Linear Discriminant Analysis

Presentation · January 2015

CITATIONS
4

READS
7,673

1 author:

Alaa Tharwat
Fachhochschule Bielefeld
120 PUBLICATIONS **5,368** CITATIONS

SEE PROFILE

Linear Discriminant Analysis: An Overview

Alaa Tharwat

Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial" AI Communications 30 (2017) 169-190
Email: engalaatharwat@hotmail.com

December 30, 2017

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- The goal for any dimensional reduction method is to reduce the dimensions of the original data for different purposes such as visualization, decrease CPU time, ..etc..
- Dimensionality reduction techniques are important in many applications related to machine learning, data mining, Bioinformatics, biometric and information retrieval.
- There are two types of dimensionality reduction methods, namely, supervised and unsupervised.
  - **Supervised** (e.g. LDA).
  - **Unsupervised** (e.g. PCA).

- The Linear Discriminant Analysis (LDA) technique is developed to transform the features into a lower dimensional space, which maximizes the ratio of the between-class variance to the within-class variance, thereby guaranteeing maximum class separability.

$$J(W) = \frac{W^T S_B \, W}{W^T S_W \, W} \tag{1}$$

- There are two types of LDA technique to deal with classes: *class-dependent* and *class-independent*.
  - In the **class-dependent** LDA, one separate lower dimensional space is calculated for each class to project its data on it,
  - In the **class-independent** LDA, each class will be considered as a separate class against the other classes. In this type, there is just one lower dimensional space for all classes to project their data on it.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- Given the original data matrix $X = \{x_1, x_2, \ldots, x_N\}$, where x_i represents the i^{th} sample, pattern, or observation and N is the total number of samples.
- Each sample is represented by M features ($x_i \in \mathcal{R}^M$). In other words, each sample is represented as a point in M-dimensional space.
- The data matrix is partitioned into c classes as follows, $X = [\omega_1, \omega_2, \ldots, \omega_c]$. Thus, Each class has n_i samples.
- The total number of samples (N) is calculated as follows, $N = \sum_{i=1}^{c} n_i$.

- The goal of the LDA technique is to project the original data matrix onto a lower dimensional space. To achieve this goal, three steps needed to be performed.
  1. The first step is to calculate the separability between different classes (i.e. the distance between the means of different classes), which is called the **between-class variance** or **between-class matrix**.
  2. The second step is to calculate the distance between the mean and the samples of each class, which is called the **within-class variance** or **within-class matrix**.
  3. The third step is to construct the **lower dimensional space** which maximizes the between-class variance and minimizes the within-class variance.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- To calculate the between-class variance (S_B), the separation distance between different classes which is denoted by $(m_i - m)$ will be calculated as follows,

$$(m_i - m)^2 = (W^T \mu_i - W^T \mu)^2 = W^T (\mu_i - \mu)(\mu_i - \mu)^T W \quad (2)$$

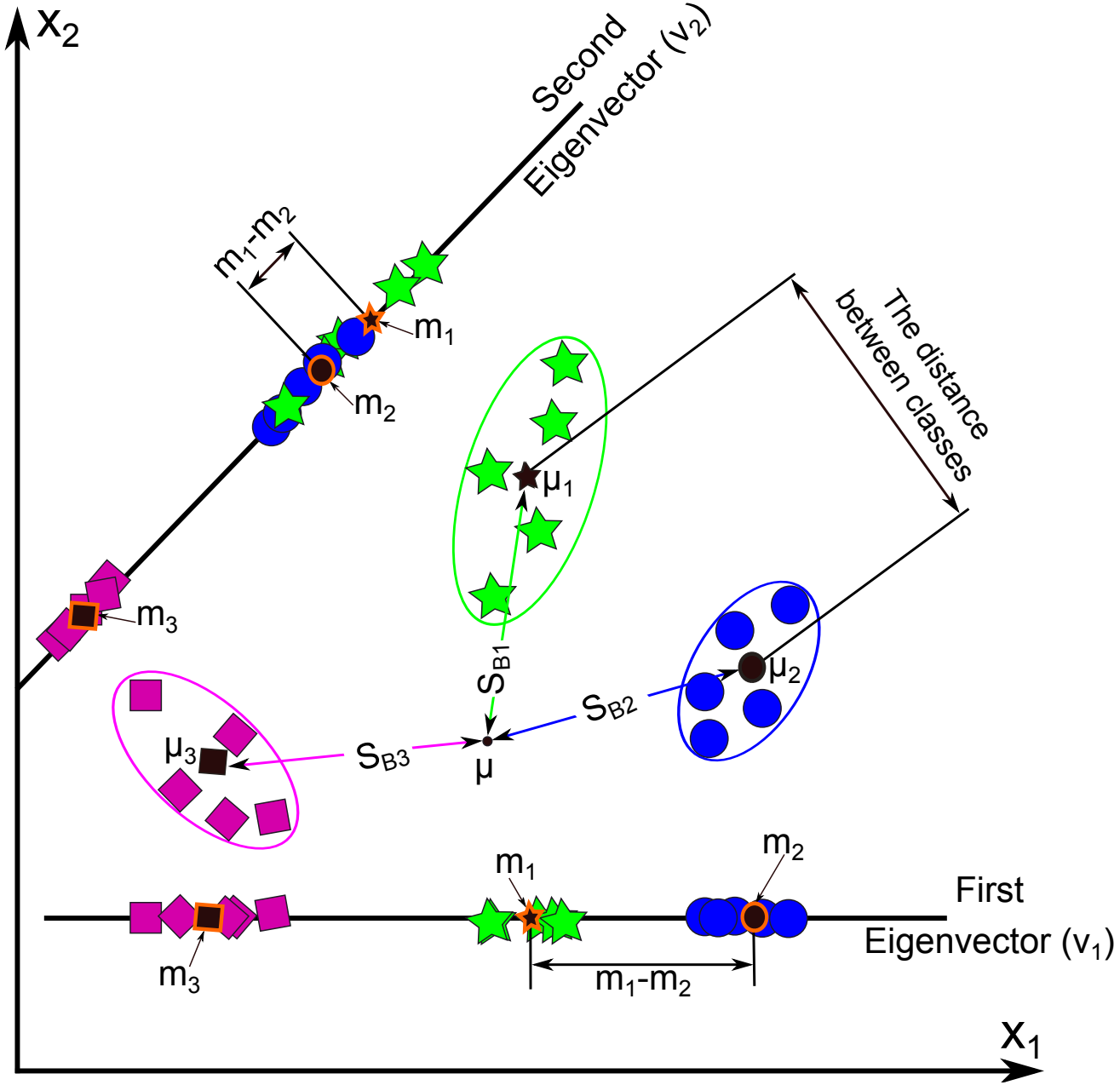
where,

- m_i represents the projection of the mean of the i^{th} class and it is calculated as follows, $m_i = W^T \mu_i$,
- m is the projection of the total mean of all classes and it is calculated as follows, $m = W^T \mu$,
- W represents the transformation matrix of LDA,
- $\mu_j (1 \times M)$ represents the mean of the i^{th} class ($\mu_j = \frac{1}{n_j} \sum_{x_i \in \omega_j} x_i$), and
- $\mu (1 \times M)$ is the total mean of all classes ($\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{c} \sum_{j=1}^{c} \mu_j$), where c represents the total number of classes.

- The term $(\mu_i - \mu)(\mu_i - \mu)^T$ in Equation (2) represents the separation distance between the mean of the i^{th} class (μ_i) and the total mean (μ), or simply it represents the between-class variance of the i^{th} class (S_{B_i}).
- Substitute S_{B_i} into Equation (2) as follows,

$$(m_i - m)^2 = W^T S_{Bi}\, W \tag{3}$$

- The total between-class variance is calculated as follows, ($S_B = \sum_{i=1}^{c} n_i S_{B_i}$).


