Fisher Linear Discriminant Analysis (2 Classes)

April 27, 2025

Contents

Chapter 1

Introduction to Model

Population dataset $\mathcal{P}$ exists. Given available dataset $\mathcal{A}$ and its splits training dataset $\mathcal{I}$ and test dataset $\mathcal{S}$, each dataset contains samples with features $\boldsymbol{x} = [[\boldsymbol{x}]_1, [\boldsymbol{x}]_2, \cdots, [\boldsymbol{x}]_p]^\top$ and $y \in \{0, 1\}$. We want to establish a model:

$$M(\boldsymbol{x}) \tag{1.1}$$

which:

$$\text{maximize} \quad \sum_{i=1}^{N(\mathcal{P})} \mathrm{I}_{\mathrm{acc}} \left(M\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}} \right), y|_{\diamond i}^{\mathcal{I}} \right) \qquad \text{(Accuracy)} \tag{1.2}$$

$$\mathrm{I}_{\mathrm{acc}}(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

Fisher LDA is a **supervised** dimensionality reduction method that projects data onto a lower-dimensional space while:

- Maximizing **between-class scatter** (separation between class means)
- Minimizing **within-class scatter** (variance within each class)

which tends to fit a good classifier.

Mathematically, $\boldsymbol{w}^\top \boldsymbol{x}$ means projecting $\boldsymbol{x}$ on a direction $\boldsymbol{w}$.

Define $\mathcal{I}_0$ as the training data set that contains only samples from class 0 ($y|_{\diamond i}^{\mathcal{I}} = 0$), and $\mathcal{I}_1$ as the training data set that contains only samples from class 1 ($y|_{\diamond i}^{\mathcal{I}} = 1$).

The objective is to find a projection matrix $\boldsymbol{w}$ that maximizes the ratio between **between-class scatter** and **within-class scatter**:

$$J_1(\boldsymbol{w}) = \frac{\left\| \boldsymbol{w}^\top \boldsymbol{\mu}_0 - \boldsymbol{w}^\top \boldsymbol{\mu}_1 \right\|_2^2}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Sigma}_1 \boldsymbol{w}} \tag{1.4}$$

where:

$$\boldsymbol{\mu}_0 = \frac{1}{N(\mathcal{I}_0)} \sum_{i=1}^{N(\mathcal{I}_0)} \boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} \tag{1.5}$$

$$\boldsymbol{\mu}_1 = \frac{1}{N(\mathcal{I}_1)} \sum_{i=1}^{N(\mathcal{I}_1)} \boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} \tag{1.6}$$

$$\boldsymbol{\Sigma}_0 = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \tag{1.7}$$

$$\boldsymbol{\Sigma}_1 = \sum_{i=1}^{N(\mathcal{I}_1)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1 \right)^\top \tag{1.8}$$

- $\boldsymbol{\mu}_0$ is the class center of class 0 in training dataset $\mathcal{I}$.
- $\boldsymbol{\mu}_1$ is the class center of class 1 in training dataset $\mathcal{I}$.
- $\boldsymbol{\Sigma}_0$ is the class variance of class 0 in training dataset $\mathcal{I}$.
- $\boldsymbol{\Sigma}_1$ is the class variance of class 1 in training dataset $\mathcal{I}$.

Define the within-class scatter matrix:

$$\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 \tag{1.9}$$

We require $\boldsymbol{S}_{\mathrm{w}}$ to be non-zero as this is not naturally guaranteed, so that the mathematical programming problem (P1) can be well-defined.

Define the between-class scatter matrix:

$$\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top} \tag{1.10}$$

Substitute the (1.9) and (1.10) into (1.11) yields:

$$J_1(\boldsymbol{w}) = \frac{\left\|\boldsymbol{w}^{\top}\boldsymbol{\mu}_0 - \boldsymbol{w}^{\top}\boldsymbol{\mu}_1\right\|_2^2}{\boldsymbol{w}^{\top}\boldsymbol{\Sigma}_0\boldsymbol{w} + \boldsymbol{w}^{\top}\boldsymbol{\Sigma}_1\boldsymbol{w}} = \frac{\boldsymbol{w}^{\top}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top}\boldsymbol{w}}{\boldsymbol{w}^{\top}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\boldsymbol{w}} = \frac{\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}} \tag{1.11}$$

The J_1 is the objective we want to optimize. Define the problem:

$$\begin{aligned} \text{(P1)} \quad & \underset{\boldsymbol{w}}{\text{Maximize}} \quad J_1(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}} && (1.11) \\ & \text{Subject to} \quad \boldsymbol{w}^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} > \mathbf{0} && (1.12) \end{aligned}$$

The problem (P1) is the target of LDA model fitting.

We need to admit that there can be more than one optimal solution of (P1) when $\boldsymbol{S}_{\mathrm{w}}$ is singular. In this case, we only need to find one of the optimal solutions for (P1). However, for interest, we also discuss about the optimal solution sets in the following several chapters.

In the next section we are going to talk about how to maximize J_1.

✏️ Problem Solution Status

We define the following three problem solution status:

- Non-trivial Optimal Solution Status: (P1) has a non-empty optimal solution set, and the optimal solution set is different from the feasible solution set.
- Trivial Optimal Solution Status: (P1) has a non-empty optimal solution set, and the optimal solution set is equivalent to the feasible solution set.
- Infinite Optimal Solution Status: (P1) has an empty optimal solution set, and the objective value has no upper bound.

In details:

Status	Optimal Set	Optimal Objective Value	Relationship between Feasible and Optimal Sets
Non-trivial	Non-empty	Finite	Different
Trivial	Non-empty	Finite	Same
Infinite	Empty	Unbounded	Different

The trivial optimal solution status implies that (P1) is meaningless at all, as all feasible solutions can be optimal solutions.
The infinite optimal solution status implies that (P1) has no upper bound.
You may find the examples in the Example Chapter.

Chapter 2

Model Fitting

2.1 Reasoning

The model fitting aims to find an optimal solution, and the solution status for (P1).

We discuss the following two cases:

$\exists\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$:

When there $\exists\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$, the objective function $J_1(\boldsymbol{w}) = \gamma$ for all feasible solutions, which means that the optimal solution status is trivial. Therefore, $\boldsymbol{J}_1(\boldsymbol{w}) = 0$ for any feasible $\boldsymbol{w}$, which implies that the problem has a trivial optimal solution.

In this case, the feasible solution satisfies $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} > 0$, which is equivalent to $\boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} \neq \boldsymbol{0}$.

We know that since $\boldsymbol{S}_{\mathrm{w}}$ is non-zero, then there must exists at least one non-zero column. We just choose a non-zero column of $\boldsymbol{S}_{\mathrm{w}}$ to be a feasible as well as an optimal solution of (P1).

$\forall\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathrm{b}} \neq \gamma \boldsymbol{S}_{\mathrm{w}}$:

$\forall\gamma$, $\boldsymbol{S}_{\mathrm{b}} \neq \gamma \boldsymbol{S}_{\mathrm{w}}$, implies that the $\boldsymbol{S}_{\mathrm{b}} \neq \boldsymbol{0}$. According to the definition of $\boldsymbol{S}_{\mathrm{b}}$: $\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top$, $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$.

We notice that the numerator and denominator are the quadratic form of $\boldsymbol{w}$, which means that the solution is irrelevant to the magnitude of $\boldsymbol{w}$ but its direction. Without loss of generality, we let $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = 1$. Then we can remove the condition $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} > \boldsymbol{0}$ because it is already included in $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = 1$. Then we can solve the (P2) instead of (P1):

$$\boxed{\begin{aligned} (\mathrm{P2}) \quad \underset{\boldsymbol{w}}{\text{Minimize}} \quad & J_2(\boldsymbol{w}) = -\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} && (2.1) \\ \text{Subject to} \quad & \boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = 1 && (2.2) \end{aligned}}$$

Then we can apply the Generalized Fritz Johns Condition (**Theorem 4**):

$$\begin{aligned} \lambda \cdot \nabla_{\boldsymbol{w}}(-\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w}) + v \cdot \nabla_{\boldsymbol{w}}(\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} - 1) = 0 && (2.3) \\ -\lambda \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} + v \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = 0 && (2.4) \\ v \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = \lambda \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} && (2.5) \end{aligned}$$

and also the:

$$\begin{aligned} \lambda \geq 0 && (2.6) \\ [\lambda, v]^\top \neq \boldsymbol{0} && (2.7) \end{aligned}$$

yields an equivalent problem (P3) of (P2):

$$\boxed{\begin{aligned} (\mathrm{P3}) \quad \underset{\boldsymbol{w},\lambda,v}{\text{Minimize}} \quad & J_3(\boldsymbol{w}, \lambda, v) = -\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} && (2.8) \\ \text{Subject to} \quad & \lambda \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} = v \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} && (2.9) \\ & \boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = 1 && (2.10) \\ & \lambda \geq 0 && (2.11) \\ & [\lambda, v]^\top \neq \boldsymbol{0} && (2.12) \end{aligned}}$$

$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} = 1 > 0$ implies that $\boldsymbol{S}_\mathrm{w} \boldsymbol{w} \neq \boldsymbol{0}$ (**Proposition 4**). We claim that $\lambda \neq 0$. If $\lambda = 0$, then $v\boldsymbol{S}_\mathrm{w}\boldsymbol{w} = \lambda \boldsymbol{S}_\mathrm{b}\boldsymbol{w} = \boldsymbol{0}$. Since $\boldsymbol{S}_\mathrm{w}\boldsymbol{w} \neq \boldsymbol{0}$ then $v = 0$, which contradicts that $[\lambda, v]^\top \neq \boldsymbol{0}$. Therefore, $\lambda > 0$, $\lambda \boldsymbol{S}_\mathrm{b}\boldsymbol{w} = v\boldsymbol{S}_\mathrm{w}\boldsymbol{w}$ can be converted $\boldsymbol{S}_\mathrm{b}\boldsymbol{w} = \frac{v}{\lambda}\boldsymbol{S}_\mathrm{w}\boldsymbol{w}$.

Since $\lambda > 0$, then $[\lambda, v]^\top \neq \boldsymbol{0}$ holds for all $v \in \mathbb{R}$. We can remove this condition.

Now, since $\frac{v}{\lambda}$ ranges in $\mathbb{R}$, we can replace it with α, which yields (P4):

$$\boxed{\begin{aligned} \text{(P4)} \quad \underset{\boldsymbol{w},\alpha}{\text{Minimize}} \quad & J_4(\boldsymbol{w}, \alpha) = -\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} && (2.13) \\ \text{Subject to} \quad & \boldsymbol{S}_\mathrm{b}\boldsymbol{w} = \alpha \boldsymbol{S}_\mathrm{w}\boldsymbol{w} && (2.14) \\ & \boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} = 1 && (2.15) \end{aligned}}$$

Then:

$$\boldsymbol{S}_\mathrm{b}\boldsymbol{w} = \alpha \boldsymbol{S}_\mathrm{w}\boldsymbol{w} \tag{2.16}$$

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w} = \alpha \boldsymbol{S}_\mathrm{w}\boldsymbol{w} \tag{2.17}$$

Suppose that $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w} \neq 0$:

$$\alpha \boldsymbol{S}_\mathrm{w}\boldsymbol{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w} \tag{2.18}$$

$$\frac{\alpha}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}} \boldsymbol{S}_\mathrm{w}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{2.19}$$

Notice that $\frac{\alpha}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}}$ is a scalar and therefore we know that $\boldsymbol{S}_\mathrm{w}\boldsymbol{w}$ is collinear with $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. Since in (P1) the magnitude of $\boldsymbol{w}$ does not affect the feasibility and optimality of the result, then we solve the (P5) instead:

$$\boxed{\text{(P5)} \quad \underset{\boldsymbol{w},\,\beta}{\text{Solve}} \quad \beta \cdot (\boldsymbol{S}_\mathrm{w}\boldsymbol{w}) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1} \tag{2.20}$$

Since we just want to find a solution for (P5), then we can let $\beta = 1$ and solve the following (P6) instead:

$$\boxed{\text{(P6)} \quad \underset{\boldsymbol{w}}{\text{Solve}} \quad \boldsymbol{S}_\mathrm{w}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1} \tag{2.21}$$

Then, a solution of (P6) is the optimal solution of (P1).

