, \lambda\, \boldsymbol{S}_W\, \boldsymbol{u} \overset{\text{set}}{=} \boldsymbol{0} \\ \implies 2\, \boldsymbol{S}_B\, \boldsymbol{u} = 2\, \lambda\, \boldsymbol{S}_W\, \boldsymbol{u} &\implies \boxed{\boldsymbol{S}_B\, \boldsymbol{u} = \lambda\, \boldsymbol{S}_W\, \boldsymbol{u},} \end{aligned} \tag{22}$$

which is a generalized eigenvalue problem $(\boldsymbol{S}_B, \boldsymbol{S}_W)$ according to [4]. The $\boldsymbol{u}$ is the eigenvector with the largest eigenvalue (because the optimization is maximization) and the λ is the corresponding eigenvalue.

- The $\boldsymbol{u}$ is referred to as the **Fisher direction** or **Fisher axis**.

Fisher Subspace: Variant 1

- One possible solution to the generalized eigenvalue problem $(\boldsymbol{S}_B, \boldsymbol{S}_W)$ is [4]:

$$\boldsymbol{S}_B\, \boldsymbol{u} = \lambda\, \boldsymbol{S}_W\, \boldsymbol{u} \implies \boldsymbol{S}_W^{-1} \boldsymbol{S}_B\, \boldsymbol{u} = \lambda\, \boldsymbol{u} \\ \implies \boldsymbol{u} = \mathbf{eig}(\boldsymbol{S}_W^{-1} \boldsymbol{S}_B), \tag{23}$$

  where **eig**(.) denotes the eigenvector of the matrix with the largest eigenvalue. Although the solution in Eq. (23) is a little dirty [4] because $\boldsymbol{S}_w$ might be singular and not invertible, but this solution is very common for FDA.

- In some researches, the diagonal of $\boldsymbol{S}_W$ is strengthened slightly to make it full rank and invertible [4]:

$$\boldsymbol{u} = \mathbf{eig}((\boldsymbol{S}_W + \varepsilon \boldsymbol{I})^{-1} \boldsymbol{S}_B), \tag{24}$$

  where ε is a very small positive number, large enough to make $\boldsymbol{S}_W$ full rank.

Projection and Reconstruction in FDA

- The projection, projection of out-of-sample, reconstruction, and reconstruction of out-of-sample in SPCA are:

$$\begin{cases} \widetilde{\boldsymbol{x}} = \boldsymbol{U}^\top \boldsymbol{x}, & (25) \\ \widetilde{\boldsymbol{x}}_t = \boldsymbol{U}^\top \boldsymbol{x}_t, & (26) \\ \widehat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} = \boldsymbol{U}\widetilde{\boldsymbol{x}}, & (27) \\ \widehat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_t = \boldsymbol{U}\widetilde{\boldsymbol{x}}_t, & (28) \end{cases}$$

  respectively.
- In FDA, there is no need to center the data, in contrast to PCA.

Fisher Subspace: Variant 2

- Another way to find the FDA direction is to consider another version of Fisher criterion. According to Eq. (15) for $\boldsymbol{S}_B$, the Fisher criterion becomes [1]:

$$f(\boldsymbol{u}) = \frac{\boldsymbol{u}^\top \boldsymbol{S}_B\, \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}} \overset{(15)}{=} \frac{\boldsymbol{u}^\top (\boldsymbol{S}_T - \boldsymbol{S}_W)\, \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}} = \frac{\boldsymbol{u}^\top \boldsymbol{S}_T\, \boldsymbol{u} - \boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}} = \frac{\boldsymbol{u}^\top \boldsymbol{S}_T\, \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u}} - \cancel{1}. \tag{29}$$

- The -1 is a constant and is dropped in the optimization; therefore:

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{maximize}} \quad & \boldsymbol{u}^\top \boldsymbol{S}_T\, \boldsymbol{u} \\ \text{subject to} \quad & \boldsymbol{u}^\top \boldsymbol{S}_W\, \boldsymbol{u} = 1, \end{aligned} \tag{30}$$

whose solution is similarly obtained as:

$$\underbrace{\boldsymbol{S}_T\, \boldsymbol{u} = \lambda\, \boldsymbol{S}_W\, \boldsymbol{u},} \tag{31}$$

which is a generalized eigenvalue problem $\underbrace{(\boldsymbol{S}_T, \boldsymbol{S}_W)}$ according to [4].

Multi-dimensional Subspace

Multi-dimensional Subspace

- In case the Fisher subspace is the span of several Fisher directions, $\{\boldsymbol{u}_j\}_{j=1}^{p}$ where $\boldsymbol{u}_j \in \mathbb{R}^d$, the d_B and d_W are defined as:

$$\mathbb{R} \ni d_B := \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_B\, \boldsymbol{U}), \tag{32}$$

$$\mathbb{R} \ni d_W := \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_W\, \boldsymbol{U}), \tag{33}$$

where $\mathbb{R}^{d\times p} \ni \boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_p]$. In this case, maximizing the *Fisher criterion* is:

$$\underset{\boldsymbol{U}}{\text{maximize}} \quad f(\boldsymbol{U}) := \frac{d_B(\boldsymbol{U})}{d_W(\boldsymbol{U})} = \frac{\mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_B\, \boldsymbol{U})}{\mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_W\, \boldsymbol{U})}. \tag{34}$$

- The Fisher criterion $f(\boldsymbol{U})$ is a generalized Rayleigh-Ritz quotient (see preliminaries). According to preliminaries, the optimization in Eq. (34) is **approximately** equivalent to:

$$\begin{cases} \underset{\boldsymbol{U}}{\text{maximize}} & \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_B\, \boldsymbol{U}) \\ \text{subject to} & \boldsymbol{U}^\top \boldsymbol{S}_W\, \boldsymbol{U} = \boldsymbol{I}. \end{cases} \tag{35}$$

- Note that it is exactly true for one projection vector $\boldsymbol{u}$ but it approximately holds for the projection matrix $\boldsymbol{U}$ having multiple projection directions.