X2
Second
Eigenvector (v2)
m1-m2
m1
m2
The distance
between classes
μ1
m3
SB1
SB2
μ2
μ3
SB3
μ
m1
m2
First
Eigenvector (v1)
m3
m1-m2
X1

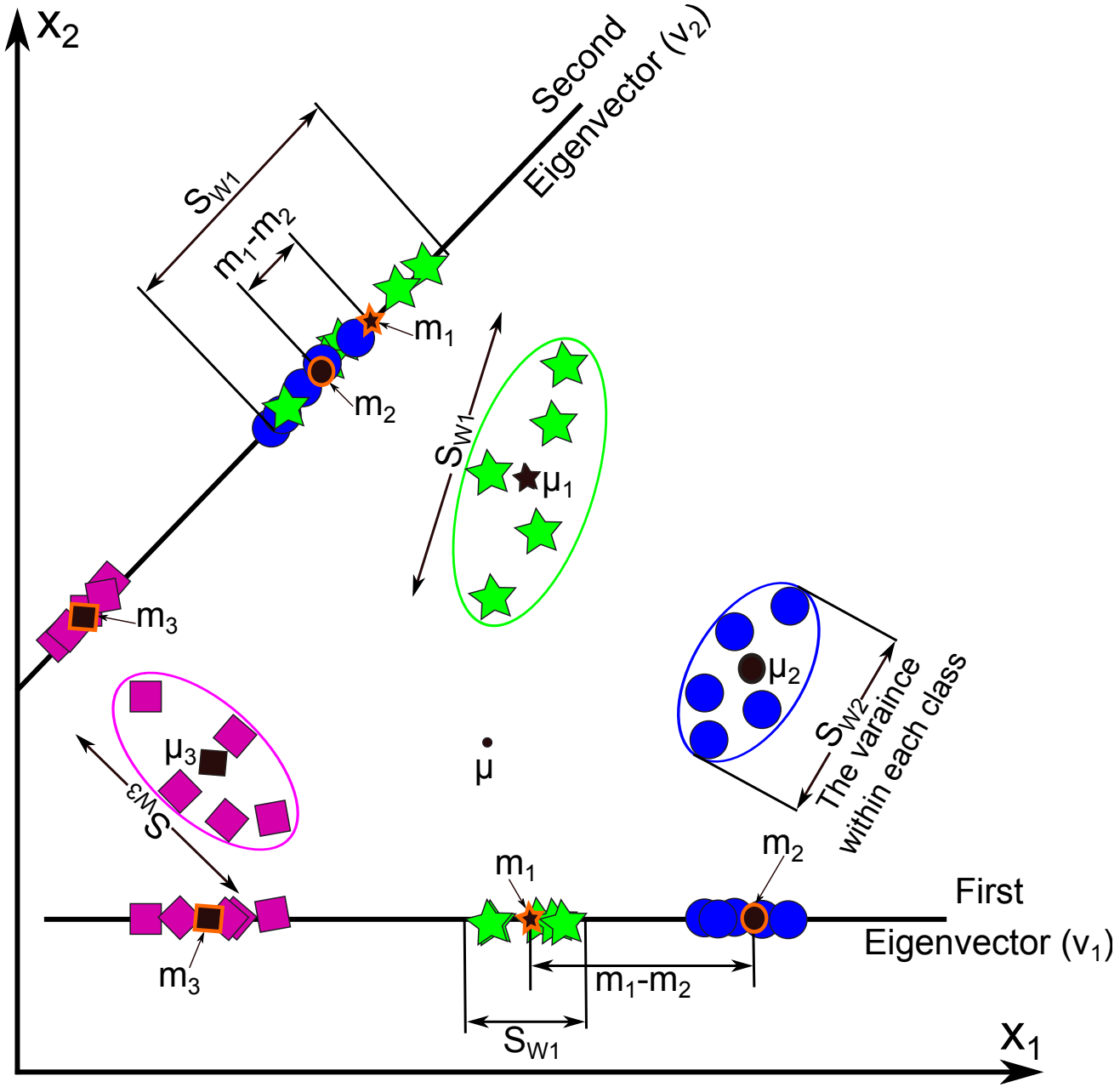
- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- The within-class variance of the i^{th} class (S_{W_i}) represents the difference between the mean and the samples of that class.
- LDA technique searches for a lower-dimensional space, which is used to minimize the difference between the projected mean (m_i) and the projected samples of each class ($W^T x_i$), or simply minimizes the within-class variance.

$$\begin{aligned}
&\sum_{x_i \in \omega_j} (W^T x_i - m_j)^2 \\
&= \sum_{x_i \in \omega_j} (W^T x_i - W^T \mu_j)^2 \\
&= \sum_{x_i \in \omega_j} W^T (x_i - \mu_j)^2 W \\
&= \sum_{x_i \in \omega_j} W^T (x_i - \mu_j)(x_i - \mu_j)^T W \\
&= \sum_{x_i \in \omega_j} W^T S_{W_j} W
\end{aligned} \tag{4}$$

- From Equation (4), the within-class variance for each class can be calculated as follows, $S_{W_j} = d_j^T * d_j = \sum_{i=1}^{n_j}(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$, where x_{ij} represents the i^{th} sample in the j^{th} class as shown in Fig. (2, step (**E, F**)), and d_j is the centering data of the j^{th} class, i.e. $d_j = \omega_j - \mu_j = \{x_i\}_{i=1}^{n_j} - \mu_j$.
- Step (**F**) in the figure illustrates how the within-class variance of the first class (S_{W_1}) in our example is calculated.
- The total within-class variance represents the sum of all within-class matrices of all classes (see Fig. (2, step (**F**))), and it can be calculated as in Equation (5).

$$S_W = \sum_{i=1}^{c} S_{W_i} = \sum_{x_i \in \omega_1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{x_i \in \omega_2} (x_i - \mu_2)(x_i - \mu_2)^T + \cdots + \sum_{x_i \in \omega_c} (x_i - \mu_c)(x_i - \mu_c)^T \quad (5)$$


x2
Second
Eigenvector (v2)
SW1
m1-m2
m1
m2
SW1
μ1
m3
μ2
SW2
The varaince
within each class
μ3
μ
SW3
m2
m1
First
Eigenvector (v1)
m3
m1-m2
SW1
x1

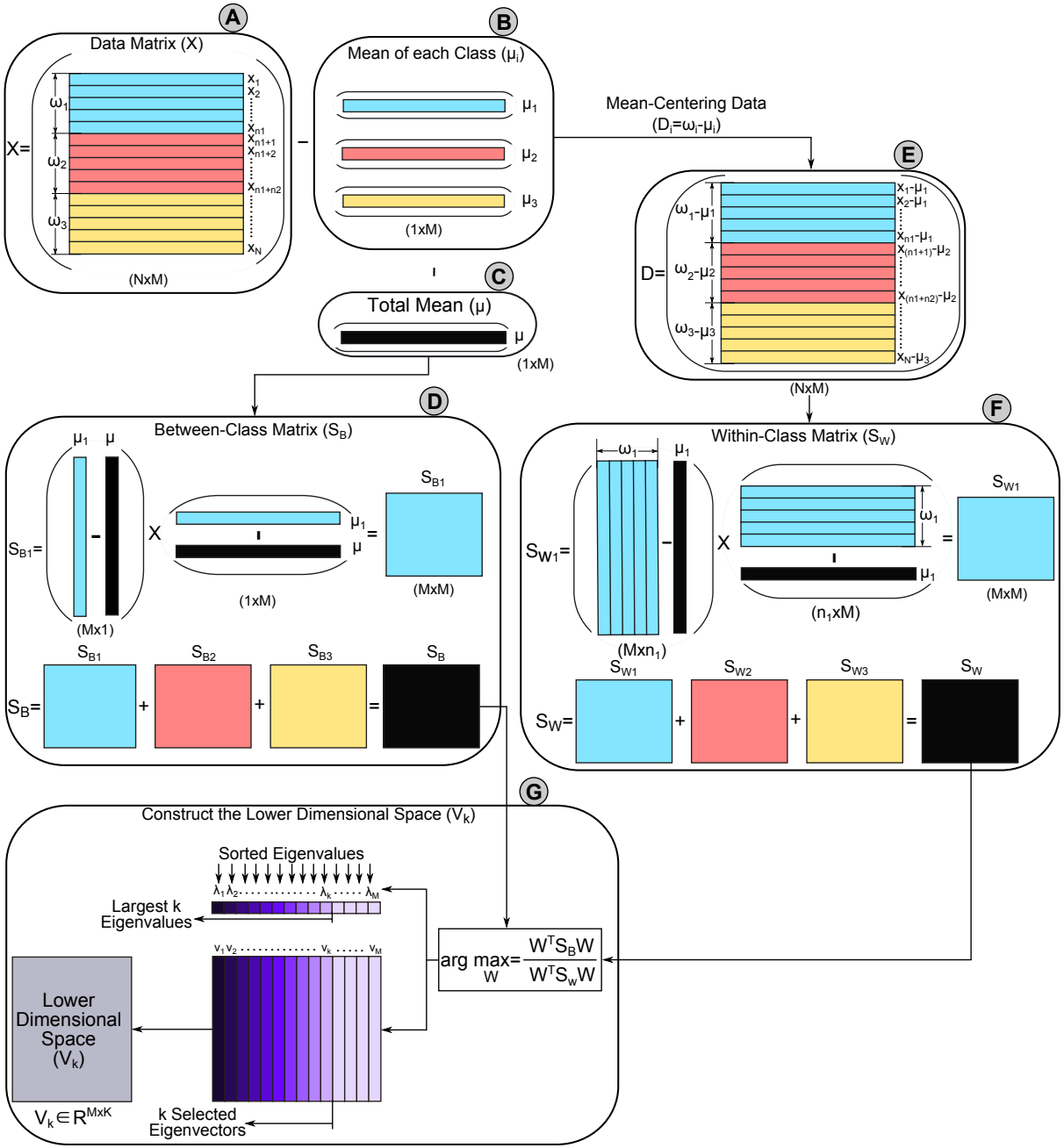
- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- After calculating the between-class variance (S_B) and within-class variance (S_W), the transformation matrix (W) of the LDA technique can be calculated as in Equation (1), which can be reformulated as an optimization problem as in Equation (6).