However, the (P6) is not guaranteed to have a solution. When (P6) does not have a solution we can show that the objective value of (P1) is unbounded, and therefore have an infinite optimal solution status.

If (P6) has a solution, we still need to discuss whether the optimal solution status of (P6) is trivial or not.

By observation, we notice that if there $\exists \gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$, the $J_1(\boldsymbol{w})$ is not relevant to the $\boldsymbol{w}$, which implies that this is a key condition for trivial optimal solution.

2.2 Algorithm

The optimal solution status and an optimal solution can be described as follows:

- $\exists \gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$: (P1) has trivial optimal solution status, and an optimal solution is a non-zero column of $\boldsymbol{S}_\mathrm{w}$.
- $\forall \gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} \neq \gamma \boldsymbol{S}_\mathrm{w}$, and (P6) has a solution: (P1) has non-trivial optimal solution status, and an optimal solution is a solution of (P6).
- $\forall \gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} \neq \gamma \boldsymbol{S}_\mathrm{w}$, and (P6) has no solution: (P1) has infinite optimal solution status, and the optimal solution set of (P1) is empty.

The proof can be found in **Claim 14**, **Claim 15**, **Claim 16**.

Based on this, the algorithm is described as follows:

- Step 1: Split the training dataset $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$
- Step 2: Compute $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{S}_\mathrm{w}$. Check if $\boldsymbol{S}_\mathrm{w} = \boldsymbol{0}$. If yes, return that (P1) is ill-defined. Else if no, turn to Step 3.

- Step 3: Compute $\boldsymbol{S}_{\mathrm{b}}$. Check if there exists $\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$. If yes, return that (P1) has trivial optimal solution status, and a non-zero column of $\boldsymbol{S}_{\mathrm{w}}$ is an optimal solution of (P1). Else if no, turn to Step 4.
- Step 4: Solve (P6), if (P6) has solutions, then the (P1) has non-trivial optimal solution and the solution of (P6) is the optimal solution of (P1). Else if (P6) has no solution, then (P1) has infinite optimal solution status, and the optimal solution set is $\varnothing$.

2.3 Proof of Algorithm Feasibility

It is obvious that all steps in the algorithm are feasible.

2.4 Proof of Algorithm Optimality

Claim 1 $\boldsymbol{S}_{\mathrm{w}}$ *is symmetric* ($\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{S}_{\mathrm{w}}^{\top}$)

Proof:

$$\boldsymbol{S}_{\mathrm{w}}^{\top} = \left(\sum_{i=1}^{N(\mathcal{I}_0)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \right)^{\top} \tag{2.22}$$

$$\boldsymbol{S}_{\mathrm{w}}^{\top} = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \right)^{\top} + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \right)^{\top} \tag{2.23}$$

$$\boldsymbol{S}_{\mathrm{w}}^{\top} = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \tag{2.24}$$

$$\boldsymbol{S}_{\mathrm{w}}^{\top} = \boldsymbol{S}_{\mathrm{w}} \tag{2.25}$$

- (2.22): This comes directly from the definition: (1.9),(1.7), (1.8).
- (2.22) $\Rightarrow$ (2.23): Transpose of summation equals to the summation of transpose.
- (2.23) $\Rightarrow$ (2.24): Rearrangement of terms.
- (2.24) $\Rightarrow$ (2.25): This also comes from the definition: (1.9),(1.7), (1.8).

which completes the proof.

Claim 2 $\boldsymbol{S}_{\mathrm{w}}$ *is positive semi-definite.*

Proof:
For any $\boldsymbol{v} \in \mathbb{R}^p$:

$$\boldsymbol{v}^{\top} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} = \boldsymbol{v}^{\top} \left(\sum_{i=1}^{N(\mathcal{I}_0)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \right) + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \right) \right) \boldsymbol{v} \tag{2.26}$$

$$\boldsymbol{v}^{\top} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} = \sum_{i=1}^{N(\mathcal{I}_0)} \boldsymbol{v}^{\top} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \boldsymbol{v} + \sum_{i=1}^{N(\mathcal{I}_1)} \boldsymbol{v}^{\top} \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right) \left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \boldsymbol{v} \tag{2.27}$$

$$\boldsymbol{v}^{\top} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \boldsymbol{v} \right)^{\top} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \boldsymbol{v} \right) + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \boldsymbol{v} \right)^{\top} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \boldsymbol{v} \right) \tag{2.28}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{v} = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \boldsymbol{v} \right) \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \boldsymbol{v} \right) + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right) \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right) \tag{2.29}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{v} = \sum_{i=1}^{N(\mathcal{I}_0)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \boldsymbol{v} \right)^2 + \sum_{i=1}^{N(\mathcal{I}_1)} \left(\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right)^2 \tag{2.30}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{v} \geq 0 \tag{2.31}$$

- (2.26): This comes directly from the definition: (1.9),(1.7), (1.8).
- (2.26) ⇒ (2.27): By Distributive Law of Matrix (**Proposition 2**).
- (2.28) ⇒ (2.29): Since $\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \boldsymbol{v}$ is a scalar, its transpose equals to itself.

Combining $\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{v} \geq 0$ (2.31) and $\boldsymbol{S}_\mathrm{w}$ is symmetric (**Claim 1**), the $\boldsymbol{S}_\mathrm{w}$ is positive semi-definite by **Definition 2**.

Claim 3 $\boldsymbol{S}_\mathrm{b}$ *is symmetric* ($\boldsymbol{S}_\mathrm{b} = \boldsymbol{S}_\mathrm{b}^\top$)

Proof:

$$\boldsymbol{S}_\mathrm{b}^\top = \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \right)^\top \tag{2.32}$$

$$\boldsymbol{S}_\mathrm{b}^\top = \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \tag{2.33}$$

$$\boldsymbol{S}_\mathrm{b}^\top = \boldsymbol{S}_\mathrm{b} \tag{2.34}$$

- (2.32): This comes directly (1.10).
- (2.33) ⇒ (2.34): This comes directly (1.10).

which completes the proof.

Claim 4 $\boldsymbol{S}_\mathrm{b}$ *is positive semi-definite.*

Proof: For any $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} = \boldsymbol{v}^\top \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \tag{2.35}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} = \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right)^\top \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right) \tag{2.36}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} = \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right) \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right) \tag{2.37}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} = \left(\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^\top \boldsymbol{v} \right)^2 \tag{2.38}$$

$$\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} \geq 0 \tag{2.39}$$

- (2.35): This comes directly (1.10).
- (2.35) ⇒ (2.36): Rearrangement of terms.
- (2.36) ⇒ (2.37): Since $\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0 \right)^\top \boldsymbol{v}$ is a scalar, its transpose equals to itself.

Combining $\boldsymbol{v}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{v} \geq 0$ (2.39) and $\boldsymbol{S}_\mathrm{b}$ is symmetric (**Claim 3**), the $\boldsymbol{S}_\mathrm{b}$ is positive semi-definite by **Definition 2**.

Claim 5 (Cauchy Schwarz Inequality) *If $\boldsymbol{S}_{\mathrm{w}}$ is positive semi-definite, then $(\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^2 \leq (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}) \cdot (\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})$. The equality holds when $\boldsymbol{u} = \boldsymbol{v}$.*

Proof: Let $t \in \mathbb{R}$, since $\boldsymbol{S}_{\mathrm{w}}$ is positive semi-definite:

$$(\boldsymbol{u} + t\boldsymbol{v})^\top \boldsymbol{S}_{\mathrm{w}} (\boldsymbol{u} + t\boldsymbol{v}) \geq 0 \tag{2.40}$$
$$\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + t^2 \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} \geq 0 \tag{2.41}$$

Since $\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}$ is a scalar, therefore its transpose is itself and $\boldsymbol{S}_{\mathrm{w}}$ is symmetry, $\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} = (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^\top = \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}$.
Then,

$$\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + t^2 \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} \geq 0 \tag{2.42}$$
$$\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t^2 \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} \geq 0 \tag{2.43}$$
$$\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + 2t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t^2 \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} \geq 0 \tag{2.44}$$

must hold for all $t \in \mathbb{R}$, as for all $t \in \mathbb{R}$, $(\boldsymbol{u} + t\boldsymbol{v}) \in \mathbb{R}^p$ (by positive semi-definite).
Since $\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u} + 2t\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v} + t^2 \boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}$ is a quadratic function. Therefore, $\Delta = (2\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^2 - 4 \cdot (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}) \cdot (\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}) \leq 0$:

$$(2\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^2 - 4 \cdot (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}) \cdot (\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}) \leq 0 \tag{2.45}$$
$$4(\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^2 - 4 \cdot (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}) \cdot (\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}) \leq 0 \tag{2.46}$$
$$(\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v})^2 \leq (\boldsymbol{u}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{u}) \cdot (\boldsymbol{v}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{v}) \tag{2.47}$$

which completes the proof.

Claim 6 *Let $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ and* (P5) *has solution, then:*

- (P1) *has optimal solutions*
- *Denote $(\hat{\boldsymbol{w}}, \hat{\beta})$ to be the solution of* (P5)*, then $\hat{\boldsymbol{w}}$ is an optimal solution of* (P1)*.*
- *The optimal value of* (P1) *is strictly positive.*

Proof:

$$\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \neq \mathbf{0} \tag{2.48}$$
$$\hat{\beta} \cdot \boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} \neq \mathbf{0} \tag{2.49}$$
$$\boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} \neq \mathbf{0} \tag{2.50}$$
$$\hat{\boldsymbol{w}}^\top \boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} > 0 \tag{2.51}$$

- (2.48): This comes directly from the assumption $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$.
- (2.48) $\Rightarrow$ (2.49): By assumption $(\hat{\boldsymbol{w}}, \hat{\beta})$ is a solution of (P5), then $\hat{\beta} \cdot \boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$.
- (2.49) $\Rightarrow$ (2.50): If $\boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} = \mathbf{0}$, then $\hat{\beta} \cdot \boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} = 0$, which is a contradiction. Therefore, $\boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} \neq \mathbf{0}$.
- (2.50) $\Rightarrow$ (2.51): Since $\boldsymbol{S}_{\mathrm{w}}$ is positive semi-definite by **Claim 2**, and $\boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} \neq \mathbf{0}$, by **Proposition 4** we know that $\hat{\boldsymbol{w}}^\top \boldsymbol{S}_{\mathrm{w}} \hat{\boldsymbol{w}} > 0$.

(2.51) shows that $\hat{\boldsymbol{w}}$ is a feasible solution of (P1).
Denote any feasible solution of (P1) to be $\boldsymbol{w}$ then:

$$\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w} = \boldsymbol{w}^\top (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w} \tag{2.52}$$

$$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} = \boldsymbol{w}^\top (\hat{\beta} \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})(\hat{\beta} \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^\top \boldsymbol{w} \tag{2.53}$$

$$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} = \hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \cdot (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^\top \tag{2.54}$$

$$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} = \hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \cdot (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.55}$$

$$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} = \hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^2 \tag{2.56}$$

- (2.52): This comes directly from (1.10).
- (2.52) $\Rightarrow$ (2.53): By assumption $(\hat{\boldsymbol{w}}, \hat{\beta})$ is the the feasible solution of (P5), then $\hat{\beta} \cdot (\boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$.
- (2.54) $\Rightarrow$ (2.55): $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}$ is a scalar, and therefore its transpose $(\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^\top$ is still a scalar and equals to itself.

Then:

$$(\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^2 \leq (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w})(\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.57}$$

$$\hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^2 \leq \hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w})(\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.58}$$

$$\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w} \leq \hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w})(\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.59}$$

$$\frac{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}} \leq \hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.60}$$

$$J_1(\boldsymbol{w}) \leq \hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.61}$$

- (2.57): By Cauchy Schwarz Inequality (**Claim 5**).
- (2.57) $\Rightarrow$ (2.58): Multiply both sides with $\hat{\beta}^2$.
- (2.58) $\Rightarrow$ (2.59): $\hat{\beta}^2 (\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})^2 = \boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}$ comes from (2.56).
- (2.59) $\Rightarrow$ (2.60): Since $\boldsymbol{w}$ is a feasible solution of (P1), then by (1.12), $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$. Then divide both sides by $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}$.
- (2.60) $\Rightarrow$ (2.61): This comes directly from (1.11).