Multi-dimensional Subspace

- The Lagrangian [3] is:

$$\mathcal{L} = \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_B\, \boldsymbol{U}) - \mathbf{tr}\big(\boldsymbol{\Lambda}^\top (\boldsymbol{U}^\top \boldsymbol{S}_W\, \boldsymbol{U} - \boldsymbol{I})\big),$$

  where $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose diagonal entries are the Lagrange multipliers. Equating the derivative of $\mathcal{L}$ to zero gives:

$$\begin{aligned} \mathbb{R}^{d \times p} \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} &= 2\, \boldsymbol{S}_B\, \boldsymbol{U} - 2\, \boldsymbol{S}_W\, \boldsymbol{U}\boldsymbol{\Lambda} \overset{\text{set}}{=} \mathbf{0} \\ \implies 2\, \boldsymbol{S}_B\, \boldsymbol{U} &= 2\, \boldsymbol{S}_W\, \boldsymbol{U}\boldsymbol{\Lambda} \implies \boldsymbol{S}_B\, \boldsymbol{U} = \boldsymbol{S}_W\, \boldsymbol{U}\boldsymbol{\Lambda}, \end{aligned} \tag{36}$$

  which is a generalized eigenvalue problem $(\boldsymbol{S}_B, \boldsymbol{S}_W)$ according to [4]. The columns of $\boldsymbol{U}$ are the eigenvectors sorted by largest to smallest eigenvalues (because the optimization is maximization) and the diagonal entries of $\boldsymbol{\Lambda}$ are the corresponding eigenvalues.
- The columns of $\boldsymbol{U}$ are referred to as the **Fisher directions** or **Fisher axes**.

Multi-dimensional Subspace

- One possible solution to the generalized eigenvalue problem $(\boldsymbol{S}_B, \boldsymbol{S}_W)$ is [4]:

$$
\begin{aligned}
\boldsymbol{S}_B\, \boldsymbol{U} = \boldsymbol{S}_W\, \boldsymbol{U}\boldsymbol{\Lambda} &\implies \overbrace{\boldsymbol{S}_W^{-1}\boldsymbol{S}_B}\, \boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \\
&\implies \boldsymbol{U} = \mathbf{eig}(\boldsymbol{S}_W^{-1}\boldsymbol{S}_B),
\end{aligned}
\tag{37}
$$

  where $\mathbf{eig}(.)$ denotes the eigenvectors of the matrix stacked column-wise. Again, we can have [4]:

$$
\boldsymbol{U} = \mathbf{eig}((\underbrace{\boldsymbol{S}_W + \varepsilon \boldsymbol{I}})^{-1}\boldsymbol{S}_B). \tag{38}
$$

Projection and Reconstruction in FDA

- The projection, projection of out-of-sample, reconstruction, and reconstruction of out-of-sample in SPCA are:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{U}^\top \boldsymbol{X}, \tag{39}$$

$$\widetilde{\boldsymbol{X}}_t = \boldsymbol{U}^\top \boldsymbol{X}_t, \tag{40}$$

$$\widehat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X} = \boldsymbol{U}\widetilde{\boldsymbol{X}}, \tag{41}$$

$$\widehat{\boldsymbol{X}}_t = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X}_t = \boldsymbol{U}\widetilde{\boldsymbol{X}}_t, \tag{42}$$

  respectively.
- In FDA, there is no need to center the data, in contrast to PCA.

Discussion on Dimensionality of the Fisher Subspace

Discussion on Dimensionality of the Fisher Subspace

- In general, the rank of a covariance (scatter) matrix over the d-dimensional data with sample size n is at most $\min(d, n-1)$. The d is because the covariance matrix is a $d \times d$ matrix and the n is because we iterate over n data instances for calculating the covariance matrix. The -1 is because of subtracting the mean in calculation of the covariance matrix.
- For clarification, assume we only have one instance which becomes zero after removing the mean. This makes the covariance matrix a zero matrix.
- According to Eq. (10), the rank of the $\boldsymbol{S}_W$ is at most $\min(d, n-1)$ because all the instances of all the classes are considered. Hence, the rank of $\boldsymbol{S}_W$ is also at most $\min(d, n-1)$. According to Eq. (6), the rank of the $\boldsymbol{S}_B$ is at most $\min(d, c-1)$ because we have c iterations in its calculation.

$$\mathbb{R}^{d\times d} \ni \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

- In Eq. (37), we have $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ whose rank is:

$$\begin{aligned} \textbf{rank}(\boldsymbol{S}_W^{-1}\boldsymbol{S}_B) &\leq \min\big(\textbf{rank}(\boldsymbol{S}_W^{-1}), \textbf{rank}(\boldsymbol{S}_B)\big) \\ &\leq \min\big(\min(d, n-1), \min(d, c-1)\big) \\ &= \min(d, n-1, c-1) \overset{(a)}{=} c-1, \end{aligned} \tag{43}$$

  where (a) is because we usually have $c < d, n$. Therefore, the rank of $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$ is limited because of the rank of $\boldsymbol{S}_B$ which is at most $c-1$.
- According to Eq. (37), the $c-1$ leading eigenvalues will be valid and the rest are zero or very small. Therefore, the p, which is the dimensionality of the Fisher subspace, is at most $c-1$. The $c-1$ leading eigenvectors are considered as the Fisher directions and the rest of eigenvectors are invalid and ignored.