$$S_W \, W = \lambda S_B \, W \tag{6}$$

- where λ represents the eigenvalues of the transformation matrix (W).
- The solution of this problem can be obtained by calculating the eigenvalues (λ) and eigenvectors ($V = \{v_1, v_2, \ldots, v_M\}$) of $W = S_W^{-1} S_B$, **if S_W is non-singular.**

- The eigenvalues are scalar values, while the eigenvectors are non-zero vectors provides us with the information about the LDA space.
- The eigenvectors represent the directions of the new space, and the corresponding eigenvalues represent the scaling factor, length, or the magnitude of the eigenvectors.
- Thus, each eigenvector represents one axis of the LDA space and the associated eigenvalue represents the robustness of this eigenvector. The robustness of the eigenvector reflects its ability to discriminate between different classes and decreases the within-class variance of each class, hence meets the LDA goal.
- Thus, the eigenvectors with the k highest eigenvalues are used to construct a lower dimensional space (V_k), while the other eigenvectors ($\{v_{k+1}, v_{k+2}, v_M\}$) are neglected.


A
Data Matrix (X)
X=
ω1
ω2
ω3
x1
x2
xn1
xn1+1
xn1+2
xn1+n2
xN
(NxM)
B
Mean of each Class (μi)
μ1
μ2
μ3
(1xM)
C
Total Mean (μ)
μ
(1xM)
Mean-Centering Data
(Di=ωi-μi)
E
D=
ω1-μ1
ω2-μ2
ω3-μ3
x1-μ1
x2-μ1
xn1-μ1
x(n1+1)-μ2
x(n1+n2)-μ2
xN-μ3
(NxM)
D
Between-Class Matrix (SB)
SB1=
(Mx1)
X
(1xM)
=
SB1
(MxM)
SB=
SB1
SB2
SB3
SB
F
Within-Class Matrix (SW)
SW1=
(Mxn1)
X
(n1xM)
=
SW1
(MxM)
SW=
SW1
SW2
SW3
SW
G
Construct the Lower Dimensional Space (Vk)
Sorted Eigenvalues
λ1 λ2 ..... λk ..... λM
Largest k Eigenvalues
v1 v2 ..... vk ..... vM
Lower Dimensional Space (Vk)
Vk ∈ RMxK
k Selected Eigenvectors
arg max W = WTSBW / WTSwW

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- The dimension of the original data matrix ($X \in \mathcal{R}^{N \times M}$) is reduced by projecting it onto the lower dimensional space of LDA ($V_k \in \mathcal{R}^{M \times k}$) as denoted in Equation (7).
- The dimension of the data after projection is k; hence, $M - k$ features are ignored or deleted from each sample.
- Each sample (x_i) which was represented as a point a M-dimensional space will be represented in a k-dimensional space by projecting it onto the lower dimensional space (V_k) as follows, $y_i = x_i V_k$.

$$Y = XV_k \tag{7}$$

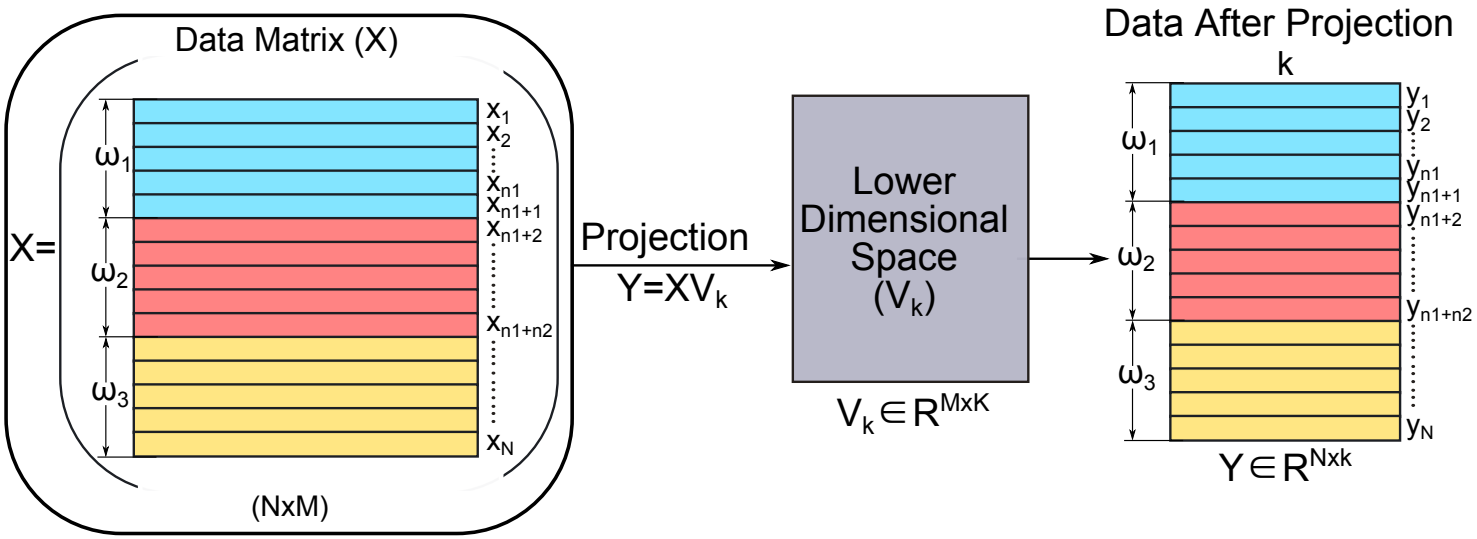


Figure: Projection of the original samples (i.e. data matrix) on the lower dimensional space of LDA (V_k).