Similarly:

$$\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{b} \hat{\boldsymbol{w}} = \hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}})(\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.62}$$

$$\frac{\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{b} \hat{\boldsymbol{w}}}{\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}} = \hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.63}$$

$$J_1(\hat{\boldsymbol{w}}) = \hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) \tag{2.64}$$

- (2.57): (2.51) shows that $\hat{\boldsymbol{w}}$ is a feasible solution of (P1), then we can apply (2.56), by substitute $\boldsymbol{w}$ with $\hat{\boldsymbol{w}}$.
- (2.62) $\Rightarrow$ (2.63): Since $\boldsymbol{w}$ is a feasible solution of (P1), then by (1.12), $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$. Then divide both sides by $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}$.
- (2.63) $\Rightarrow$ (2.64): This comes directly from (1.11).

Since (2.51) shows that $\hat{\boldsymbol{w}}$ is a feasible solution of (P1), (2.61) shows that $J_1(\boldsymbol{w})$ has a upper bound and (2.64) shows that the upper bound is achieved at solutions $\hat{\boldsymbol{w}}$, which is exactly from the solution of (P5).
Therefore, we proved that if $(\hat{\boldsymbol{w}}, \hat{\beta})$ is a solution of (P5), then $\hat{\boldsymbol{w}}$ is an optimal solution of (P1).
Since $(\hat{\boldsymbol{w}}, \hat{\beta})$ is a solution of (P5), it must satisfy (2.20): $\hat{\beta} \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. Since $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \neq \mathbf{0}$, then $\hat{\beta} \neq 0$.
The optimal value $\hat{\beta}^2 (\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}}) > 0$ because $\hat{\beta} \neq 0$ and we have already shown that $\hat{\boldsymbol{w}}^\top \boldsymbol{S}_\mathrm{w} \hat{\boldsymbol{w}} > 0$, which completes the proof.

Claim 7 *If $\boldsymbol{w}^*$ is an optimal solution of* (P1)*, then there $\exists \lambda^*, v^* \in \mathbb{R}$, such that $\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}, \lambda^*, v^*\right)$ is an optimal solution of* (P3).

Proof:
For all feasible solutions $\boldsymbol{w}$ in (P2), since it satisfies $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = 1$ (2.2), it must satisfy $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} > 0$ (1.12), which means that $\boldsymbol{w}$ is also the feasible solution of (P1).

$$J_1(\boldsymbol{w}^*) \geq J_1(\boldsymbol{w}) \tag{2.65}$$

$$\frac{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^*}{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*} \geq \frac{\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}} \tag{2.66}$$

$$\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right)^\top \boldsymbol{S}_{\mathrm{b}} \left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \geq \frac{\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}} \tag{2.67}$$

$$\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right)^\top \boldsymbol{S}_{\mathrm{b}} \left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \geq \frac{\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{1} \tag{2.68}$$

$$\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right)^\top \boldsymbol{S}_{\mathrm{b}} \left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \geq \boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w} \tag{2.69}$$

$$-\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right)^\top \boldsymbol{S}_{\mathrm{b}} \left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \leq -\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w} \tag{2.70}$$

$$J_2\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \leq J_2(\boldsymbol{w}) \tag{2.71}$$

- (2.65): Since $\boldsymbol{w}^*$ is the optimal solution of (P1) and $\boldsymbol{w}$ is the feasible solution of (P1), $J_1(\boldsymbol{w}^*) \geq J_1(\boldsymbol{w})$
- (2.65) $\Rightarrow$ (2.66): This comes directly from the definition of $J_1(\boldsymbol{w})$ (1.11).
- (2.67) $\Rightarrow$ (2.68): Since $\boldsymbol{w}$ is defined to be a feasible solution of (P2),it must satisfy $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = 1$ (2.2)
- (2.70) $\Rightarrow$ (2.71): This comes directly from the definition of $J_2(\boldsymbol{w}^*)$ (2.1).

Also:

$$\frac{\boldsymbol{w}^{*\top}}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} \cdot \boldsymbol{S}_{\mathrm{w}} \cdot \frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} = 1 \tag{2.72}$$

(2.72) implies that $\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}$ satisfies the (P2) constraint (2.2), which means that $\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}$ is a feasible solution of (P2). Also, we have shown that for all $\boldsymbol{w}$ which are feasible solutions of (P2), there is $J_2\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}\right) \leq J_2(\boldsymbol{w})$ (2.71), which means that $\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}$ has minimality among all feasible solutions. Combine the feasibility and minimality, $\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}$ is an optimal solution of (P2).
By Generalized Fritz John condition (**Proposition 1**), since $\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}$ is an optimal solution of (P2), we conclude that there exists $\lambda^*, v^* \in \mathbb{R}$ such that $\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}, \lambda^*, v^*\right)$ is an optimal solution of (P3).
This completes the proof.

Claim 8 *Let $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ and* (P1) *has optimal solutions: If $\boldsymbol{w}^*$ is an optimal solution of* (P1)*, then there $\exists \beta^* \in \mathbb{R}$, such that $(\boldsymbol{w}^*, \beta^*)$ is a solution of* (P5)*.*

Proof:
Since $\boldsymbol{w}^*$ is an optimal solution of (P1), by **Claim 7** we conclude that there exists $\lambda^*, v^* \in \mathbb{R}$ such that $\left(\frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}}, \lambda^*, v^*\right)$ is an optimal solution of (P3). Therefore, it should be a feasible solution of (P3) as well, which means that it should satisfy the constraints (2.9),(2.12) ((2.10), (2.11) are ignored because they are useless here):

$$\lambda^* \boldsymbol{S}_{\mathrm{b}} \cdot \frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} = v^* \boldsymbol{S}_{\mathrm{w}} \cdot \frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} \tag{2.73}$$

$$[\lambda^*, v^*]^\top \neq \mathbf{0} \tag{2.74}$$

We also have:

$$J_1(\boldsymbol{w}^*) > 0 \tag{2.75}$$

$$\frac{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^*}{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*} > 0 \tag{2.76}$$

$$\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^* > 0 \tag{2.77}$$

$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^* > 0 \tag{2.78}$$

$$\left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*\right)^\top \left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*\right) > 0 \tag{2.79}$$

$$\left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*\right)\left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*\right) > 0 \tag{2.80}$$

$$\left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*\right)^2 > 0 \tag{2.81}$$

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^* \neq 0 \tag{2.82}$$

- (2.75): In **Claim 6** we have shown that the optimal value of (P1) is strictly positive. Since $\boldsymbol{w}^*$ is an optimal solution of (P1), $J_1(\boldsymbol{w}^*) > 0$.
- (2.75) $\Rightarrow$ (2.76): This comes directly from the definition of $J_1(\boldsymbol{w})$ (1.11).
- (2.76) $\Rightarrow$ (2.77): Since $\boldsymbol{w}^*$ is the optimal solution of (P1), it must be feasible solution of (P1) and therefore satisfy $\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^* > 0$ (1.12).
- (2.77) $\Rightarrow$ (2.78): This comes directly from $\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top$ (1.10).
- (2.78) $\Rightarrow$ (2.79): Rearrangement of terms.
- (2.79) $\Rightarrow$ (2.80): Since $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}^*$ is a scalar, its transpose equals to itself.
- (2.81) $\Rightarrow$ (2.82): If the square of a scalar is strictly greater than 0, then this scalar cannot be 0.

We claim that $\lambda^* \neq 0$ and prove it by contradiction:
If $\lambda^* = 0$, then:

$$v^* \boldsymbol{S}_{\mathrm{w}} \cdot \frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} = \lambda^* \boldsymbol{S}_{\mathrm{b}} \cdot \frac{\boldsymbol{w}^*}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^*}} \tag{2.83}$$

$$v^* \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^* = \lambda^* \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^* \tag{2.84}$$

$$v^* \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^* = 0 \cdot \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^* \tag{2.85}$$

$$v^* \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^* = \mathbf{0} \tag{2.86}$$

$$v^* = 0 \tag{2.87}$$

- (2.83): This comes directly from (2.73).

- (2.83) ⇒ (2.84): Multiply both sides with $\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}}$.
- (2.84) ⇒ (2.85): $\lambda^{*}=0$ as the contradiction assumption.
- (2.86) ⇒ (2.87): Since $\boldsymbol{w}^{*}$ is an optimal solution of (P1) then it must be feasible satisfying (1.12) $\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}>0$. By **Proposition 4** we know that $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}\neq\boldsymbol{0}$. Therefore, $v^{*}=0$ must hold so that $v^{*}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}=0$.

However, now $[\lambda^{*},v^{*}]^{\top}=[0,0]^{\top}=\boldsymbol{0}$ contradicts (2.74). Therefore:

$$\lambda^{*}\neq 0 \tag{2.88}$$

Then:

$$\lambda^{*}\boldsymbol{S}_{\mathrm{b}}\frac{\boldsymbol{w}^{*}}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}}}=v^{*}\boldsymbol{S}_{\mathrm{w}}\frac{\boldsymbol{w}^{*}}{\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}}} \tag{2.89}$$

$$\lambda^{*}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}^{*}=v^{*}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*} \tag{2.90}$$

$$\lambda^{*}(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}^{*}=v^{*}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*} \tag{2.91}$$

$$\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}=\frac{v^{*}}{\lambda^{*}(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}^{*}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*} \tag{2.92}$$

- (2.89): This comes directly from (2.73).
- (2.89) ⇒ (2.90): Multiply both sides with $\sqrt{\boldsymbol{w}^{*\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}^{*}}$.
- (2.90) ⇒ (2.91): This comes directly from (1.10).
- (2.91) ⇒ (2.92): Since $\lambda^{*}\neq 0$ (2.88) and $(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}^{*}\neq 0$ (2.82), we can divide both sides by $\lambda^{*}(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}^{*}$

By observation we see that $(\boldsymbol{w},\beta)=\left(\boldsymbol{w}^{*},\frac{v^{*}}{\lambda^{*}(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}^{*}}\right)$ is a solution of (P5).
This shows that if $\boldsymbol{w}^{*}$ is the optimal solution of (P1), then there $\exists\beta^{*}\in\mathbb{R}$, such that $(\boldsymbol{w}^{*},\beta^{*})$ is a solution of (P5).

Claim 9 *Let $\boldsymbol{\mu}_{0}\neq\boldsymbol{\mu}_{1}$ and* (P5) *has solutions, then:*

- (P1) *has optimal solutions.*
- *Denote $\boldsymbol{w}^{*}$ to be the optimal solution of* (P1)*, then there $\exists\beta^{*}\in\mathbb{R}$, such that $(\boldsymbol{w}^{*},\beta^{*})$ is a solution of* (P5)*.*

Proof:
In **Claim 6** we have shown that if (P5) has solutions, then (P1) must have optimal solutions.
Since (P1) has optimal solutions, we can define any optimal solution to be $\boldsymbol{w}^{*}$.
Since $\boldsymbol{w}^{*}$ is an optimal solution of (P1), by **Claim 8** we conclude that there exists $\beta^{*}\in\mathbb{R}$ such that $(\boldsymbol{w}^{*},\beta^{*})$ is an optimal solution of (P5), which completes the proof.