Comparison of FDA and PCA Directions

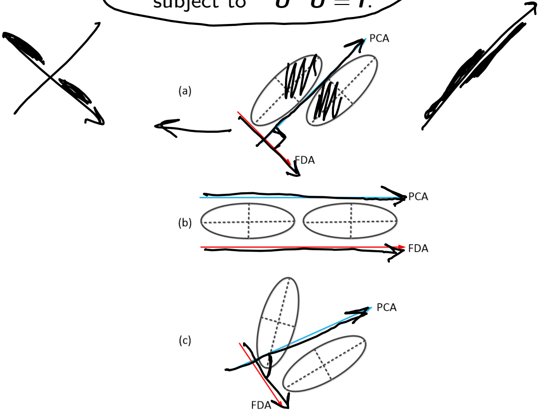
Comparison of FDA and PCA Directions

- FDA optimization:

$$\begin{aligned} \underset{\boldsymbol{U}}{\text{maximize}} \quad & \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_T\, \boldsymbol{U}) \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{S}_W \boldsymbol{U} = \boldsymbol{I}. \end{aligned} \tag{44}$$

- PCA optimization: [5]:

$$\begin{aligned} \underset{\boldsymbol{U}}{\text{maximize}} \quad & \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}_T\, \boldsymbol{U}) \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}. \end{aligned} \tag{45}$$



FDA $\overset{?}{\equiv}$ LDA

FDA $\stackrel{?}{\equiv}$ LDA

- The FDA is also referred to as **Linear Discriminant Analysis (LDA)** and **Fisher LDA (FLDA)**.
- Note that FDA is a manifold (subspace) learning method and LDA [6] is a classification method. However, LDA can be seen as a metric learning method [6] and as metric learning is a subspace learning method, there is a connection between FDA and LDA.
- We know that FDA is a projection-based subspace learning method. Consider the projection vector $\boldsymbol{u}$. The projection of data $\boldsymbol{x}$ is:

$$\boldsymbol{x} \mapsto \boldsymbol{u}^\top \boldsymbol{x}, \tag{46}$$

which can be done for all the data instances of every class. Thus, the mean and the covariance matrix of the class are transformed as:

$$\boldsymbol{\mu} \mapsto \boldsymbol{u}^\top \boldsymbol{\mu}, \tag{47}$$

$$\boldsymbol{\Sigma} \mapsto \boldsymbol{u}^\top \boldsymbol{\Sigma}\, \boldsymbol{u}, \tag{48}$$

$(u^\top \mu_2 - \mu_1^\top u)(u^\top \mu_2 - \mu_1^\top u)$

respectively, because of characteristics of mean and variance.

- According to Eq. (19), the Fisher criterion is the ratio of the between-class variance, σ_b^2, and within-class variance, σ_w^2:

$$f := \frac{\sigma_b^2}{\sigma_w^2} = \frac{(\boldsymbol{u}^\top \boldsymbol{\mu}_2 - \boldsymbol{u}^\top \boldsymbol{\mu}_1)^2}{\boldsymbol{u}^\top \boldsymbol{\Sigma}_2\, \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{\Sigma}_1\, \boldsymbol{u}} = \frac{(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1))^2}{\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}}, \tag{49}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the means of the two classes and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the covariances of the two classes.

FDA $\stackrel{?}{\equiv}$ LDA

- We had:

$$f := \frac{\sigma_b^2}{\sigma_w^2} = \frac{(\boldsymbol{u}^\top \boldsymbol{\mu}_2 - \boldsymbol{u}^\top \boldsymbol{\mu}_1)^2}{\boldsymbol{u}^\top \boldsymbol{\Sigma}_2\, \boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{\Sigma}_1\, \boldsymbol{u}} = \frac{\big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2}{\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}}.$$

- The FDA maximizes the Fisher criterion:

$$\underset{\boldsymbol{u}}{\text{maximize}} \quad \frac{\big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2}{\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u}}, \tag{50}$$

which can be restated as:

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{maximize}} \quad & \big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2, \\ \text{subject to} \quad & \boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u} = 1, \end{aligned} \tag{51}$$

according to Rayleigh-Ritz quotient method [7].

- The Lagrangian [3] is:

$$\mathcal{L} = \big(\boldsymbol{u}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^2 - \lambda \big(\boldsymbol{u}^\top (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\, \boldsymbol{u} - 1\big),$$

where λ is the Lagrange multiplier.

FDA $\stackrel{?}{\equiv}$ LDA

- Equating the derivative of $\mathcal{L}$ to zero gives:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{u}} = 2\,(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{u} - 2\,\lambda\,(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\,\boldsymbol{u} \stackrel{\text{set}}{=} \mathbf{0}$$

$$\implies (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{u} = \lambda\,(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\,\boldsymbol{u},$$

which is a generalized eigenvalue problem $\big((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top, (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)\big)$ according to [4].