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

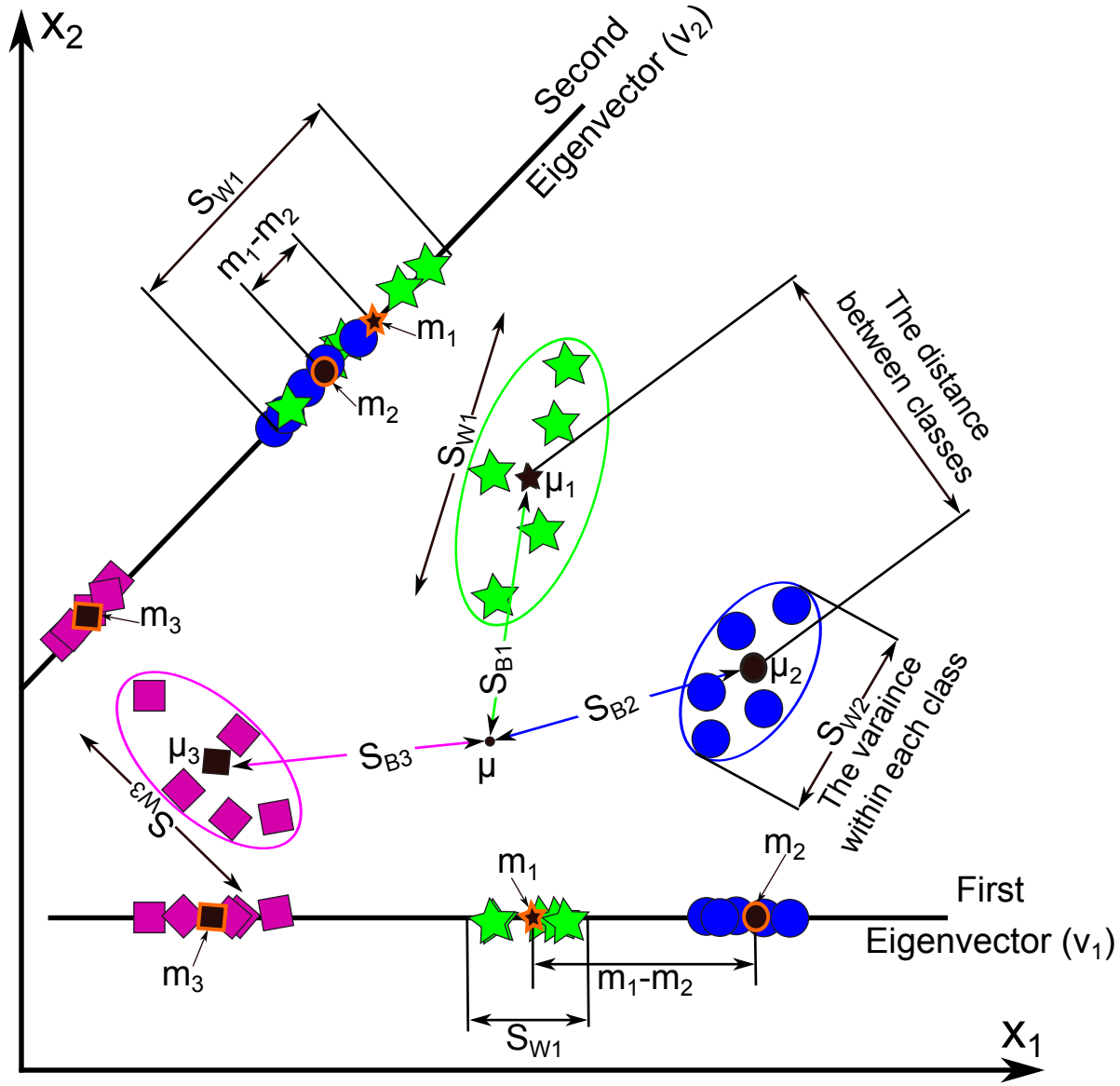


Figure: A visualized comparison between the two lower-dimensional sub-spaces which are calculated using three different classes.

- The above Figure shows a comparison between two lower-dimensional sub-spaces.
- Each class has five samples, and all samples are represented by two features only ($x_i \in \mathcal{R}^2$) to be visualized. Thus, each sample is represented as a point in two-dimensional space.
- The transformation matrix ($W(2 \times 2)$) is calculated as mentioned before.
- The eigenvalues (λ_1 and λ_2) and eigenvectors (i.e. sub-spaces) ($V = \{v_1, v_2\}$) of W are then calculated. Thus, there are two eigenvectors or sub-spaces.

- A comparison between the two lower-dimensional sub-spaces shows the following notices:
  - First, the separation distance between different classes when the data are projected on the first eigenvector (v_1) is much greater than when the data are projected on the second eigenvector (v_2). As shown in the figure, the three classes are efficiently discriminated when the data are projected on v_1. Moreover, the distance between the means of the first and second classes ($m_1 - m_2$) when the original data are projected on v_1 is much greater than when the data are projected on v_2, which reflects that the first eigenvector discriminates the three classes better than the second one.
  - Second, the within-class variance when the data are projected on v_1 is much smaller than when it projected on v_2. For example, S_{W_1} when the data are projected on v_1 is much smaller than when the data are projected on v_2. Thus, projecting the data on v_1 minimizes the within-class variance much better than v_2.
- From these two notes, we conclude that the first eigenvector meets the goal of the lower-dimensional space of the LDA technique than the second eigenvector; hence, it is selected to construct a lower-dimensional space.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- The computational complexity for the first four steps, common in both class-dependent and class-independent methods, are computed as follows.
- As illustrated in Algorithm (1) (see the paper), in step (2), to calculate the mean of the i^{th} class, there are $n_i M$ additions and M divisions, i.e., in total, there are $(NM + cM)$ operations.
- In step (3), there are NM additions and M divisions, i.e., there are $(NM + M)$ operations.
- The computational complexity of the fourth step is $c(M + M^2 + M^2)$, where M is for $\mu_i - \mu$, M^2 for $(\mu_i - \mu)(\mu_i - \mu)^T$, and the last M^2 is for the multiplication between n_i and the matrix $(\mu_i - \mu)(\mu_i - \mu)^T$.
- In the fifth step, there are $N(M + M^2)$ operations, where M is for $(x_{ij} - \mu_j)$ and M^2 is for $(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$.
- In the sixth step, there are M^3 operations to calculate S_W^{-1}, M^3 is for the multiplication between S_W^{-1} and S_B, and M^3 to calculate the eigenvalues and eigenvectors.

- In class-independent method, the computational complexity is $O(NM^2)$ if $N > M$; otherwise, the complexity is $O(M^3)$.
- In class-dependent algorithm, the number of operations to calculate the within-class variance for each class S_{W_j} in the sixth step is $n_j(M + M^2)$, and to calculate S_W, $N(M + M^2)$ operations are needed. Hence, calculating the within-class variance for both LDA methods are the same. In the seventh step and eighth, there are M^3 operations for the inverse, M^3 for the multiplication of $S_{W_i}^{-1}S_B$, and M^3 for calculating eigenvalues and eigenvectors. These two steps are repeated for each class which increases the complexity of the class-dependent algorithm. Totally, the computational complexity of the class-dependent algorithm is $O(NM^2)$ if $N > M$; otherwise, the complexity is $O(cM^3)$. Hence, the class-dependent method needs computations more than class-independent method.
- Given 40 classes and each class has ten samples. Each sample is represented by 4096 features ($M > N$). Thus, the computational complexity of the class-independent method is $O(M^3) = 4096^3$, while the class-dependent method needs $O(cM^3) = 40 \times 4096^3$.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- The aim of the two methods (class-dependent vs. class-independent) of the LDA is to calculate the LDA space.
- In the class-dependent LDA, one separate lower dimensional space is calculated for each class as follows, $W_i = S_{W_i}^{-1} S_B$, where W_i represents the transformation matrix for the ith class. Thus, eigenvalues and eigenvectors are calculated for each transformation matrix separately. Hence, the samples of each class are projected on their corresponding eigenvectors.
- In the class-independent method, one lower dimensional space is calculated for all classes. Thus, the transformation matrix is calculated for all classes, and the samples of all classes are projected on the selected eigenvectors.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

Given two different classes, $\omega_1(5 \times 2)$ and $\omega_2(6 \times 2)$ have $(n_1 = 5)$ and $(n_2 = 6)$ samples, respectively. Each sample in both classes is represented by two features (i.e. $M = 2$) as follows:

$$\omega_1 = \begin{bmatrix} 1.00 & 2.00 \\ 2.00 & 3.00 \\ 3.00 & 3.00 \\ 4.00 & 5.00 \\ 5.00 & 5.00 \end{bmatrix} \text{ and } \omega_2 = \begin{bmatrix} 4.00 & 2.00 \\ 5.00 & 0.00 \\ 5.00 & 2.00 \\ 3.00 & 2.00 \\ 5.00 & 3.00 \\ 6.00 & 3.00 \end{bmatrix} \tag{8}$$

$$\mu_1 = \begin{bmatrix} 3.00 & 3.60 \end{bmatrix} \text{ , } \mu_2 = \begin{bmatrix} 4.67 & 2.00 \end{bmatrix} \text{ , and} \tag{9}$$

$$\mu = \begin{bmatrix} \frac{5}{11}\mu_1 & \frac{6}{11}\mu_2 \end{bmatrix} = \begin{bmatrix} 3.91 & 2.727 \end{bmatrix} \tag{10}$$