Claim 10 *If* (P5) *does not have a solution, then* (P1) *has infinite maximum.*

Proof: Since (P5) does not have solution, then a special case when $\beta=1$ of (P5) does not have solution as well: $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}=\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}$ does not have a solution, this means that $\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}\notin\mathrm{Col}(\boldsymbol{S}_{\mathrm{w}})=\mathrm{Row}(\boldsymbol{S}_{\mathrm{w}})$, as $\boldsymbol{S}_{\mathrm{w}}$ is symmetric.
By **Theorem 1** we have $\mathrm{Row}(\boldsymbol{S}_{\mathrm{w}})^{\perp}=\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})$, then by **Theorem 2** we have $\mathrm{Row}(\boldsymbol{S}_{\mathrm{w}})=(\mathrm{Row}(\boldsymbol{S}_{\mathrm{w}})^{\perp})^{\perp}=\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})^{\perp}$.
Therefore, $\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}\notin\mathrm{Col}(\boldsymbol{S}_{\mathrm{w}})=\mathrm{Row}(\boldsymbol{S}_{\mathrm{w}})=\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})^{\perp}$.
Recall the definition of $\perp$, $\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})^{\perp}=\{\boldsymbol{y}\in\mathbb{R}^{p}\mid\forall\boldsymbol{v}\in\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}}),\,\boldsymbol{y}^{\top}\boldsymbol{v}=0\}$. Since $\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1}\notin\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})^{\perp}$, by definition, there $\exists\boldsymbol{w}_{0}\in\mathrm{Ker}(\boldsymbol{S}_{\mathrm{w}})$ such that:

$$(\boldsymbol{\mu}_{0}-\boldsymbol{\mu}_{1})^{\top}\boldsymbol{w}_{0}\neq\boldsymbol{0} \tag{2.93}$$

Since $\boldsymbol{w}_0 \in \text{Ker}(\boldsymbol{S}_\text{w})$, by definition:

$$\boldsymbol{S}_\text{w}\boldsymbol{w}_0 = \boldsymbol{0} \tag{2.94}$$

Since $\text{rank}(\boldsymbol{S}_\text{w}) \geq 1$ (as $\boldsymbol{S}_\text{w}$ is nonzero required by algorithm). Then there must $\exists \boldsymbol{v}_0$ such that $\boldsymbol{S}_\text{w}\boldsymbol{v}_0 \neq \boldsymbol{0}$. Since $\boldsymbol{S}_\text{w}$ is positive semi-definite, by **Proposition 4**, we have $\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 > \boldsymbol{0}$.
Then we can define $\boldsymbol{w}_\epsilon := \boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0$. Then:

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ \big((\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0)^\top \boldsymbol{S}_\text{b}(\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0)\big) \tag{2.95}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ \big(\boldsymbol{w}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_0 + \epsilon\boldsymbol{w}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{v}_0 + \epsilon\boldsymbol{v}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_0 + \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{v}_0\big) \tag{2.96}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ \boldsymbol{w}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_0 + \lim_{\epsilon\to 0}\ \epsilon\boldsymbol{w}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{v}_0 + \lim_{\epsilon\to 0}\ \epsilon\boldsymbol{v}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_0 + \lim_{\epsilon\to 0}\ \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{v}_0 \tag{2.97}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = \boldsymbol{w}_0^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_0 + 0 + 0 + 0 \tag{2.98}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = \boldsymbol{w}_0^\top (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \tag{2.99}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0)^\top ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0) \tag{2.100}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0)((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0) \tag{2.101}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon = ((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0)^2 \tag{2.102}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{b}\boldsymbol{w}_\epsilon > 0 \tag{2.103}$$

$$\tag{2.104}$$

- (2.95): This comes directly from the definition $\boldsymbol{w}_\epsilon := \boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0$.
- (2.96) $\Rightarrow$ (2.97): Distributive Law of Limit Addition.
- (2.98) $\Rightarrow$ (2.99): This comes directly from the (1.10).
- (2.99) $\Rightarrow$ (2.100): Rearrangement of terms.
- (2.100) $\Rightarrow$ (2.101): Since $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0$ is a scalar, the transpose of a scalar equals to itself.
- (2.102) $\Rightarrow$ (2.103): Since (2.93), then the square of a non-zero scalar is greater than 0.

There is an important property to show:

$$\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 = (\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0)^\top \tag{2.105}$$

$$\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 = \boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}^\top \boldsymbol{w}_0 \tag{2.106}$$

$$\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 = \boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 \tag{2.107}$$

- (2.105): Since $\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0$ is a scalar, the transpose of scalar equals to itself.
- (2.106) $\Rightarrow$ (2.107): This comes from **Claim 1**.

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ \big((\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0)^\top \boldsymbol{S}_\text{w}(\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0)\big) \tag{2.108}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ (\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 + \epsilon\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 + \epsilon\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 + \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0) \tag{2.109}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ (\boldsymbol{w}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 + \epsilon\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_0 + \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0) \tag{2.110}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ (\boldsymbol{w}_0^\top \boldsymbol{0} + \epsilon\boldsymbol{v}_0^\top \boldsymbol{0} + \epsilon\boldsymbol{v}_0^\top \boldsymbol{0} + \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0) \tag{2.111}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = \lim_{\epsilon\to 0}\ \epsilon^2\boldsymbol{v}_0^\top \boldsymbol{S}_\text{w}\boldsymbol{v}_0 \tag{2.112}$$

$$\lim_{\epsilon\to 0}\ \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\text{w}\boldsymbol{w}_\epsilon = 0 \tag{2.113}$$

- (2.108): This comes directly from the definition $\boldsymbol{w}_\epsilon := \boldsymbol{w}_0 + \epsilon \boldsymbol{v}_0$.
- (2.109) $\Rightarrow$ (2.110): This comes from $\boldsymbol{w}_0^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{v}_0 = \boldsymbol{v}_0^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}_0$ (2.107).
- (2.110) $\Rightarrow$ (2.111): This comes from (2.94)

Then:

$$\lim_{\epsilon \to 0} \frac{\boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}_\epsilon}{\boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}_\epsilon} = +\infty \tag{2.114}$$

$$\lim_{\epsilon \to 0} J_1(\boldsymbol{w}_\epsilon) = +\infty \tag{2.115}$$

- (2.114): Since $\boldsymbol{S}_\mathrm{w}$ and $\boldsymbol{S}_\mathrm{b}$ are positive semi-definite, $\boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}_\epsilon \geq 0$ and $\boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}_\epsilon \geq 0$ for $\forall \epsilon \in \mathbb{R}$, $\lim_{\epsilon \to 0} \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}_\epsilon > 0$, and $\lim_{\epsilon \to 0} \boldsymbol{w}_\epsilon^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}_\epsilon = 0$.
- (2.114) $\Rightarrow$ (2.115): This comes directly from the definition of $J_1(\boldsymbol{w})$ (1.11).

which completes the proof.

Claim 11 *There exists $\gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$ $\Rightarrow$ (P1) has trivial solution status.*

Proof:
Since $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$ by constraint of (P1) we know that the denominator of $J_1(\boldsymbol{w})$ should be non-zero and therefore well-defined.

$$J_1(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{b} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}} \tag{2.116}$$

$$J_1(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top (\gamma \boldsymbol{S}_\mathrm{w}) \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}} \tag{2.117}$$

$$J_1(\boldsymbol{w}) = \gamma \cdot \frac{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w}} \tag{2.118}$$

$$J_1(\boldsymbol{w}) = \gamma \tag{2.119}$$

- (2.116): This comes directly from the definition of J_1 (1.11).
- (2.116) $\Rightarrow$ (2.117): This comes from the assumption that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$.

which shows that $J_1(\boldsymbol{w})$ for all feasible $\boldsymbol{w}$ has the same value γ.
Since all feasible solutions have the same value, all feasible solutions are optimal solutions. Therefore, the feasible solution set is the same as the optimal solution set, which means that (P1) has trivial solution status.

Claim 12 (P1) *has trivial solution status $\Rightarrow$ There exists $\gamma \in \mathbb{R}$, such that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$.*

Proof:
Case $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$:
If $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, then $\boldsymbol{S}_\mathrm{b} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \boldsymbol{0}$.
This means that there exists $\gamma = 0$ such that $\boldsymbol{S}_\mathrm{b} = \gamma \boldsymbol{S}_\mathrm{w}$
Case $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$:
If $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$: The (P1) has trivial solution status implies that $J_1(\boldsymbol{w})$ is the same for all feasible solutions (In other words, $\boldsymbol{w}$ satisfying $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$). Then this means that (P1) has optimal solutions and all $\boldsymbol{w}$ satisfying $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$ are optimal solutions. Let us denote the same objective value to be γ_0.
Since $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ all $\boldsymbol{w}$ satisfying $\boldsymbol{w}^\top \boldsymbol{S}_\mathrm{w} \boldsymbol{w} > 0$ are optimal solutions of (P1), by **Claim 9** we conclude that for each $\boldsymbol{w}$, there exists a corresponding $\beta \in \mathbb{R}$ such that $(\boldsymbol{w}, \beta)$ is a solution of (P5) (2.20), which means that:

$$\beta \cdot (\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{2.120}$$

Then:

$$\beta \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{2.121}$$
$$\beta \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} \neq \boldsymbol{0} \tag{2.122}$$
$$\beta \neq 0 \tag{2.123}$$

- (2.121): This comes directly from (2.120).
- (2.121) $\Rightarrow$ (2.122): Now we are in the case that $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$.
- (2.122) $\Rightarrow$ (2.123): $\beta \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} \neq \boldsymbol{0}$ implies that $\beta \neq 0$ and $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} \neq \boldsymbol{0}$.

Now, we show by contradiction that $\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}}) \geq 2$ does not hold:
Suppose that $\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}}) \geq 2$, then there exists two non-zero linearly independent columns $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ of $\boldsymbol{S}_{\mathrm{w}}$:

$$\boldsymbol{s}_i = \boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i \tag{2.124}$$
$$\boldsymbol{s}_j = \boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j \tag{2.125}$$

where $\boldsymbol{e}_i$ is a vector that the i-th elememt is 1 and the others are 0s, and $\boldsymbol{e}_j$ is a vector that the j-th elememt is 1 and the others are 0s.
Continue:

$$\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i = \boldsymbol{s}_i \tag{2.126}$$
$$\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i \neq \boldsymbol{0} \tag{2.127}$$
$$\boldsymbol{e}_i^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i > \boldsymbol{0} \tag{2.128}$$

- (2.126): This comes directly from (2.124).
- (2.126) $\Rightarrow$ (2.127): This is because $\boldsymbol{s}_i$ is non-zero.
- (2.127) $\Rightarrow$ (2.128): $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i \neq \boldsymbol{0}$, then $\boldsymbol{e}_i^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i > \boldsymbol{0}$ by **Proposition 4**.

$$\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j = \boldsymbol{s}_j \tag{2.129}$$
$$\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j \neq \boldsymbol{0} \tag{2.130}$$
$$\boldsymbol{e}_j^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j > \boldsymbol{0} \tag{2.131}$$

- (2.129): This comes directly from (2.125).
- (2.129) $\Rightarrow$ (2.130): This is because $\boldsymbol{s}_j$ is non-zero.
- (2.130) $\Rightarrow$ (2.131): $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j \neq \boldsymbol{0}$, then $\boldsymbol{e}_j^{\top}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j > \boldsymbol{0}$ by **Proposition 4**.

Since we have already concluded that for each optimal solution $\boldsymbol{w}$ of (P1) there exists $\beta \in \mathbb{R}$ such that (2.120) holds, and $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are optimal solutions, the following two equations hold:

$$\beta_i \boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_i = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{2.132}$$
$$\beta_j \boldsymbol{S}_{\mathrm{w}}\boldsymbol{e}_j = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{2.133}$$

where $(\boldsymbol{e}_i, \beta_i)$ and $(\boldsymbol{e}_j, \beta_j)$ are the solution of (P5) and $\beta_i, \beta_j \neq 0$.

Then:

$$\beta_i \boldsymbol{S}_{\mathrm{w}} \boldsymbol{e}_i = \beta_j \boldsymbol{S}_{\mathrm{w}} \boldsymbol{e}_j \tag{2.134}$$

$$\boldsymbol{S}_{\mathrm{w}} \boldsymbol{e}_i = \frac{\beta_j}{\beta_i} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{e}_j \tag{2.135}$$

$$\boldsymbol{s}_i = \frac{\beta_j}{\beta_i} \boldsymbol{s}_j \tag{2.136}$$

- (2.134): Combine (2.132) and (2.133).
- (2.134) $\Rightarrow$ (2.135): In (2.123) we have proved that $\beta \neq 0$, and therefore β_i and β_j are special case of it, and therefore we can divide β_i and β_j on both sides.
- (2.135) $\Rightarrow$ (2.136): This comes directly from (2.124) and (2.125).