- The projection vector is the eigenvector of $(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top$; therefore, we can say:

$$\boldsymbol{u} \propto (\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top. \tag{52}$$

- On the other hand, in LDA, the decision function is [6]:

$$2\,\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big)^\top \boldsymbol{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + 2\ln(\frac{\pi_1}{\pi_2}) = 0, \tag{53}$$

where π_1 and π_2 are the prior distributions of the two classes. Moreover, in LDA, the covariance matrices are assumed to be equal [6]: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Therefore, in LDA, the Eq. (52) becomes [6]:

$$\begin{aligned} \boldsymbol{u} &\propto (\cancel{2}\,\boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \\ &\propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top. \end{aligned} \tag{54}$$

According to Eq. (46), we have:

$$\boldsymbol{u}^\top \boldsymbol{x} \propto \big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top\big)^\top \boldsymbol{x}. \tag{55}$$

FDA $\stackrel{?}{\equiv}$ LDA

- Comparing Eq. (53) and Eq. (55):

$$\underbrace{2\,(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1))^\top \boldsymbol{x} + \overbrace{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+2\ln(\frac{\pi_1}{\pi_2})}}\,=0,$$

$$\boldsymbol{u}^\top\boldsymbol{x} \propto \underbrace{(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top)^\top\boldsymbol{x}},$$

  shows that LDA and FDA are equivalent up to a scaling factor $\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+2\pi_1/\pi_2$.
- Note that this term is multiplied as an exponential factor before taking logarithm to obtain Eq. (53), so this term is a scaling factor (see the LDA lecture or [6] for more details).
- It should be noted that in manifold (subspace) learning, the scale does not matter because all the distances can scale similarly in the subspace, without impacting the relative distances of points.
- Hence, we can say that LDA and FDA are equivalent:

$$\text{LDA} \equiv \text{FDA}. \tag{56}$$

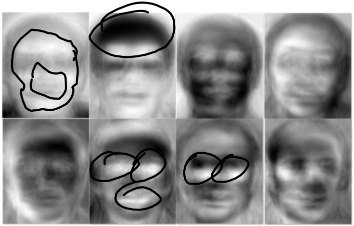
  Therefore, **the two subspaces of FDA and LDA are the same subspace.**
- In other words, **FDA followed by the use of Euclidean distance for classification in the subspace is equivalent to LDA.** This sheds light on why LDA and FDA are used interchangeably in the literature.
- Note that LDA assumes *one* (and not several) Gaussian for every class [6] and so does the FDA because they are equivalent. That is why FDA faces problem for multi-modal data [8].

Eigenfaces vs. Fisherfaces

Eigenfaces vs. Fisherfaces

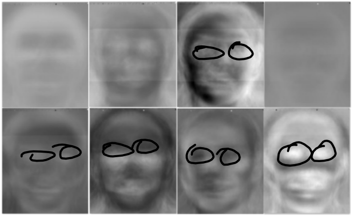
gost faces

eigenfaces (1991) [9, 10] and *Fisherfaces* (1997) [11, 12, 13]



eigenfaces

u vectors (eigenvectors of S) (solution of PCA)



Fisherfaces

u vectors in FDA

at most (c−1) vectors u

Kernel Fisher Discriminant Analysis

Kernel Fisher Discriminant Analysis

- The Eq. (3) in the feature space is:

$$\mathbb{R}^{t\times t} \ni \boldsymbol{\Phi}(\boldsymbol{S}_B) := \big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)\big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)^\top, \tag{57}$$

where the mean of the j-th class in the feature space is:

$$\mathbb{R}^t \ni \phi(\boldsymbol{\mu}_j) := \underbrace{\frac{1}{n_j}\sum_{i=1}^{n_j} \phi(\boldsymbol{x}_i^{(j)})}. \tag{58}$$

- According to the **representation theory** [14], **any solution (direction) $\phi(\boldsymbol{u}) \in \mathcal{H}$ must lie in the span of "all" the training vectors mapped to $\mathcal{H}$, i.e., $\boldsymbol{\Phi}(\boldsymbol{X}) = [\phi(\boldsymbol{x}_1), \dots, \phi(\boldsymbol{x}_n)] \in \mathbb{R}^{t\times n}$ (usually $t \gg d$).** Note that $\mathcal{H}$ denotes the Hilbert space (feature space). Therefore, we can state that:

$$\mathbb{R}^t \ni \phi(\boldsymbol{u}) = \underbrace{\sum_{i=1}^{n} \theta_i \phi(\boldsymbol{x}_i)} = \underbrace{\boldsymbol{\Phi}(\boldsymbol{X})\,\boldsymbol{\theta}}, \tag{59}$$

where $\mathbb{R}^n \ni \underbrace{\boldsymbol{\theta} := [\theta_1, \dots, \theta_n]^\top}$ is the unknown vector of coefficients, and $\phi(\boldsymbol{u}) \in \mathbb{R}^t$ is the pulled Fisher direction to the feature space.