- To calculate S_{B_1}

$$S_{B_1} = n_1(\mu_1 - \mu)^T(\mu_1 - \mu) = 5[-0.91 \;\; 0.87]^T[-0.91 \;\; 0.87] = \begin{bmatrix} 4.13 & -3.97 \\ -3.97 & 3.81 \end{bmatrix} \tag{11}$$

- Similarly, S_{B_1}

$$S_{B_2} = \begin{bmatrix} 3.44 & -3.31 \\ -3.31 & 3.17 \end{bmatrix} \tag{12}$$

- The total between-class variance is calculated a follows:

$$S_B = S_{B_1} + S_{B_2} = \begin{bmatrix} 4.13 & -3.97 \\ -3.97 & 3.81 \end{bmatrix} + \begin{bmatrix} 3.44 & -3.31 \\ -3.31 & 3.17 \end{bmatrix} = \begin{bmatrix} 7.58 & -7.27 \\ -7.27 & 6.98 \end{bmatrix} \tag{13}$$

- To calculate S_W, first calculate mean-centering data.

$$d_1 = \begin{bmatrix} -2.00 & -1.60 \\ -1.00 & -0.60 \\ 0.00 & -0.60 \\ 1.00 & 1.40 \\ 2.00 & 1.40 \end{bmatrix} \text{ and } d_2 = \begin{bmatrix} -0.67 & 0.00 \\ 0.33 & -2.00 \\ 0.33 & 0.00 \\ -1.67 & 0.00 \\ 0.33 & 1.00 \\ 1.33 & 1.00 \end{bmatrix} \tag{14}$$

- After centering the data, in class-independent method, the within-class variance for each class ($S_{W_i}(2 \times 2)$) is calculated as follows, $S_{W_j} = d_j^T * d_j = \sum_{i=1}^{n_j} (x_{ij} - \mu_j)^T (x_{ij} - \mu_j)$, where x_{ij} represents the i^{th} sample in the j^{th} class.
- The total within-class matrix ($S_W(2 \times 2)$) is then calculated as follows, $S_W = \sum_{i=1}^{c} S_{W_i}$.

$$S_{W_1} = \begin{bmatrix} 10.00 & 8.00 \\ 8.00 & 7.20 \end{bmatrix}, \; S_{W_2} = \begin{bmatrix} 5.33 & 1.00 \\ 1.00 & 6.00 \end{bmatrix}, \tag{15}$$
$$S_W = \begin{bmatrix} 15.33 & 9.00 \\ 9.00 & 13.20 \end{bmatrix}$$

- The transformation matrix (W) can be obtained as follows, $W = S_W^{-1} S_B$, and the values of (S_W^{-1}) and (W) are as follows:

$$S_W^{-1} = \begin{bmatrix} 0.11 & -0.07 \\ -0.07 & 0.13 \end{bmatrix} \text{ and } W = \begin{bmatrix} 1.37 & -1.32 \\ -1.49 & 1.43 \end{bmatrix} \quad (16)$$

- The eigenvalues $(\lambda(2 \times 2))$ and eigenvectors $(V(2 \times 2))$ of W are then calculated as follows:

$$\lambda = \begin{bmatrix} 0.00 & 0.00 \\ 0.00 & 2.81 \end{bmatrix} \text{ and } V = \begin{bmatrix} -0.69 & 0.68 \\ -0.72 & -0.74 \end{bmatrix} \quad (17)$$

- The second eigenvector (V_2) has corresponding eigenvalue more than the first one (V_1), which reflects that, the second eigenvector is more robust than the first one; hence, it is selected to construct the lower dimensional space.

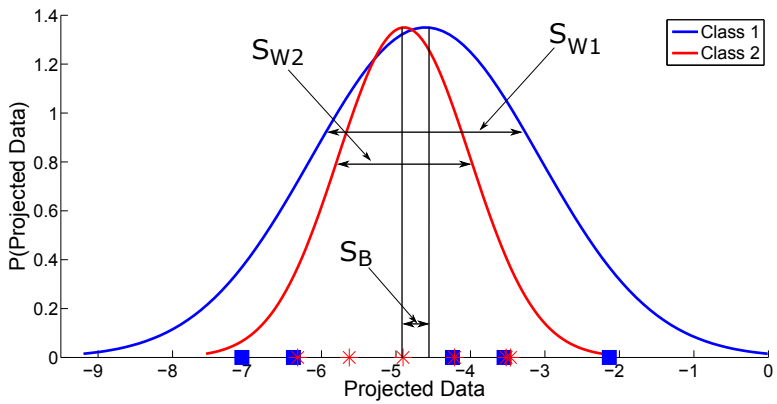
- The original data is projected on the lower dimensional space, as follows, $y_i = \omega_i \, V_2$, where $y_i(n_i \times 1)$ represents the data after projection of the i^{th} class, and its values will be as follows:

$$y_1 = \omega_1 V_2 = \begin{bmatrix} 1.00 & 2.00 \\ 2.00 & 3.00 \\ 3.00 & 3.00 \\ 4.00 & 5.00 \\ 5.00 & 5.00 \end{bmatrix} \begin{bmatrix} 0.68 \\ -0.74 \end{bmatrix} = \begin{bmatrix} -0.79 \\ -0.85 \\ -0.18 \\ -0.97 \\ -0.29 \end{bmatrix} \tag{18}$$

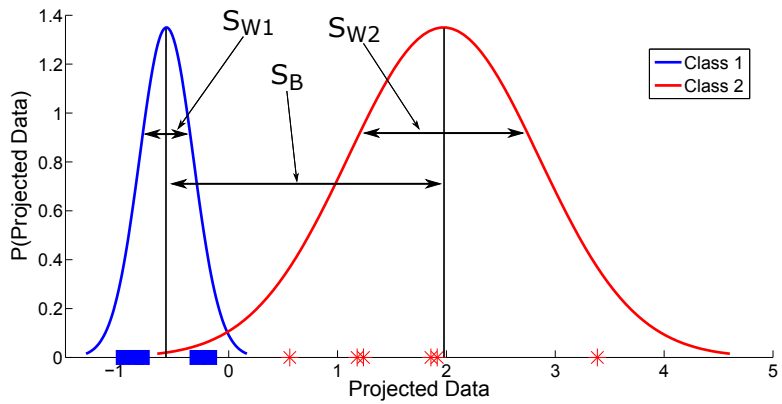
- Similarly, y_2 is as follows:

$$y_2 = \omega_2 V_2 = \begin{bmatrix} 1.24 \\ 3.39 \\ 1.92 \\ 0.56 \\ 1.18 \\ 1.86 \end{bmatrix} \tag{19}$$

- The data of each class is completely discriminated when it is projected on the second eigenvector (see Fig.)**(b)**) than the first one (see Fig. **a**)).
- The within-class variance (i.e. the variance between the same class samples) of the two classes are minimized when the data are projected on the second eigenvector. The within-class variance of the first class is small compared with as shown Fig. **(a)**.



(a)



(b)

Figure: Probability density function of the projected data of the first example, (a) the projected data on V_1, (b) the projected data on V_2.

- For class-dependent method, the aim is to calculate a separate transformation matrix (W_i) for each class.
- The within-class variance for each class ($S_{W_i}(2 \times 2)$) is calculated as in class-independent method.
- The transformation matrix (W_i) for each class is then calculated as follows, $W_i = S_{W_i}^{-1} S_B$. The values of the two transformation matrices (W_1 and W_2) will be as follows:

$$
\begin{aligned}
W_1 = S_{W_1}^{-1} S_B &= \begin{bmatrix} 10.00 & 8.00 \\ 8.00 & 7.20 \end{bmatrix}^{-1} \begin{bmatrix} 7.58 & -7.27 \\ -7.27 & 6.98 \end{bmatrix} \\
&= \begin{bmatrix} 0.90 & -1.00 \\ -1.00 & 1.25 \end{bmatrix} \begin{bmatrix} 7.58 & -7.27 \\ -7.27 & 6.98 \end{bmatrix} \\
&= \begin{bmatrix} 14.09 & -13.53 \\ -16.67 & 16.00 \end{bmatrix}
\end{aligned} \tag{20}
$$

- Similarly, W_2 is calculated as follows:

$$W_2 = \begin{bmatrix} 1.70 & -1.63 \\ -1.50 & 1.44 \end{bmatrix} \tag{21}$$

- The eigenvalues (λ_i) and eigenvectors (V_i) for each transformation matrix (W_i) are calculated, and the values of the eigenvalues and eigenvectors are shown below.

$$\lambda_{\omega_1} = \begin{bmatrix} 0.00 & 0.00 \\ 0.00 & 30.01 \end{bmatrix} \text{ and } V_{\omega_1} = \begin{bmatrix} -0.69 & 0.65 \\ -0.72 & -0.76 \end{bmatrix} \tag{22}$$

$$\lambda_{\omega_2} = \begin{bmatrix} 3.14 & 0.00 \\ 0.00 & 0.00 \end{bmatrix} \text{ and } V_{\omega_2} = \begin{bmatrix} 0.75 & 0.69 \\ -0.66 & 0.72 \end{bmatrix} \tag{23}$$

- where λ_{ω_i} and V_{ω_i} represent the eigenvalues and eigenvectors of the i^{th} class, respectively.