In (2.136) we proved that $\boldsymbol{s}_i = \frac{\beta_j}{\beta_i} \boldsymbol{s}_j$, which means that $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are collinear. However, $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are linearly independent (because $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are two non-zero linearly independent columns of $\boldsymbol{S}_{\mathrm{w}}$), which results in a contradiction. Therefore, $\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}}) \geq 2$ does not hold.
Therefore, $\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}}) = 1$: coupled with the fact that $\boldsymbol{S}_{\mathrm{w}}$ is positive semi-definite (**Claim 2**) then $\boldsymbol{S}_{\mathrm{w}}$ can be represented as $\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{u}\boldsymbol{u}^\top$ (**Proposition 3**).
Then:

$$\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 = \beta \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} \tag{2.137}$$

$$\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 = \beta \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{w} \tag{2.138}$$

$$\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 = \beta \boldsymbol{u}^\top \boldsymbol{w} \boldsymbol{u} \tag{2.139}$$

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \left(\beta \boldsymbol{u}^\top \boldsymbol{w}\right)^2 \boldsymbol{u}\boldsymbol{u}^\top \tag{2.140}$$

$$\boldsymbol{S}_{\mathrm{b}} = \left(\beta \boldsymbol{u}^\top \boldsymbol{w}\right)^2 \boldsymbol{S}_{\mathrm{w}} \tag{2.141}$$

- (2.137): This comes directly from (2.120).
- (2.137) $\Rightarrow$ (2.138): Since the $\boldsymbol{S}_{\mathrm{w}}$ has been represented as: $\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{u}\boldsymbol{u}^\top$.
- (2.140) $\Rightarrow$ (2.141): This comes from (1.10) and $\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{u}\boldsymbol{u}^\top$.

The $\left(\beta \boldsymbol{u}^\top \boldsymbol{w}\right)^2$ is a scalar and therefore we prove that there $\exists \gamma = \left(\beta \boldsymbol{u}^\top \boldsymbol{w}\right)^2$ such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$.

There are some interesting and special cases which satisfying that: there exists $\gamma \in \mathbb{R}$, such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$.

An interesting case is $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, which means that the centers of two classes are the same. Since $\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \boldsymbol{0} \cdot \boldsymbol{0}^\top = \boldsymbol{0}$, and the corresponding $\gamma = 0$, $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$. Therefore, the (P1) is trivial when $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$.

Another interesting case is that $p = 1$, which means that there is only one feature. In this case, $\boldsymbol{S}_{\mathrm{b}}$ and $\boldsymbol{S}_{\mathrm{w}} \in \mathbb{R}^{1 \times 1}$. Then, there exists $\gamma = \frac{B_{1,1}}{S_{1,1}}$ such that $\boldsymbol{S}_{\mathrm{b}} = [B_{1,1}] = [\gamma S_{1,1}] = \gamma [S_{1,1}] = \gamma \boldsymbol{S}_{\mathrm{w}}$. Therefore, the (P1) is trivial when $p = 1$.

Claim 13 *A non-zero column of $\boldsymbol{S}_{\mathrm{w}}$ satisfies $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} > \boldsymbol{0}$.*

Proof:
Let us denote the i-th column of $\boldsymbol{S}_{\mathrm{w}}$ to be $\boldsymbol{s}_i \in \mathbb{R}^p$. Since $\boldsymbol{S}_{\mathrm{w}}$ is symmetric **Claim 1**, then the i-th row of $\boldsymbol{S}_{\mathrm{w}}$ is $\boldsymbol{s}_i^\top \in \mathbb{R}^{1 \times p}$.
By assumption, suppose that the k-th column of $\boldsymbol{S}_{\mathrm{w}}$ is non-zero, which implies that: $\boldsymbol{s}_k \neq \boldsymbol{0}$.
Therefore:

$$\boldsymbol{S}_{\mathrm{w}}\boldsymbol{s}_k = \begin{bmatrix} \boldsymbol{s}_1^\top \\ \boldsymbol{s}_2^\top \\ \vdots \\ \boldsymbol{s}_k^\top \\ \vdots \\ \boldsymbol{s}_p^\top \end{bmatrix} \boldsymbol{s}_k = \begin{bmatrix} \star \\ \star \\ \vdots \\ ||\boldsymbol{s}_k||_2^2 \\ \vdots \\ \star \end{bmatrix} \tag{2.142}$$

Since $\boldsymbol{s}_k \neq \boldsymbol{0}$, $||\boldsymbol{s}_k||_2^2 > 0$, which shows that $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{s}_k \neq \boldsymbol{0}$.
Since $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{s}_k \neq \boldsymbol{0}$ and $\boldsymbol{S}_{\mathrm{w}}$ is positive semi-definite (**Claim 2**), by **Proposition 4** we know that $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} > \boldsymbol{0}$.

Claim 14 *If $\exists \gamma \in \mathbb{R}$, such that $\boldsymbol{S}_b = \gamma \boldsymbol{S}_w$, then:*

- (P1) *has trivial optimal solution status.*
- *A non-zero column of $\boldsymbol{S}_w$ is an optimal solution of* (P1).

Proof:
By **Claim 11** we know that • (P1) has trivial optimal solution status holds.
By **Claim 13** a non-zero column of $\boldsymbol{S}_{\mathrm{w}}$ satisfies $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} > \boldsymbol{0}$. Since $\boldsymbol{w}^\top \boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} > \boldsymbol{0}$ is the only constraint of (P1), then $\boldsymbol{w}$ is a feasible solution of (P1).
Since (P1) has trivial optimal solution status, by definition all feasible solutions of (P1) are optimal solutions of (P1).
Therefore, $\boldsymbol{w}$ is an optimal solution of (P1).

Claim 15 *If $\forall \gamma \in \mathbb{R}$, $\boldsymbol{S}_b \neq \gamma \boldsymbol{S}_w$, and* (P6) *has a solution, then:*

- (P1) *has non-trivial optimal solution status.*
- *Denote $\hat{\boldsymbol{w}}$ to be the solution of* (P6)*, then $\hat{\boldsymbol{w}}$ is an optimal solution of* (P1).

Proof:
The contrapositive of **Claim 12** tells us that if $\forall \gamma \in \mathbb{R}$, $\boldsymbol{S}_{\mathrm{b}} \neq \gamma \boldsymbol{S}_{\mathrm{w}}$, then (P1) has non-trivial optimal solution status.
Since $\hat{\boldsymbol{w}}$ is a solution of (P6), let $\hat{\beta} = 1$ and it is obvious that $(\hat{\boldsymbol{w}}, \hat{\beta})$ is a solution of (P5). By **Claim 6** we know that $\hat{\boldsymbol{w}}$ is an optimal solution of (P1).

Claim 16 *If $\forall \gamma \in \mathbb{R}$, $\boldsymbol{S}_b \neq \gamma \boldsymbol{S}_w$, and* (P6) *has no solution, then:*

- (P1) *has infinite optimal solution status.*
- *The optimal solution set of* (P1) *is empty.*

Proof:
Suppose that (P5) has a solution $(\hat{\boldsymbol{w}}, \hat{\beta})$, then it satisfies: $\hat{\beta}\boldsymbol{S}_{\mathrm{w}}\hat{\boldsymbol{w}} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. Rearrange the terms we have: $\boldsymbol{S}_{\mathrm{w}}(\hat{\beta}\hat{\boldsymbol{w}}) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, which implies that $\hat{\beta}\hat{\boldsymbol{w}}$ is a solution of (P6).
In a nutshell, if (P5) has a solution, then (P6) has a solution. By contrapositive we know that if (P6) has no solution, then (P5) has no solution.
Since the assumption has said that (P5) has no solution, **Claim 10** tells us that (P1) has infinite optimal solution status, and it is natural that the optimal solution set is empty.

Chapter 3

Model-based Prediction

3.1 Simplified Model

Suppose that the minimizer of $J_1(\boldsymbol{w})$ is $\boldsymbol{w}^*$, then the model is formulated as:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right| \leq \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right| \\ 1 & \text{if } \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right| > \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right| \end{cases} \tag{3.1}$$

which classifies the samples according to the distance to each projected cluster center.

We can simplify the model.

Since $\left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right|$ measures the distance between $\boldsymbol{w}^{*\top}\boldsymbol{x}$ and $\boldsymbol{w}^{*\top}\boldsymbol{\mu}_0$, and $\left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right|$ measures the distance between $\boldsymbol{w}^{*\top}\boldsymbol{x}$ and $\boldsymbol{w}^{*\top}\boldsymbol{\mu}_1$. Then $\left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right| \leq \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right|$ implies that $\boldsymbol{w}^{*\top}\boldsymbol{x}$ is closer to $\boldsymbol{w}^{*\top}\boldsymbol{\mu}_0$. Therefore, it is obvious that the decision boundary is the middle point $\frac{1}{2}\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$.

Therefore, the model is simplified as:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0 \\ 1 & \text{if } \left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) > 0 \end{cases} \tag{3.2}$$

3.2 Proof of Simplified Model

Claim 17 *Let*

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right| \leq \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right| \\ 1 & \text{if } \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0\right| > \left|\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1\right| \end{cases} \tag{3.1}$$

Then it can be rewritten as:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0 \\ 1 & \text{if } \left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) > 0 \end{cases} \tag{3.2}$$

Proof:

Let

$$z := \boldsymbol{w}^{*\top}\boldsymbol{x}, \tag{3.3}$$

$$m_0 := \boldsymbol{w}^{*\top}\boldsymbol{\mu}_0, \tag{3.4}$$

$$m_1 := \boldsymbol{w}^{*\top}\boldsymbol{\mu}_1. \tag{3.5}$$

We begin with the original definition of the classifier:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } |z - m_0| \leq |z - m_1| \\ 1 & \text{if } |z - m_0| > |z - m_1| \end{cases} \tag{3.6}$$

To proceed, we first square both sides:

$$|z - m_0| \leq |z - m_1| \tag{3.7}$$
$$(z - m_0)^2 \leq (z - m_1)^2 \tag{3.8}$$
$$z^2 - 2zm_0 + m_0^2 \leq z^2 - 2zm_1 + m_1^2 \tag{3.9}$$
$$-2z(m_0 - m_1) + (m_0^2 - m_1^2) \leq 0 \tag{3.10}$$
$$-2z(m_0 - m_1) + (m_0 - m_1)(m_0 + m_1) \leq 0 \tag{3.11}$$
$$(2z - (m_0 + m_1))\,(m_1 - m_0) \leq 0 \tag{3.12}$$

Substituting back the original terms:

$$\left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0 \tag{3.13}$$

This completes the proof.