- The pulled directions can be put together in $\mathbb{R}^{t\times p} \ni \boldsymbol{\Phi}(\boldsymbol{U}) := [\phi(\boldsymbol{u}_1), \dots, \phi(\boldsymbol{u}_p)]$:

$$\underbrace{\mathbb{R}^{t\times p}} \ni \boxed{\boldsymbol{\Phi}(\boldsymbol{U}) = \boldsymbol{\Phi}(\boldsymbol{X})\,\boldsymbol{\Theta}}, \tag{60}$$

where $\underbrace{\boldsymbol{\Theta} := [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p] \in \mathbb{R}^{n\times p}}$.

$\Phi(U)^\top \Phi(X) = \Theta^\top \Phi(X)^\top \Phi(X)$

Kernel Fisher Discriminant Analysis

- The d_B in the feature space is:

$$\mathbb{R} \ni d_B := \phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_B)\,\phi(\boldsymbol{u}) \tag{61}$$

$$\overset{(a)}{=} \boldsymbol{\theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)\big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)^\top \boldsymbol{\Phi}(\boldsymbol{X})\,\boldsymbol{\theta}, \tag{62}$$

where (a) is because of Eqs. (57). and (59).

- For the j-th class (here $j \in \{1, 2\}$), we have:

$$\boldsymbol{\theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \phi(\boldsymbol{\mu}_j) \overset{(59)}{=} \sum_{i=1}^{n} \theta_i\, \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{\mu}_j) \overset{(58)}{=} \frac{1}{n_j} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \theta_i\, \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_k^{(j)})$$

$$= \frac{1}{n_j} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \theta_i\, k(\boldsymbol{x}_i, \boldsymbol{x}_k^{(j)}) = \boldsymbol{\theta}^\top \boldsymbol{m}_j, \tag{63}$$

where $\boldsymbol{m}_j \in \mathbb{R}^n$ whose i-th entry is:

$$\boldsymbol{m}_j(i) := \frac{1}{n_j} \sum_{k=1}^{n_j} k(\boldsymbol{x}_i, \boldsymbol{x}_k^{(j)}). \tag{64}$$

Kernel Fisher Discriminant Analysis

- We had:

$$d_B = \boldsymbol{\theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)\big(\phi(\boldsymbol{\mu}_1) - \phi(\boldsymbol{\mu}_2)\big)^\top \boldsymbol{\Phi}(\boldsymbol{X})\, \boldsymbol{\theta},$$

$$\boldsymbol{\theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \phi(\boldsymbol{\mu}_j) = \frac{1}{n_j} \sum_{i=1}^{n} \sum_{k=1}^{n_j} \theta_i\, k(\boldsymbol{x}_i, \boldsymbol{x}_k^{(j)}) = \boldsymbol{\theta}^\top \boldsymbol{m}_j,$$

$$\boldsymbol{m}_j(i) := \frac{1}{n_j} \sum_{k=1}^{n_j} k(\boldsymbol{x}_i, \boldsymbol{x}_k^{(j)}).$$

- Hence, Eq. (62) becomes:

$$d_B \overset{(63)}{=} \boldsymbol{\theta}^\top \overbrace{(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^\top}^{M} \boldsymbol{\theta} = \boldsymbol{\theta}^\top \boldsymbol{M} \boldsymbol{\theta}, \tag{65}$$

where:

$$\mathbb{R}^{n \times n} \ni \boldsymbol{M} := (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^\top, \tag{66}$$

is the **between-scatter** in kernel FDA. Hence, the Eq. (62) becomes:

$$d_B = \phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_B)\, \phi(\boldsymbol{u}) = \boldsymbol{\theta}^\top \boldsymbol{M} \boldsymbol{\theta}. \tag{67}$$

Kernel Fisher Discriminant Analysis

- The Eq. (10) in the feature space is:

$$\mathbb{R}^{t\times t} \ni \boldsymbol{\Phi}(\boldsymbol{S}_W) := \sum_{j=1}^{c}\sum_{i=1}^{n_j}\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)^\top. \quad (68)$$

- The d_W in the feature space is:

$$\mathbb{R} \ni d_W := \phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_W)\,\phi(\boldsymbol{u})$$

$$\overset{(a)}{=} \Big(\sum_{\ell=1}^{n}\theta_\ell\,\phi(\boldsymbol{x}_\ell)^\top\Big)\Big(\sum_{j=1}^{c}\sum_{i=1}^{n_j}\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)^\top\Big)\Big(\sum_{k=1}^{n}\theta_k\,\phi(\boldsymbol{x}_k)\Big)$$

$$= \sum_{j=1}^{c}\sum_{\ell=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}\Big(\theta_\ell\,\phi(\boldsymbol{x}_\ell)^\top\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)\big(\phi(\boldsymbol{x}_i^{(j)}) - \phi(\boldsymbol{\mu}_j)\big)^\top\theta_k\,\phi(\boldsymbol{x}_k)\Big)$$

$$\overset{(58)}{=} \sum_{j=1}^{c}\sum_{\ell=1}^{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n}\Big(\theta_\ell\,\phi(\boldsymbol{x}_\ell)^\top\big(\phi(\boldsymbol{x}_i^{(j)}) - \frac{1}{n_j}\sum_{e=1}^{n_j}\phi(\boldsymbol{x}_e^{(j)})\big)$$

$$\big(\phi(\boldsymbol{x}_i^{(j)}) - \frac{1}{n_j}\sum_{z=1}^{n_j}\phi(\boldsymbol{x}_z^{(j)})\big)^\top\theta_k\,\phi(\boldsymbol{x}_k)\Big)$$