- From the results shown (above) it can be seen that, the second eigenvector of the first class ($V_{\omega_1}^{\{2\}}$) has corresponding eigenvalue more than the first one; thus, the second eigenvector is used as a lower dimensional space for the first class as follows, $y_1 = \omega_1 * V_{\omega_1}^{\{2\}}$, where y_1 represents the projection of the samples of the first class.
- The first eigenvector in the second class ($V_{\omega_2}^{\{1\}}$) has corresponding eigenvalue more than the second one. Thus, $V_{\omega_2}^{\{1\}}$ is used to project the data of the second class as follows, $y_2 = \omega_2 * V_{\omega_2}^{\{1\}}$, where y_2 represents the projection of the samples of the second class.
- The values of y_1 and y_2 will be as follows:

$$y_1 = \begin{bmatrix} -0.88 \\ -1.00 \\ -0.35 \\ -1.24 \\ -0.59 \end{bmatrix} \text{ and } y_2 = \begin{bmatrix} 1.68 \\ 3.76 \\ 2.43 \\ 0.93 \\ 1.77 \\ 2.53 \end{bmatrix} \tag{24}$$

- The Figure (below) shows a pdf graph of the projected data (i.e. y_1 and y_2) on the two eigenvectors ($V_{\omega_1}^{\{2\}}$ and $V_{\omega_2}^{\{1\}}$) and a number of findings are revealed the following:
  - First, the projection data of the two classes are efficiently discriminated.
  - Second, the within-class variance of the projected samples is lower than the within-class variance of the original samples.

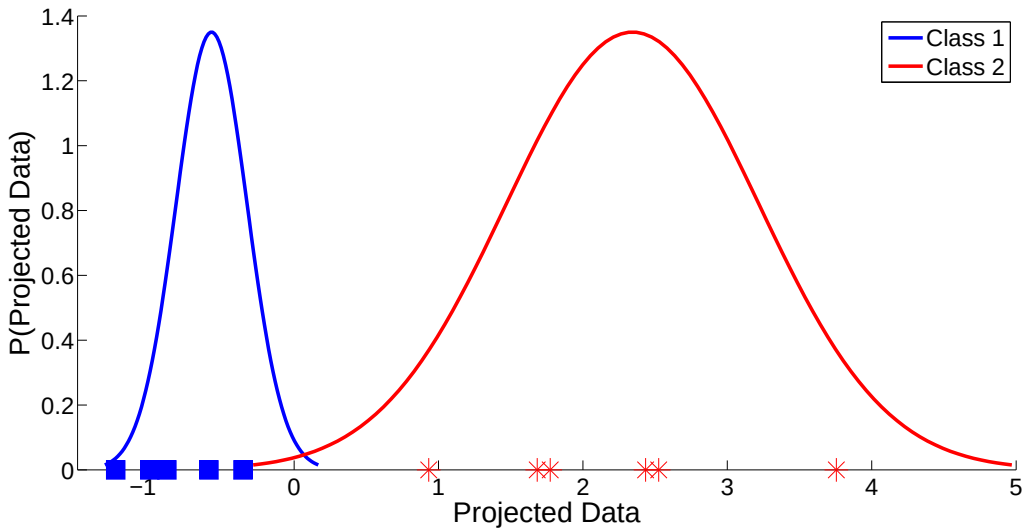


Figure: Probability density function (pdf) of the projected data using class-dependent method, the first class is projected on $V_{\omega_1}^{\{2\}}$, while the second class is projected on $V_{\omega_2}^{\{1\}}$.

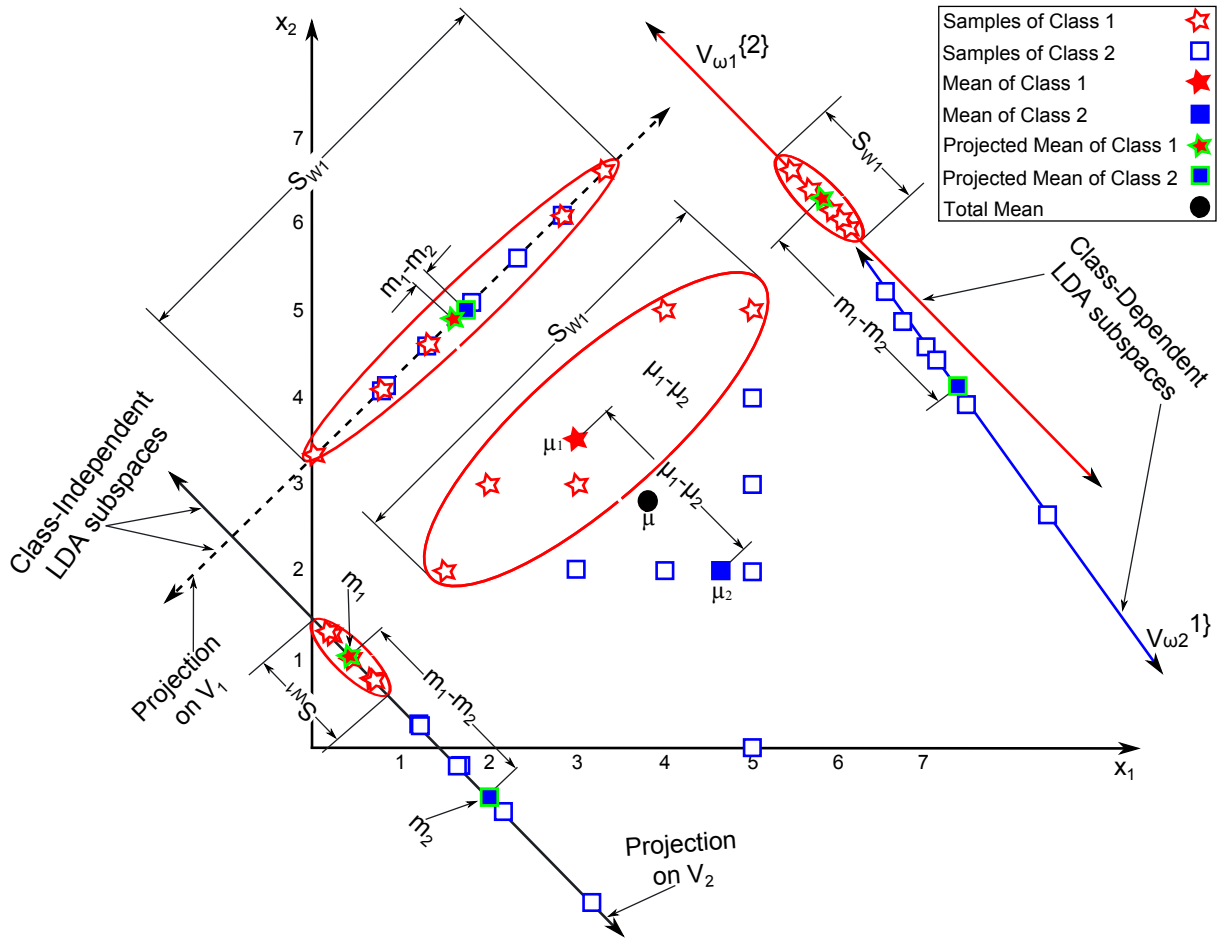


Figure: Illustration of the example of the two different methods of LDA methods. The blue and red lines represent the first and second eigenvectors of the class-dependent approach, respectively, while the solid and dotted black lines represent the second and first eigenvectors of class-independent approach, respectively.