Chapter 4

Example

4.1 Example 1

Question

Given a training dataset $\mathcal{I}$:

- $\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}} = \begin{bmatrix}0\\1\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 2}^{\mathcal{I}} = \begin{bmatrix}1\\2\end{bmatrix}$ $\boldsymbol{x}|_{\diamond 3}^{\mathcal{I}} = \begin{bmatrix}3\\5\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 4}^{\mathcal{I}} = \begin{bmatrix}3\\7\end{bmatrix}$
- $\boldsymbol{y}|_{\diamond 1}^{\mathcal{I}} = 0$, $y|_{\diamond 2}^{\mathcal{I}} = 0$ $y|_{\diamond 3}^{\mathcal{I}} = 1$, $y|_{\diamond 4}^{\mathcal{I}} = 1$

Model Fitting

Step 1: Split Dataset $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$

$\mathcal{I}_0_{\diamond 1} = \mathcal{I}_{\diamond 1}$, $\mathcal{I}_0_{\diamond 2} = \mathcal{I}_{\diamond 2}$, $\mathcal{I}_1_{\diamond 1} = \mathcal{I}_{\diamond 3}$, $\mathcal{I}_1_{\diamond 2} = \mathcal{I}_{\diamond 4}$

Step 2: Compute $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{S}_{\mathbf{w}}$

$$
\begin{aligned}
\boldsymbol{\mu}_0 &= \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_0} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_0}\right) = \frac{1}{2}\cdot\begin{bmatrix}1\\3\end{bmatrix} = \begin{bmatrix}0.5\\1.5\end{bmatrix} \\
\boldsymbol{\mu}_1 &= \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_1} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_1}\right) = \frac{1}{2}\cdot\begin{bmatrix}6\\12\end{bmatrix} = \begin{bmatrix}3\\6\end{bmatrix} \\
\boldsymbol{\Sigma}_0 &= \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^\top \\
&= \begin{bmatrix}0.25 & 0.25\\0.25 & 0.25\end{bmatrix} + \begin{bmatrix}0.25 & 0.25\\0.25 & 0.25\end{bmatrix} = \begin{bmatrix}0.5 & 0.5\\0.5 & 0.5\end{bmatrix} \\
\boldsymbol{\Sigma}_1 &= \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^\top \\
&= \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\0 & 2\end{bmatrix} \\
\boldsymbol{S}_{\mathrm{w}} &= \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \begin{bmatrix}0.5 & 0.5\\0.5 & 2.5\end{bmatrix}
\end{aligned}
$$

Step 3: Compute $\boldsymbol{S}_{\mathrm{b}}$, check if there exists $\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathbf{b}} = \gamma \boldsymbol{S}_{\mathbf{w}}$

$$
\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \begin{bmatrix}-2.5\\-4.5\end{bmatrix}\begin{bmatrix}-2.5 & -4.5\end{bmatrix} = \begin{bmatrix}6.25 & 11.25\\11.25 & 20.25\end{bmatrix}
$$

Does not exist.

Step 4: Solve $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 2.5 \end{bmatrix} \cdot \boldsymbol{w}^* = \begin{bmatrix} -2.5 \\ -4.5 \end{bmatrix}$$
$$\boldsymbol{w}^* = \begin{bmatrix} -4 \\ -1 \end{bmatrix}$$

(P1) has non-trivial optimal solution status, and $\boldsymbol{w}^*$ is an optimal solution of (P1).

Prediction

We consider the inequality:

$$\left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0$$

Given:

$$\boldsymbol{w}^* = \begin{bmatrix} -4 \\ -1 \end{bmatrix}, \quad \boldsymbol{\mu}_0 = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

Step 1:

Compute $\boldsymbol{w}^{*\top}\boldsymbol{x}$

$$\boldsymbol{w}^{*\top}\boldsymbol{x} = [-4, -1] \cdot \begin{bmatrix} [\boldsymbol{x}]_1 \\ [\boldsymbol{x}]_2 \end{bmatrix} = -4[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2$$

Step 2:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$

$$\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 = \begin{bmatrix} 0.5 + 3 \\ 1.5 + 6 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 7.5 \end{bmatrix}$$
$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1) = [-4, -1] \cdot \begin{bmatrix} 3.5 \\ 7.5 \end{bmatrix} = -4 \cdot 3.5 - 1 \cdot 7.5 = -14 - 7.5 = -21.5$$

Step 3:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \begin{bmatrix} 3 - 0.5 \\ 6 - 1.5 \end{bmatrix} = \begin{bmatrix} 2.5 \\ 4.5 \end{bmatrix}$$
$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = [-4, -1] \cdot \begin{bmatrix} 2.5 \\ 4.5 \end{bmatrix} = -4 \cdot 2.5 - 1 \cdot 4.5 = -10 - 4.5 = -14.5$$

Step 4:

Substitute all values into the inequality

$$(2(-4[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2) - (-21.5)) \cdot (-14.5) \leq 0$$

Simplify:

$$16[\boldsymbol{x}]_1 + 4[\boldsymbol{x}]_2 - 43 \leq 0$$

Final result:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \ 16[\boldsymbol{x}]_1 + 4[\boldsymbol{x}]_2 - 43 \leq 0 \\ 1 & \text{if } \ 16[\boldsymbol{x}]_1 + 4[\boldsymbol{x}]_2 - 43 > 0 \end{cases}$$

4.2 Example 2

Question

Given a training dataset $\mathcal{I}$:

- $\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}} = \begin{bmatrix}1\\2\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 2}^{\mathcal{I}} = \begin{bmatrix}2\\4\end{bmatrix}$ $\boldsymbol{x}|_{\diamond 3}^{\mathcal{I}} = \begin{bmatrix}3\\6\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 4}^{\mathcal{I}} = \begin{bmatrix}4\\8\end{bmatrix}$
- $\boldsymbol{y}|_{\diamond 1}^{\mathcal{I}} = 0$, $y|_{\diamond 2}^{\mathcal{I}} = 0$ $y|_{\diamond 3}^{\mathcal{I}} = 1$, $y|_{\diamond 4}^{\mathcal{I}} = 1$

Model Fitting

Step 1: Split Dataset $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$

$\mathcal{I}_{\diamond 1}^{0} = \mathcal{I}_{\diamond 1}$, $\mathcal{I}_{\diamond 2}^{0} = \mathcal{I}_{\diamond 2}$, $\mathcal{I}_{\diamond 1}^{1} = \mathcal{I}_{\diamond 3}$, $\mathcal{I}_{\diamond 2}^{1} = \mathcal{I}_{\diamond 4}$

Step 2: Compute $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{S}_{\mathbf{w}}$

$$
\begin{aligned}
\boldsymbol{\mu}_0 &= \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_0} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_0}\right) = \frac{1}{2}\left(\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}2\\4\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}3\\6\end{bmatrix} = \begin{bmatrix}1.5\\3\end{bmatrix} \\
\boldsymbol{\mu}_1 &= \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_1} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_1}\right) = \frac{1}{2}\left(\begin{bmatrix}3\\6\end{bmatrix} + \begin{bmatrix}4\\8\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}7\\14\end{bmatrix} = \begin{bmatrix}3.5\\7\end{bmatrix} \\
\boldsymbol{\Sigma}_0 &= \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^{\top} \\
&= \begin{bmatrix}0.25 & 0.5\\0.5 & 1\end{bmatrix} + \begin{bmatrix}0.25 & 0.5\\0.5 & 1\end{bmatrix} = \begin{bmatrix}0.5 & 1\\1 & 2\end{bmatrix} \\
\boldsymbol{\Sigma}_1 &= \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^{\top} \\
&= \begin{bmatrix}0.25 & 0.5\\0.5 & 1\end{bmatrix} + \begin{bmatrix}0.25 & 0.5\\0.5 & 1\end{bmatrix} = \begin{bmatrix}0.5 & 1\\1 & 2\end{bmatrix} \\
\boldsymbol{S}_{\mathrm{w}} &= \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \begin{bmatrix}1 & 2\\2 & 4\end{bmatrix}
\end{aligned}
$$

Step 3: Compute $\boldsymbol{S}_{\mathrm{b}}$, check if there exists $\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathbf{b}} = \gamma \boldsymbol{S}_{\mathbf{w}}$

$$
\begin{aligned}
\boldsymbol{S}_{\mathrm{b}} &= (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top} \\
&= \begin{bmatrix}4 & 8\\8 & 16\end{bmatrix}
\end{aligned}
$$

Thus, there exists $\gamma = 4$ such that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$.
(P1) has trivial optimal solution status, and a non-zero column of $\boldsymbol{S}_{\mathrm{w}}$ is an optimal solution of (P1).

$$\boldsymbol{w}^* = \begin{bmatrix}1\\2\end{bmatrix}$$

Prediction

We consider the inequality:

$$\left(2 \cdot \boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right) \cdot \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0$$

Given:

$$\boldsymbol{w}^* = \begin{bmatrix}1\\2\end{bmatrix}, \quad \boldsymbol{\mu}_0 = \begin{bmatrix}1.5\\3.5\end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix}3.5\\7\end{bmatrix}$$

Step 1:

Compute $\boldsymbol{w}^{*\top}\boldsymbol{x}$

$$\boldsymbol{w}^{*\top}\boldsymbol{x} = [1,2] \cdot \begin{bmatrix}[\boldsymbol{x}]_1 \\ [\boldsymbol{x}]_2\end{bmatrix} = 1[\boldsymbol{x}]_1 + 2[\boldsymbol{x}]_2$$

Step 2:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$

$$\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 = \begin{bmatrix}1.5 + 3.5 \\ 3.5 + 7\end{bmatrix} = \begin{bmatrix}5 \\ 10.5\end{bmatrix}$$

$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1) = [1,2] \cdot \begin{bmatrix}5 \\ 10.5\end{bmatrix} = 1 \times 5 + 2 \times 10.5 = 5 + 21 = 26$$

Step 3:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \begin{bmatrix}3.5 - 1.5 \\ 7 - 3.5\end{bmatrix} = \begin{bmatrix}2 \\ 3.5\end{bmatrix}$$

$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = [1,2] \cdot \begin{bmatrix}2 \\ 3.5\end{bmatrix} = 1 \times 2 + 2 \times 3.5 = 2 + 7 = 9$$

Step 4:

Substitute all values into the inequality

$$(2(1[\boldsymbol{x}]_1 + 2[\boldsymbol{x}]_2) - 26) \cdot 9 \leq 0$$

Simplify:

$$(2[\boldsymbol{x}]_1 + 4[\boldsymbol{x}]_2 - 26) \times 9 \leq 0$$

which is equivalent to

$$[\boldsymbol{x}]_1 + 2[\boldsymbol{x}]_2 - 13 \leq 0$$

Final result:

$$M(\boldsymbol{x}) = \begin{cases}0 & \text{if } [\boldsymbol{x}]_1 + 2[\boldsymbol{x}]_2 - 13 \leq 0 \\ 1 & \text{if } [\boldsymbol{x}]_1 + 2[\boldsymbol{x}]_2 - 13 > 0\end{cases}$$

‘

4.3 Example 3

Question

Given a training dataset $\mathcal{I}$:

- $\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}} = \begin{bmatrix}1 \\ 1 \\ 0\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 2}^{\mathcal{I}} = \begin{bmatrix}0 \\ -1 \\ 0\end{bmatrix}$ $\boldsymbol{x}|_{\diamond 3}^{\mathcal{I}} = \begin{bmatrix}-1 \\ 0 \\ 0\end{bmatrix}$, $\boldsymbol{x}|_{\diamond 4}^{\mathcal{I}} = \begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}$
- $\boldsymbol{y}|_{\diamond 1}^{\mathcal{I}} = 0$, $y|_{\diamond 2}^{\mathcal{I}} = 0$ $y|_{\diamond 3}^{\mathcal{I}} = 1$, $y|_{\diamond 4}^{\mathcal{I}} = 1$

Model Fitting

Step 1: Split Dataset $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$

$\mathcal{I}_{\diamond 1}^{0} = \mathcal{I}_{\diamond 1}$, $\mathcal{I}_{\diamond 2}^{0} = \mathcal{I}_{\diamond 2}$, $\mathcal{I}_{\diamond 1}^{1} = \mathcal{I}_{\diamond 3}$, $\mathcal{I}_{\diamond 2}^{1} = \mathcal{I}_{\diamond 4}$

Step 2: Compute $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{S}_{\mathbf{w}}$

$$\boldsymbol{\mu}_0 = \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_0} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_0}\right) = \frac{1}{2}\cdot\begin{bmatrix}1\\0\\0\end{bmatrix} = \begin{bmatrix}0.5\\0\\0\end{bmatrix}$$

$$\boldsymbol{\mu}_1 = \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_1} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_1}\right) = \frac{1}{2}\cdot\begin{bmatrix}-1\\0\\0\end{bmatrix} = \begin{bmatrix}-0.5\\0\\0\end{bmatrix}$$

$$\boldsymbol{\Sigma}_0 = \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^\top = \begin{bmatrix}0.5 & 1 & 0\\1 & 2 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\boldsymbol{\Sigma}_1 = \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^\top = \begin{bmatrix}0.5 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

$$\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \begin{bmatrix}1 & 1 & 0\\1 & 2 & 0\\0 & 0 & 0\end{bmatrix}$$

Step 3: Compute $\boldsymbol{S}_{\mathbf{b}}$, check if there exists $\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathbf{b}} = \gamma \boldsymbol{S}_{\mathbf{w}}$

$$\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

Does not exist.