Kernel Fisher Discriminant Analysis

$\Phi(x_i)^T \Phi(x_k)$
$= \Phi(x_k)^T \Phi(x_i)$

$$= \sum_{j=1}^{c} \sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\theta_\ell\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)}) - \frac{1}{n_j} \sum_{e=1}^{n_j} \theta_\ell\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_e^{(j)})\Big)$$

$$\Big(\theta_k\, k(\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_k) - \frac{1}{n_j} \sum_{z=1}^{n_j} \theta_k\, k(\boldsymbol{x}_z^{(j)}, \boldsymbol{x}_k)\Big)$$

$$\overset{(b)}{=} \sum_{j=1}^{c} \sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\theta_\ell\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)}) - \frac{1}{n_j} \sum_{e=1}^{n_j} \theta_\ell\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_e^{(j)})\Big)\Big(\theta_k\, k(\boldsymbol{x}_k, \boldsymbol{x}_i^{(j)}) - \frac{1}{n_j} \sum_{z=1}^{n_j} \theta_k\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)})\Big)$$

$$= \sum_{j=1}^{c} \sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\theta_\ell\, \theta_k\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_i^{(j)}) - \frac{2\,\theta_\ell\, \theta_k}{n_j} \sum_{z=1}^{n_j} k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)})$$

$$+ \frac{\theta_\ell\, \theta_k}{n_j^2} \sum_{e=1}^{n_j} \sum_{z=1}^{n_j} k(\boldsymbol{x}_\ell, \boldsymbol{x}_e^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)})\Big)$$

$$= \sum_{j=1}^{c} \sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\theta_\ell\, \theta_k\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_i^{(j)}) - \frac{\theta_\ell\, \theta_k}{n_j} \sum_{z=1}^{n_j} k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)})\Big)$$

Kernel Fisher Discriminant Analysis

$$
\begin{aligned}
&= \sum_{j=1}^{c} \sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\theta_\ell\, \theta_k\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_i^{(j)}) - \frac{\theta_\ell\, \theta_k}{n_j} \sum_{z=1}^{n_j} k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)}) \Big) \\
&= \sum_{j=1}^{c} \Big(\underbrace{\sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \big(\theta_\ell\, \theta_k\, k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_i^{(j)}) \big)}_{\star} \\
&\qquad - \underbrace{\sum_{\ell=1}^{n} \sum_{i=1}^{n_j} \sum_{k=1}^{n} \Big(\frac{\theta_\ell\, \theta_k}{n_j} \sum_{z=1}^{n_j} k(\boldsymbol{x}_\ell, \boldsymbol{x}_i^{(j)})\, k(\boldsymbol{x}_k, \boldsymbol{x}_z^{(j)}) \Big)}_{\star} \Big) \\
&\overset{(c)}{=} \sum_{j=1}^{c} \big(\underbrace{\boldsymbol{\theta}^\top \boldsymbol{K}_j \boldsymbol{K}_j^\top \boldsymbol{\theta}}_{\star} - \underbrace{\boldsymbol{\theta}^\top \boldsymbol{K}_j \frac{1}{n_j} \mathbf{1}\mathbf{1}^\top \boldsymbol{K}_j^\top \boldsymbol{\theta}}_{\star} \big) = \sum_{j=1}^{c} \boldsymbol{\theta}^\top \boldsymbol{K}_j \underbrace{\big(\boldsymbol{I} - \frac{1}{n_j} \mathbf{1}\mathbf{1}^\top \big)}_{\star} \boldsymbol{K}_j^\top \boldsymbol{\theta} \\
&\overset{(d)}{=} \sum_{j=1}^{c} \boldsymbol{\theta}^\top \boldsymbol{K}_j \boldsymbol{H}_j \boldsymbol{K}_j^\top \boldsymbol{\theta} = \boldsymbol{\theta}^\top \boxed{\Big(\sum_{j=1}^{c} \boldsymbol{K}_j \boldsymbol{H}_j \boldsymbol{K}_j^\top \Big)} \boldsymbol{\theta},
\end{aligned}
$$

where (a) is because of Eqs. (68) and (59), (b) is because $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k(\boldsymbol{x}_2, \boldsymbol{x}_1) \in \mathbb{R}$, and (c) is because $\boldsymbol{K}_j \in \mathbb{R}^{n \times n_j}$ is the kernel matrix of the whole training data and the training data of the j-th class. The (a, b)-th element of $\boldsymbol{K}_j$ is:

$$\boldsymbol{K}_j(a, b) := k(\boldsymbol{x}_a, \boldsymbol{x}_b^{(j)}). \tag{69}$$

Kernel Fisher Discriminant Analysis

- The (d) is because:

$$\mathbb{R}^{n_j \times n_j} \ni \boldsymbol{H}_j := \boldsymbol{I} - \frac{1}{n_j}\mathbf{1}\mathbf{1}^\top, \tag{70}$$

  is the **centering matrix**.
- We define:

$$\mathbb{R}^{n \times n} \ni \boldsymbol{N} := \sum_{j=1}^{c} \boldsymbol{K}_j \boldsymbol{H}_j \boldsymbol{K}_j^\top, \tag{71}$$

  as the **within-scatter** in kernel FDA. Hence, the d_W becomes:

$$d_W = \phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_W)\,\phi(\boldsymbol{u}) = \underbrace{\boldsymbol{\theta}^\top \boldsymbol{N}\boldsymbol{\theta}}. \tag{72}$$