- Figure (above) shows a further explanation of the two methods as following:
  - Class-Independent: As shown from the figure, there are two eigenvectors, V_1 (dotted black line) and V_2 (solid black line). The differences between the two eigenvectors are as follows:
    - The projected data on the second eigenvector (V_2) which has the highest corresponding eigenvalue will discriminate the data of the two classes better than the first eigenvector. As shown in the figure, the distance between the projected means $m_1 - m_2$ which represents S_B, increased when the data are projected on V_2 than V_1.
    - The second eigenvector decreases the within-class variance much better than the first eigenvector. The above figure illustrates that the within-class variance of the first class (S_{W_1}) was much smaller when it was projected on V_2 than V_1.
    - As a result of the above two findings, V_2 is used to construct the LDA space.

- Figure (above) shows a further explanation of the two methods as following:
  - Class-Dependent: As shown from the figure, there are two eigenvectors, $V_{\omega_1}^{\{2\}}$ (red line) and $V_{\omega_2}^{\{1\}}$ (blue line), which represent the first and second classes, respectively. The differences between the two eigenvectors are as following:
    - Projecting the original data on the two eigenvectors discriminates between the two classes. As shown in the figure, the distance between the projected means $m_1 - m_2$ is larger than the distance between the original means $\mu_1 - \mu_2$.
    - The within-class variance of each class is decreased. For example, the within-class variance of the first class (S_{W_1}) is decreased when it is projected on its corresponding eigenvector.
    - As a result of the above two findings, $V_{\omega_1}^{\{2\}}$ and $V_{\omega_2}^{\{1\}}$ are used to construct the LDA space.

- Figure (above) shows a further explanation of the two methods as following:
  - Class-Dependent vs. Class-Independent: The two LDA methods are used to calculate the LDA space, but a class-dependent method calculates separate lower dimensional spaces for each class which has two main limitations: (1) it needs more CPU time and calculations more than class-independent method; (2) it may lead to *SSS* problem because the number of samples in each class affects the singularity of S_{W_i}.
- These findings reveal that the standard LDA technique used the class-independent method rather than using the class-dependent method.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- Although LDA is one of the most common data reduction techniques, it suffers from two main problems:
  - **Small Sample Size** (SSS) and
  - **linearity problems**.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- LDA technique is used to find a linear transformation that discriminates between different classes.
- If the classes are non-linearly separable, LDA cannot find a lower dimensional space. In other words, LDA fails to find the LDA space when the discriminatory information are not in the means of classes.
- The Figure (below) shows how the discriminatory information does not exist in the mean, but in the variance of the data. This is because the means of the two classes are equal.
- The mathematical interpretation for this problem is as follows: if the means of the classes are approximately equal, so the S_B and W will be zero. Hence, the LDA space cannot be calculated.

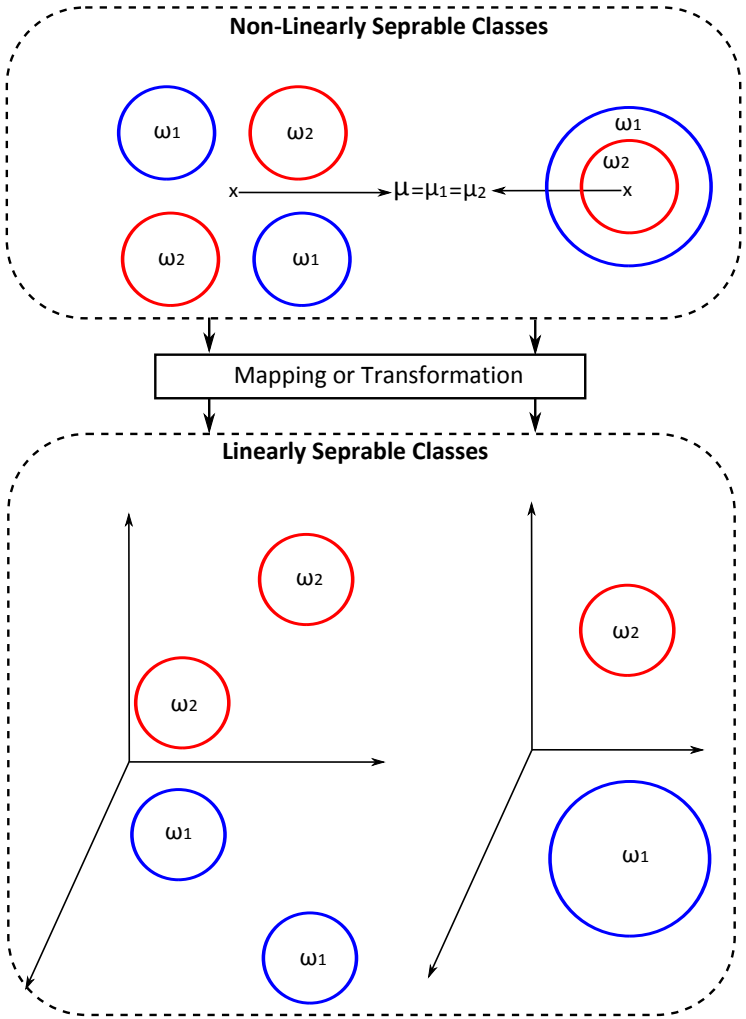


Figure: Two examples of two non-linearly separable classes, top panel shows how the two classes are non-separable, while the bottom shows how the transformation solves this problem and the two classes are linearly separable.

- One of the solutions of the linearity problem is based on the transformation concept, which is known as **a kernel methods or functions**.
- The kernel idea is applied in Support Vector Machine (SVM), Support Vector Regression (SVR), PCA, and LDA.
- The previous figure illustrates how the transformation is used to map the original data into a higher dimensional space; hence, the data will be linearly separable, and the LDA technique can find the lower dimensional space in the new space.
- The figure (next slide) graphically and mathematically shows how two non-separable classes in one-dimensional space are transformed into a two-dimensional space (i.e. higher dimensional space); thus, allowing linear separation.

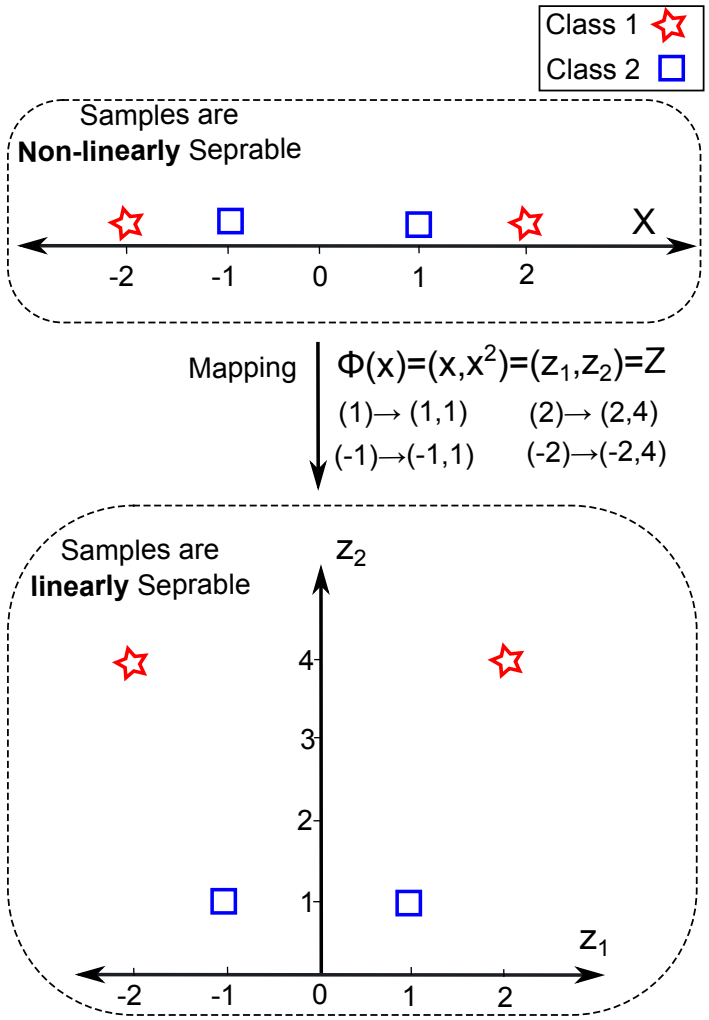


Figure: Example of kernel functions, the samples lie on the top panel (X) which are represented by a line (i.e. one-dimensional space) are non-linearly separable, where the samples lie on the bottom panel (Z) which are generated from mapping the samples of the top space are linearly separable.