Step 4: Solve $\boldsymbol{S}_{\mathbf{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$

We solve:

$$\begin{bmatrix}1 & 1 & 0\\1 & 2 & 0\\0 & 0 & 0\end{bmatrix}\boldsymbol{w}^* = \begin{bmatrix}1\\0\\0\end{bmatrix}$$

$$\boldsymbol{w}^* = \begin{bmatrix}2\\-1\\0\end{bmatrix} + \xi\begin{bmatrix}0\\0\\1\end{bmatrix}, \quad \xi \in \mathbb{R}.$$

We choose:

$$\boldsymbol{w}^* = \begin{bmatrix}2\\-1\\0\end{bmatrix}$$

(P1) has non-trivial optimal solution status, and $\boldsymbol{w}^*$ is an optimal solution of (P1).

Prediction

We consider the inequality:

$$\left(2\cdot\boldsymbol{w}^{*\top}\boldsymbol{x} - \boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\right)\cdot\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \leq 0$$

Given:

$$\boldsymbol{w}^* = \begin{bmatrix}2\\-1\\0\end{bmatrix}, \quad \boldsymbol{\mu}_0 = \begin{bmatrix}0.5\\0\\0\end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix}-0.5\\0\\0\end{bmatrix}$$

Step 1:

Compute $\boldsymbol{w}^{*\top}\boldsymbol{x}$

$$\boldsymbol{w}^{*\top}\boldsymbol{x} = [2, -1, 0] \cdot \begin{bmatrix} [\boldsymbol{x}]_1 \\ [\boldsymbol{x}]_2 \\ [\boldsymbol{x}]_3 \end{bmatrix} = 2[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2$$

Step 2:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$

$$\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 = \begin{bmatrix} 0.5 + (-0.5) \\ 0 + 0 \\ 0 + 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1) = [2, -1, 0] \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = 0$$

Step 3:

Compute $\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \begin{bmatrix} -0.5 - 0.5 \\ 0 - 0 \\ 0 - 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{w}^{*\top}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = [2, -1, 0] \cdot \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = -2$$

Step 4:

Substitute all values into the inequality

$$(2(2[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2) - 0) \times (-2) \leq 0$$

Simplify inside:

$$2[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2 \geq 0$$

Final result:

$$M(\boldsymbol{x}) = \begin{cases} 0 & \text{if } 2[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2 \geq 0 \\ 1 & \text{if } 2[\boldsymbol{x}]_1 - [\boldsymbol{x}]_2 < 0 \end{cases}$$

4.4 Example 4

Question

Given a training dataset $\mathcal{I}$:

- $\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$, $\boldsymbol{x}|_{\diamond 2}^{\mathcal{I}} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$ $\boldsymbol{x}|_{\diamond 3}^{\mathcal{I}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\boldsymbol{x}|_{\diamond 4}^{\mathcal{I}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- $\boldsymbol{y}|_{\diamond 1}^{\mathcal{I}} = 0$, $y|_{\diamond 2}^{\mathcal{I}} = 0$ $y|_{\diamond 3}^{\mathcal{I}} = 1$, $y|_{\diamond 4}^{\mathcal{I}} = 1$

Model Fitting

Step 1: Split Dataset $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$

$\mathcal{I}_{0\,\diamond 1}^{\mathcal{I}} = \mathcal{I}_{\diamond 1}^{\mathcal{I}}$, $\mathcal{I}_{0\,\diamond 2}^{\mathcal{I}} = \mathcal{I}_{\diamond 2}^{\mathcal{I}}$, $\mathcal{I}_{1\,\diamond 1}^{\mathcal{I}} = \mathcal{I}_{\diamond 3}^{\mathcal{I}}$, $\mathcal{I}_{1\,\diamond 2}^{\mathcal{I}} = \mathcal{I}_{\diamond 4}^{\mathcal{I}}$

Step 2: Compute $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{S}_{\mathbf{w}}$

$$\boldsymbol{\mu}_0 = \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_0} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_0}\right) = \frac{1}{2}\left(\begin{bmatrix}8\\6\end{bmatrix} + \begin{bmatrix}6\\2\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}14\\8\end{bmatrix} = \begin{bmatrix}7\\4\end{bmatrix}$$
$$\boldsymbol{\mu}_1 = \frac{1}{2}\left(\boldsymbol{x}|_{\diamond 1}^{\mathcal{I}_1} + \boldsymbol{x}|_{\diamond 2}^{\mathcal{I}_1}\right) = \frac{1}{2}\left(\begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$
$$\boldsymbol{\Sigma}_0 = \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_0} - \boldsymbol{\mu}_0\right)^\top = \begin{bmatrix}2 & 4\\4 & 8\end{bmatrix}$$
$$\boldsymbol{\Sigma}_1 = \sum_{i=1}^{2}\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}|_{\diamond i}^{\mathcal{I}_1} - \boldsymbol{\mu}_1\right)^\top = \begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$$
$$\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 = \begin{bmatrix}2 & 4\\4 & 8\end{bmatrix}$$

Step 3: Compute $\boldsymbol{S}_{\mathrm{b}}$, check if there exists $\gamma \in \mathbb{R}$ such that $\boldsymbol{S}_{\mathbf{b}} = \gamma \boldsymbol{S}_{\mathbf{w}}$

$$\boldsymbol{S}_{\mathrm{b}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top = \begin{bmatrix}49 & 28\\28 & 16\end{bmatrix}$$

Does not exist.

Step 4: Solve $\boldsymbol{S}_{\mathbf{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$

We solve:

$$\begin{bmatrix}2 & 4\\4 & 8\end{bmatrix}\boldsymbol{w}^* = \begin{bmatrix}7\\4\end{bmatrix}$$

This problem has no solution.
Since (P5) $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ has no solution, (P1) has infinite optimal solution status, and the optimal set is empty.

Chapter 5

Some other Important Properties

✏️ Solution Set of (P1)

Case $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ and (P6) has a solution:

In this case, there exists $\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{p-\text{rank}(k)} \in \mathbb{R}^p$, such that $\mathbb{D}^p$ is the solution set of (P6) $\boldsymbol{S}_{\text{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, where

$$\mathbb{D}^p = \left\{ \boldsymbol{v}_0 + \sum_{k=1}^{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \xi_k \boldsymbol{v}_k \,\middle|\, \xi_1, \cdots, \xi_{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \in \mathbb{R} \right\} \tag{5.1}$$

The optimal solution set of (P1) is the following equivalent sets:

$$\mathbb{A}^p = \{\boldsymbol{w} \in \mathbb{R}^p | \exists \beta \in \mathbb{R}, \text{such that } \beta \boldsymbol{S}_{\text{w}}\boldsymbol{w} = \boldsymbol{S}_{\text{b}}\boldsymbol{w}\} \tag{5.2}$$

$$\mathbb{B}^p = \left\{ \xi_0 \boldsymbol{v}_0 + \sum_{k=1}^{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \xi_k \boldsymbol{v}_k \,\middle|\, \text{where } \xi_0 \in \mathbb{R} \setminus \{0\} \text{ and } \xi_1, \cdots, \xi_{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \in \mathbb{R} \right\} \tag{5.3}$$

Notice that $\mathbb{A}^p$ is relevant to (P5).

Case $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$:

In this case, there exists $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{p-\text{rank}(k)} \in \mathbb{R}^p$, such that $\mathbb{W}^p$ is the solution set of (P6) $\boldsymbol{S}_{\text{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, where

$$\mathbb{W}^p = \left\{ \sum_{k=1}^{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \xi_k \boldsymbol{v}_k \,\middle|\, \xi_1, \cdots, \xi_{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \in \mathbb{R} \right\} \tag{5.4}$$

Then there exists $\boldsymbol{v}_{p-\text{rank}(\boldsymbol{S}_{\text{w}})+1}, \cdots, \boldsymbol{v}_p \in \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_p\} \subseteq \mathbb{R}^p$ such that $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_p \in \mathbb{R}^p$ are linearly independent. The readers can select $\boldsymbol{v}_{p-\text{rank}(\boldsymbol{S}_{\text{w}})+1}, \ldots, \boldsymbol{v}_p$ from $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ via a brute-force combinatorial search as follows:

- Consider all subsets $\mathbb{I} \subset \{1, 2, \ldots, p\}$ such that $|\mathbb{I}| = \text{rank}(\boldsymbol{S}_{\text{w}})$.
- For each candidate subset $\mathbb{I} = \{i_1, i_2, \ldots, i_{\text{rank}(\boldsymbol{S}_{\text{w}})}\}$, form the set of vectors $\{\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_{\text{rank}(\boldsymbol{S}_{\text{w}})}}\}$.
- Check whether
$$\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p-\text{rank}(\boldsymbol{S}_{\text{w}})}\} \cup \{\boldsymbol{e}_i : i \in \mathbb{I}\}$$
is a linearly independent set.
- If such a subset $\mathbb{I}$ is found, assign
$$(\boldsymbol{v}_{p-\text{rank}(\boldsymbol{S}_{\text{w}})+1}, \ldots, \boldsymbol{v}_p) = (\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_{\text{rank}(\boldsymbol{S}_{\text{w}})}}),$$
and terminate the search.

The optimal solution set of (P1) is the following equivalent set:

$$\mathbb{U}^p = \{\boldsymbol{w} \in \mathbb{R}^p | \boldsymbol{S}_{\text{w}}\boldsymbol{w} \neq \boldsymbol{0}\} \tag{5.5}$$

$$\mathbb{V}^p = \left\{ \sum_{k=1}^{p} \xi_k \boldsymbol{v}_k \,\middle|\, \text{where } \xi_1, \cdots, \xi_{p-\text{rank}(\boldsymbol{S}_{\text{w}})} \in \mathbb{R} \setminus \{0\} \text{ and } \xi_{p-\text{rank}(\boldsymbol{S}_{\text{w}})+1}, \cdots, \xi_p \in \mathbb{R} \right\} \tag{5.6}$$

Claim 18 *Let* $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ *and* (P5) *has solutions, then the optimal solution set of* (P1) *is:* $\{\boldsymbol{w} \in \mathbb{R}^p : \exists \beta \in \mathbb{R}, \text{such that } \beta \boldsymbol{S}_w \boldsymbol{w} = \boldsymbol{S}_b \boldsymbol{w}\}$

If $(\boldsymbol{w}, \beta)$ is the solution of (P5). By **Claim 6**, we have $\boldsymbol{w}$ is an optimal solution of (P1).
If for a specific $\boldsymbol{w}$, there does not exist $\beta \in \mathbb{R}$ such that $(\boldsymbol{w}, \beta)$ is a solution of (P5), then by contrapositive of **Claim 9** we have that $\boldsymbol{w}$ is not an optimal solution of (P1).
Therefore, the solution set is: $\{\boldsymbol{w} \in \mathbb{R}^p : \exists \beta \in \mathbb{R}, \text{such that } \beta \boldsymbol{S}_{\text{w}} \boldsymbol{w} = \boldsymbol{S}_{\text{b}} \boldsymbol{w}\}$.

We have shown that if $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ and (P5) have solutions. Then (P1) must have solutions (**Claim 6**).

Combine **Claim 6** and **Claim 9** we conclude that $\hat{\boldsymbol{w}}$ is an optimal solution of (P1) if and only if there $\exists \hat{\beta} \in \mathbb{R}$ such that $(\hat{\boldsymbol{w}}, \hat{\beta})$ is a solution of (P5).