- The **kernel Fisher criterion** is:

$$f(\boldsymbol{\theta}) := \frac{d_B(\boldsymbol{\theta})}{d_W(\boldsymbol{\theta})} = \frac{\phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_B)\,\phi(\boldsymbol{u})}{\phi(\boldsymbol{u})^\top \boldsymbol{\Phi}(\boldsymbol{S}_W)\,\phi(\boldsymbol{u})} = \underbrace{\frac{\boldsymbol{\theta}^\top \boldsymbol{M}\boldsymbol{\theta}}{\boldsymbol{\theta}^\top \boldsymbol{N}\boldsymbol{\theta}}}, \tag{73}$$

  where the $\boldsymbol{\theta} \in \mathbb{R}^n$ is the **kernel Fisher direction**.
- Similar to the solution of Eq. (19), the solution to maximization of Eq. (73) is:

$$\underbrace{\boldsymbol{M}\boldsymbol{\theta} = \lambda\, \boldsymbol{N}\boldsymbol{\theta},} \tag{74}$$

  which is a generalized eigenvalue problem $(\boldsymbol{M}, \boldsymbol{N})$ according to [4]. The $\boldsymbol{\theta}$ is the eigenvector with the largest eigenvalue (because the optimization is maximization) and the λ is the corresponding eigenvalue. The $\boldsymbol{\theta}$ is the **kernel Fisher direction** or **kernel Fisher axis**.

Kernel Fisher Discriminant Analysis

- Again, one possible solution to the generalized eigenvalue problem $(\boldsymbol{M}, \boldsymbol{N})$ is [4]:

$$\boldsymbol{\theta} = \textbf{eig}(\boldsymbol{N}^{-1}\boldsymbol{M}), \tag{75}$$

or [4]:

$$\boldsymbol{\theta} = \textbf{eig}((\boldsymbol{N} + \varepsilon \boldsymbol{I})^{-1}\boldsymbol{M}), \tag{76}$$

where **eig**(.) denotes the eigenvector of the matrix with the largest eigenvalue.

- The projection and reconstruction of the training data point $\boldsymbol{x}_i$ and the out-of-sample data point $\boldsymbol{x}_t$ are:

$$\mathbb{R} \ni \phi(\widetilde{\boldsymbol{x}}_i) = \phi(\boldsymbol{u})^\top \phi(\boldsymbol{x}_i) \overset{(59)}{=} \boldsymbol{\theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \phi(\boldsymbol{x}_i) = \boldsymbol{\theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_i), \tag{77}$$

$$\mathbb{R}^t \ni \phi(\widehat{\boldsymbol{x}}_i) = \phi(\boldsymbol{u})\phi(\boldsymbol{u})^\top \phi(\boldsymbol{x}_i) \overset{(59)}{=} \boldsymbol{\Phi}(\boldsymbol{X})\,\boldsymbol{\theta}\boldsymbol{\theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_i), \tag{78}$$

$$\mathbb{R} \ni \phi(\widetilde{\boldsymbol{x}}_t) = \boldsymbol{\theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_t), \tag{79}$$

$$\mathbb{R}^t \ni \phi(\widehat{\boldsymbol{x}}_t) = \boldsymbol{\Phi}(\boldsymbol{X})\,\boldsymbol{\theta}\boldsymbol{\theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_t). \tag{80}$$

- However, in reconstruction expressions, the $\boldsymbol{\Phi}(\boldsymbol{X})$ is not necessarily available; therefore, in kernel FDA, similar to kernel PCA [5], **reconstruction cannot be done.**
- For the whole training and out-of-sample data, the projections are:

$$\mathbb{R}^{1\times n} \ni \boldsymbol{\Phi}(\widetilde{\boldsymbol{X}}) = \boldsymbol{\theta}^\top \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}), \tag{81}$$

$$\mathbb{R}^{1\times n_t} \ni \boldsymbol{\Phi}(\widetilde{\boldsymbol{X}}_t) = \boldsymbol{\theta}^\top \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_t). \tag{82}$$

Kernel Fisher Discriminant Analysis

- In multi-dimensional kernel Fisher subspace, the within- and between-scatters are the same but the Fisher criterion is different. According to Eq. (60), the d_B and d_W are:

$$d_B = \textbf{tr}\big(\phi(\boldsymbol{U})^\top \boldsymbol{\Phi}(\boldsymbol{S}_B)\,\phi(\boldsymbol{U})\big) = \textbf{tr}(\underline{\boldsymbol{\Theta}}^\top \boldsymbol{M}\underline{\boldsymbol{\Theta}}), \quad (83)$$

$$d_W = \textbf{tr}\big(\phi(\boldsymbol{U})^\top \boldsymbol{\Phi}(\boldsymbol{S}_W)\,\phi(\boldsymbol{U})\big) = \textbf{tr}(\boldsymbol{\Theta}^\top \boldsymbol{N}\boldsymbol{\Theta}), \quad (84)$$

where $\mathbb{R}^{n\times p} \ni \boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p]$ and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{N} \in \mathbb{R}^{n\times n}$ are the between- and within-scatters, respectively, determined for either two-class or multi-class case.