- Let ϕ represents a nonlinear mapping to the new feature space $\mathcal{Z}$. The transformation matrix (W) in the new feature space $(\mathcal{Z})$ is calculated as follows:

$$F(W) = \max \left| \frac{W^T S_B^\phi W}{W^T S_W^\phi W} \right| \tag{25}$$

- where W is a transformation matrix and $\mathcal{Z}$ is the new feature space. The between-class matrix (S_B^ϕ) and the within-class matrix (S_W^ϕ) are defined as follows:

$$S_B^\phi = \sum_{i=1}^{c} n_i (\mu_i^\phi - \mu^\phi)(\mu_i^\phi - \mu^\phi)^T \tag{26}$$

$$S_W^\phi = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (\phi\{x_{ij}\} - \mu_j^\phi)(\phi\{x_{ij}\} - \mu_j^\phi)^T \tag{27}$$

- where $\mu_i^\phi = \frac{1}{n_i} \sum_{i=1}^{n_i} \phi\{x_i\}$ and $\mu^\phi = \frac{1}{N} \sum_{i=1}^{N} \phi\{x_i\} = \sum_{i=1}^{c} \frac{n_i}{N} \mu_i^\phi$

- Thus, in kernel LDA, all samples are transformed non-linearly into a new space $\mathcal{Z}$ using the function ϕ.
- In other words, the ϕ function is used to map the original features into $\mathcal{Z}$ space by creating a nonlinear combination of the original samples using a dot-products of it.
- There are many types of kernel functions to achieve this aim. Examples of these function include Gaussian or Radial Basis Function (RBF), $K(x_i, x_j) = exp(-||x_i - x_j||^2/2\sigma^2)$, where σ is a positive parameter, and the polynomial kernel of degree d, $K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d$.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- Problem Definition
  - Singularity, Small Sample Size (SSS), or under-sampled problem is one of the big problems of LDA technique.
  - This problem results from high-dimensional pattern classification tasks or a low number of training samples available for each class compared with the dimensionality of the sample space.
  - The *SSS* problem occurs when the S_W is singular[1].
  - The upper bound of the rank[2] of S_W is $N - c$, while the dimension of S_W is $M \times M$.
  - Thus, in most cases $M >> N - c$ which leads to *SSS* problem.
  - For example, in face recognition applications, the size of the face image my reach to $100 \times 100 = 10000$ pixels, which represent high-dimensional features and it leads to a singularity problem.

[1] A matrix is singular if it is square, does not have a matrix inverse, the determinant is zeros; hence, not all columns and rows are independent

[2] The rank of the matrix represents the number of linearly independent rows or columns

- Common Solutions to SSS Problem.
  - **Regularization (RLDA):**
    - In regularization method, the identity matrix is scaled by multiplying it by a regularization parameter ($\eta > 0$) and adding it to the within-class matrix to make it non-singular. Thus, the diagonal components of the within-class matrix are biased as follows, $S_W = S_W + \eta I$.
    - However, choosing the value of the regularization parameter requires more tuning and a poor choice for this parameter can degrade the performance of the method.
    - The parameter η is just added to perform the inverse of S_W and has no clear mathematical interpretation.

- Common Solutions to SSS Problem.
  - **Sub-space:**
    - In this method, a non-singular intermediate space is obtained to reduce the dimension of the original data to be equal to the rank of S_W; hence, S_W becomes full-rank[3], and then S_W can be inverted.
    - For example, Belhumeur et al. used PCA, to reduce the dimensions of the original space to be equal to $N - c$ (i.e. the upper bound of the rank of S_W).
    - However, losing some discriminant information is a common drawback associated with the use of this method.
  - **Null Space:**
    - There are many studies proposed to remove the null space of S_W to make S_W full-rank; hence, invertible.
    - The drawback of this method is that more discriminant information is lost when the null space of S_W is removed, which has a negative impact on how the lower dimensional space satisfies the LDA goal.

[3] A is a full-rank matrix if all columns and rows of the matrix are independent, (i.e. rank(A)= # rows= #cols)

- Four different variants of the LDA technique that are used to solve the SSS problem are introduced as follows:
  - **PCA + LDA** technique:
    - In this technique, the original d-dimensional features are first reduced to h-dimensional feature space using PCA, and then the LDA is used to further reduce the features to k-dimensions.
    - The PCA is used in this technique to reduce the dimensions to make the rank of S_W is $N - c$; hence, the *SSS* problem is addressed.
    - However, the PCA neglects some discriminant information, which may reduce the classification performance.
  - **Direct LDA** technique
    - Direct LDA (DLDA) is one of the well-known techniques that are used to solve the *SSS* problem.
    - This technique has two main steps.
    - In the first step, the transformation matrix, W, is computed to transform the training data to the range space of S_B.
    - In the second step, the dimensionality of the transformed data is further transformed using some regulating matrices.
    - The benefit of the DLDA is that there is no discriminative features are neglected as in PCA+LDA technique.

- Four different variants of the LDA technique that are used to solve the SSS problem are introduced as follows:
  - **Regularized LDA** technique:
    - In the Regularized LDA (RLDA), a small perturbation is add to the S_W matrix to make it non-singular. This regularization can be applied as follows:

$$(S_W + \eta I)^{-1} S_B w_i = \lambda_i w_i \tag{28}$$

    - where η represents a regularization parameter. The diagonal components of the S_W are biased by adding this small perturbation.
    - However, the regularization parameter need to be tuned and poor choice of it can degrade the generalization performance.

$$W = arg \max_{|W^T S_W W|=0} |W^T S_B W| \tag{29}$$

- Four different variants of the LDA technique that are used to solve the SSS problem are introduced as follows:
  - **Null LDA** technique
    - The aim of NLDA is to find the orientation matrix W.
    - Firstly, the range space of the S_W is neglected, and the data are projected only on the null space of S_W as follows, $S_W W = 0$.
    - Secondly, the aim is to search for W that satisfies $S_B W = 0$ and maximizes $|W^T S_B W|$.
    - The higher dimensionality of the feature space may lead to computational problems. This problem can be solved by (1) using the PCA technique as a pre-processing step, i.e. before applying the NLDA technique, to reduce the dimension of feature space to be $N - 1$; by removing the null space of $S_T = S_B + S_W$, (2) using the PCA technique before the second step of the NLDA technique.
    - Mathematically, in the Null LDA (NLDA) technique, the h column vectors of the transformation matrix $W = [w_1, w_2, \ldots, w_h]$ are taken to be the null space of the S_W as follows, $w_i^T S_W w_i = 0$, $\forall i = 1 \ldots h$, where $w_i^T S_B w_i \neq 0$. Hence, $M - (N - c)$ linearly independent vectors are used to form a new orientation matrix, which is used to maximize $|W^T S_B W|$ subject to the constraint $|W^T S_W W| = 0$ as follows, $W = arg \max_{|W^T S_W W| = 0} |W^T S_B W|$.

- Introduction to Linear Discriminant Analysis (LDA).
- Theoretical Background to LDA.
  - Definition of LDA.
  - Calculating the Between-Class Variance (S_B).
  - Calculating the Within-Class Variance (S_W).
  - Constructing the Lower Dimensional Space.
  - Projection onto the LDA space.
  - Example of Two LDA Subspaces.
  - Computational Complexity of LDA.
  - Class-Dependent vs. Class-Independent Methods.
- Numerical Example.
- Main Problems of LDA.
  - Linearity problem.
  - Small Sample Size Problem.
- Conclusions.

- LDA is an easy to implement dimensionality reduction method.
- The goals of LDA is to increase between class variance and decrease within-class variance.
- The value of eigenvalues reflect the robustness of the corresponding eigenvector.
- For more details, read the original paper "Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial" AI Communications 30 (2017) 169-190"
- For more questions, send to engalaatharwat@hotmail.com