- The Fisher criterion becomes:

$$f(\boldsymbol{\Theta}) := \frac{d_B(\boldsymbol{\Theta})}{d_W(\boldsymbol{\Theta})} = \frac{\textbf{tr}\big(\phi(\boldsymbol{U})^\top \boldsymbol{\Phi}(\boldsymbol{S}_B)\,\phi(\boldsymbol{U})\big)}{\textbf{tr}\big(\phi(\boldsymbol{U})^\top \boldsymbol{\Phi}(\boldsymbol{S}_W)\,\phi(\boldsymbol{U})\big)} = \frac{\textbf{tr}(\boldsymbol{\Theta}^\top \boldsymbol{M}\boldsymbol{\Theta})}{\textbf{tr}(\boldsymbol{\Theta}^\top \boldsymbol{N}\boldsymbol{\Theta})}, \quad (85)$$

where the columns of $\boldsymbol{\Theta}$ are the *kernel Fisher directions*.

- Similar to Eq. (34), the solution to maximization of this criterion is:

$$\boxed{\boldsymbol{M}\boldsymbol{\Theta} = \boldsymbol{N}\boldsymbol{\Theta}\boldsymbol{\Lambda},} \quad (86)$$

which is the generalized eigenvalue problem $(\boldsymbol{M}, \boldsymbol{N})$ according to [4]. The columns of $\boldsymbol{\Theta}$ are the eigenvectors sorted from the largest to smallest eigenvalues (because the optimization is maximization) and the diagonal entries of $\boldsymbol{\Lambda}$ are the corresponding eigenvalues.

Kernel Fisher Discriminant Analysis

- As mentioned before, in kernel FDA, we do not have reconstruction.
- The projection of the training data point $\boldsymbol{x}_i$ and the out-of-sample data point $\boldsymbol{x}_t$ are:

$$\mathbb{R}^p \ni \phi(\widetilde{\boldsymbol{x}}_i) = \boldsymbol{\Phi}(\boldsymbol{U})^\top \phi(\boldsymbol{x}_i) \overset{(60)}{=} \boldsymbol{\Theta}^\top \boldsymbol{\Phi}(\boldsymbol{X})^\top \phi(\boldsymbol{x}_i) = \boldsymbol{\Theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_i), \tag{87}$$

$$\mathbb{R}^p \ni \phi(\widetilde{\boldsymbol{x}}_t) = \boldsymbol{\Theta}^\top \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{x}_t). \tag{88}$$

- For the whole training and out-of-sample data, the projections are:

$$\mathbb{R}^{p \times n} \ni \boldsymbol{\Phi}(\widetilde{\boldsymbol{X}}) = \boldsymbol{\Theta}^\top \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}), \tag{89}$$

$$\mathbb{R}^{p \times n_t} \ni \boldsymbol{\Phi}(\widetilde{\boldsymbol{X}}_t) = \boldsymbol{\Theta}^\top \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_t). \tag{90}$$

Acknowledgment

- Some slides are based on our tutorial paper: "Fisher and kernel Fisher discriminant analysis: Tutorial" [15]
- Some slides of this slide deck are inspired by teachings of Prof. Ali Ghodsi at University of Waterloo, Department of Statistics and Ptof. Hoda Mohammadzade at Sharif University of Technology, Department of Electrical Engineering.
- The code of FDA in my GitHub page (in Python language): `https://github.com/bghojogh/Fisher-Discriminant-Analysis`
- FDA/LDA in sklearn: `https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html`

References

[1] M. Welling, "Fisher linear discriminant analysis," tech. rep., Department of Computer Science, University of Toronto, 2005.

[2] Y. Xu and G. Lu, "Analysis on Fisher discriminant criterion and linear separability of feature space," in *2006 International Conference on Computational Intelligence and Security*, vol. 2, pp. 1671–1676, IEEE, 2006.

[3] S. Boyd and L. Vandenberghe, *Convex optimization*.
Cambridge university press, 2004.

[4] B. Ghojogh, F. Karray, and M. Crowley, "Eigenvalue and generalized eigenvalue problems: Tutorial," *arXiv preprint arXiv:1903.11240*, 2019.

[5] B. Ghojogh and M. Crowley, "Unsupervised and supervised principal component analysis: Tutorial," *arXiv preprint arXiv:1906.03148*, 2019.

[6] B. Ghojogh and M. Crowley, "Linear and quadratic discriminant analysis: Tutorial," *arXiv preprint arXiv:1906.02590*, 2019.

[7] E. Croot, "The Rayleigh principle for finding eigenvalues," tech. rep., Georgia Institute of Technology, School of Mathematics, 2005.
Online: http://people.math.gatech.edu/∼ecroot/notes_linear.pdf, Accessed: March 2019.

References (cont.)

[8] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of machine learning research*, vol. 8, no. May, pp. 1027–1061, 2007.

[9] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[10] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591, IEEE, 1991.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 711–720, 1997.

[12] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America A*, vol. 14, no. 8, pp. 1724–1733, 1997.

[13] W. Zhao, R. Chellappa, and P. J. Phillips, *Subspace linear discriminant analysis for face recognition*.
Citeseer, 1999.

[14] J. L. Alperin, *Local representation theory: Modular representations as an introduction to the local representation theory of finite groups*, vol. 11.
Cambridge University Press, 1993.

References (cont.)

[15] B. Ghojogh, F. Karray, and M. Crowley, "Fisher and kernel fisher discriminant analysis: Tutorial," *arXiv preprint arXiv:1906.09436*, 2019.