Claim 19 *If there exists* $\gamma \in \mathbb{R}$ *such that* $\boldsymbol{S}_{\text{b}} = \gamma \boldsymbol{S}_{\text{w}}$*, then (P5) has solutions.*

Proof:
Case $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$:
We can have: $\boldsymbol{S}_{\text{b}} = \boldsymbol{0}$. Then there exists $\gamma = 0$ such that $\boldsymbol{S}_{\text{b}} = \gamma \boldsymbol{S}_{\text{w}}$.
$\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 = \boldsymbol{0}$. Therefore, $(\boldsymbol{0}, 0)$ can be a solution of (P5).
This shows that (P5) has solutions when $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$.
Case $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$:

$$\gamma \boldsymbol{S}_{\text{w}} = \boldsymbol{S}_{\text{b}} \tag{5.7}$$
$$\gamma \boldsymbol{S}_{\text{w}} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \tag{5.8}$$
$$\gamma \boldsymbol{S}_{\text{w}} \neq \boldsymbol{0} \tag{5.9}$$
$$\gamma \neq 0 \tag{5.10}$$

- (5.7): This is by assumption.
- (5.7) ⇒ (5.8): This comes directly from (1.10).
- (5.8) ⇒ (5.9): Since $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$, $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \neq \boldsymbol{0}$.
- (5.9) ⇒ (5.10): Since $\gamma \boldsymbol{S}_{\text{w}} \neq \boldsymbol{0}$, $\gamma \neq 0$.

Since $\boldsymbol{S}_{\text{w}}$ is non-zero then there exists $\boldsymbol{w}_0 \in \mathbb{R}^p$ such that:

$$\boldsymbol{S}_{\text{w}} \boldsymbol{w}_0 \neq \boldsymbol{0} \tag{5.11}$$

Then:

$$\boldsymbol{S}_{\text{w}} \boldsymbol{w}_0 \neq \boldsymbol{0} \tag{5.12}$$
$$\gamma \boldsymbol{S}_{\text{w}} \boldsymbol{w}_0 \neq \boldsymbol{0} \tag{5.13}$$
$$\boldsymbol{S}_{\text{b}} \boldsymbol{w}_0 \neq \boldsymbol{0} \tag{5.14}$$
$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \neq \boldsymbol{0} \tag{5.15}$$
$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \neq 0 \tag{5.16}$$

- (5.12): This comes directly from (5.11).
- (5.12) ⇒ (5.13): Multiply both sides with γ and the equality still holds because $\gamma \neq 0$ by (5.10).
- (5.13) ⇒ (5.14): This comes from the assumption that $\boldsymbol{S}_{\text{b}} = \gamma \boldsymbol{S}_{\text{w}}$.
- (5.14) ⇒ (5.15): This comes directly from the definition of $\boldsymbol{S}_{\text{b}}$ (1.10).
- (5.15) ⇒ (5.16): Since $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \neq \boldsymbol{0}$, $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \neq 0$.

Then:

$$\gamma \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}_0 = \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w}_0 \tag{5.17}$$

$$\frac{\gamma}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}_0 = \frac{1}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} \boldsymbol{S}_{\mathrm{b}} \boldsymbol{w}_0 \tag{5.18}$$

$$\frac{\gamma}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}_0 = \frac{1}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \tag{5.19}$$

$$\frac{\gamma}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}_0 = \frac{1}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \tag{5.20}$$

$$\frac{\gamma}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0} \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w}_0 = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.21}$$

- (5.17): This comes directly from the assumption that $\boldsymbol{S}_{\mathrm{b}} = \gamma \boldsymbol{S}_{\mathrm{w}}$
- (5.17) $\Rightarrow$ (5.18): We can divide both sides with $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0$ because $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0 \neq 0$ (5.16).
- (5.18) $\Rightarrow$ (5.19): This comes directly from the definition of $\boldsymbol{S}_{\mathrm{b}}$ (1.10).
- (5.19) $\Rightarrow$ (5.20): Since $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0$ is a scalar, we can rearrange it.

From (5.21) we can see that $\left(\frac{\gamma}{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{w}_0}, \boldsymbol{w}_0\right)$ is a solution of (P5).

Claim 20 *Let $\boldsymbol{\mu_0} \neq \boldsymbol{\mu_1}$ and (P6) has a solution, then:*

- *There exists $\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{p-\mathrm{rank}(k)} \in \mathbb{R}^p$, such that $\mathbb{D}^p$ is the solution set of (P6) $\boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$*

$$\mathbb{D}^p = \left\{ \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_w)} \xi_k \boldsymbol{v}_k \,\middle|\, \xi_1, \cdots, \xi_{p-\mathrm{rank}(\boldsymbol{S}_w)} \in \mathbb{R} \right\} \tag{5.1}$$

- *The following two sets $\mathbb{A}^p$ and $\mathbb{B}^p$ are equivalent:*

$$\mathbb{A}^p = \{ \boldsymbol{w} \in \mathbb{R}^p | \exists \beta, \textit{such that } \beta \boldsymbol{S}_{\mathrm{w}} \boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \} \tag{5.2}$$

$$\mathbb{B}^p = \left\{ \xi_0 \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_w)} \xi_k \boldsymbol{v}_k \,\middle|\, \textit{where } \xi_0 \in \mathbb{R} \setminus \{0\} \textit{ and } \xi_1, \cdots, \xi_{p-\mathrm{rank}(\boldsymbol{S}_w)} \in \mathbb{R} \right\} \tag{5.3}$$

Proof:
The proof of the first bullet point is ignored, but you can refer to the linear algebra textbook.
($\boldsymbol{w}_0 \in \mathbb{B}^p \Rightarrow \boldsymbol{w}_0 \in \mathbb{A}^p$):
Suppose that $\boldsymbol{w}_0 \in \mathbb{B}^p$. Then there exists $\bar{\xi}_0 \in \mathbb{R} \setminus \{0\}$ and $\bar{\xi}_k \in \mathbb{R}$ for $k = 1, 2, \cdots, p - \mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})$, such that:

$$\boldsymbol{w}_0 = \bar{\xi}_0 \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \bar{\xi}_k \boldsymbol{v}_k \tag{5.22}$$

Since $\xi_0 \neq 0$:

$$\frac{1}{\bar{\xi}_0} \boldsymbol{w}_0 = \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \frac{\bar{\xi}_k}{\bar{\xi}_0} \boldsymbol{v}_k \tag{5.23}$$

Since $\bar{\xi}_k \in \mathbb{R}$ and $\bar{\xi}_0 \in \mathbb{R} \setminus \{0\}$, $\frac{\bar{\xi}_k}{\bar{\xi}_0} \in \mathbb{R}$ as well. This implies that $\frac{1}{\bar{\xi}_0}\boldsymbol{w}_0 \in \mathbb{D}^p$, which is exactly the solution set of $\boldsymbol{S}_{\mathrm{w}}w = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. Therefore, $\frac{1}{\bar{\xi}_0}\boldsymbol{w}_0$ is the solution of $\boldsymbol{S}_w\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ the following must hold:

$$\boldsymbol{S}_w\left(\frac{1}{\bar{\xi}_0}\boldsymbol{w}_0\right) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.24}$$

Then there $\exists\beta = \frac{1}{\bar{\xi}_0}$, such that:

$$\beta\boldsymbol{S}_w\boldsymbol{w}_0 = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.25}$$

which shows that $\boldsymbol{w}_0 \in \mathbb{A}^p$.
($\boldsymbol{w}_0 \in \mathbb{A}^p \Rightarrow \boldsymbol{w}_0 \in \mathbb{B}^p$):
Suppose that $\boldsymbol{w}_0 \in \mathbb{A}^p$, then there exists β_0 such that:

$$\beta_0\boldsymbol{S}_{\mathrm{w}}w_0 = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.26}$$
$$\boldsymbol{S}_{\mathrm{w}}(\beta_0 w_0) = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.27}$$

which shows that $\beta_0 w_0$ is a solution of $\boldsymbol{S}_{\mathrm{w}}w = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, then it should have the form of (5.1), which means that there exists $\bar{\xi}_k \in \mathbb{R}$ for $k = 1, 2, \cdots, p - \mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})$.

$$\beta_0\boldsymbol{w}_0 = \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \bar{\xi}_k\boldsymbol{v}_k \tag{5.28}$$

By assumption $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$, then $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \neq \boldsymbol{0}$. Combine this with (5.26) we can conclude that $\beta_0 \neq 0$.
Then:

$$\beta_0\boldsymbol{w}_0 = \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \bar{\xi}_k\boldsymbol{v}_k \tag{5.29}$$
$$\boldsymbol{w}_0 = \frac{1}{\beta_0}\boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \frac{\bar{\xi}_k}{\beta_0}\boldsymbol{v}_k \tag{5.30}$$

Since $\frac{1}{\beta_0} \in \mathbb{R} \setminus \{0\}$ and $\frac{\bar{\xi}_k}{\beta_0} \in \mathbb{R}$, we can conclude that $\boldsymbol{w}_0 \in \mathbb{B}^p$.
Since ($\boldsymbol{w}_0 \in \mathbb{B}^p \Rightarrow \boldsymbol{w}_0 \in \mathbb{A}^p$) and ($\boldsymbol{w}_0 \in \mathbb{A}^p \Rightarrow \boldsymbol{w}_0 \in \mathbb{B}^p$) we can conclude that $\mathbb{A}^p$ and $\mathbb{B}^p$ are equivalent.

Claim 21 *Let $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$, and $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ has solutions, and the solution set of $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ to be:*

$$\mathbb{D}^p = \left\{ \boldsymbol{w} \in \mathbb{R}^p \,\middle|\, \boldsymbol{w} = \boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_w)} \xi_k\boldsymbol{v}_k, \text{ where } \xi_1, \cdots, \xi_{p-\mathrm{rank}(\boldsymbol{S}_w)} \in \mathbb{R} \right\} \tag{5.1}$$

where:

$$\boldsymbol{S}_w\boldsymbol{v}_0 = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \tag{5.31}$$
$$\boldsymbol{S}_w\boldsymbol{v}_k = \boldsymbol{0} \tag{5.32}$$

then the optimal solution set of (P1) *is the following two equivalent sets:*

- $\mathbb{A} = \{\boldsymbol{w} \in \mathbb{R}^p : \exists\beta \in \mathbb{R}, \text{such that } \beta\boldsymbol{S}_w\boldsymbol{w} = \boldsymbol{S}_b\boldsymbol{w}\}$
- $\mathbb{B} = \left\{ \boldsymbol{w} \in \mathbb{R}^p \,\middle|\, \boldsymbol{w} = \xi_0\boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_w)} \xi_k\boldsymbol{v}_k, \text{ where } \xi_0 \in \mathbb{R} \setminus \{0\} \text{ and } \xi_1, \cdots, \xi_{p-\mathrm{rank}(\boldsymbol{S}_w)} \in \mathbb{R} \right\}$

Proof:
In the assumption $\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ has solutions, which also implies that (P5) $\beta\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ has solutions.
Since $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ and (P5) has solutions, we can apply **Claim 18** which claims that the solution set of (P1) is $\{\boldsymbol{w} \in \mathbb{R}^p : \exists\beta \in \mathbb{R}, \text{such that } \beta\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}\}$.
Then we can apply **Claim 20** which shows that $\{\boldsymbol{w} \in \mathbb{R}^p : \exists\beta \in \mathbb{R}, \text{such that } \beta\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w} = \boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}\}$ is equivalent to $\left\{\boldsymbol{w} \in \mathbb{R}^p \,\middle|\, \boldsymbol{w} = \xi_0\boldsymbol{v}_0 + \sum_{k=1}^{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \xi_k\boldsymbol{v}_k, \text{ where } \xi_0 \in \mathbb{R} \setminus \{0\} \text{ and } \xi_1, \cdots, \xi_{p-\mathrm{rank}(\boldsymbol{S}_{\mathrm{w}})} \in \mathbb{R}\right\}$.
This completes the proof.

Chapter 6

Appendix: Calculus and Linear Algebra

Definition 1 *Let $\boldsymbol{A}$ be an $n \times n$ square matrix over a field $\mathbb{F}$. A matrix $\boldsymbol{S}_{\mathrm{b}}$ is called the **inverse** of $\boldsymbol{A}$, denoted $\boldsymbol{A}^{-1}$, if it satisfies:*

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.
*A matrix possessing an inverse is said to be **invertible** or **non-singular**.*
*If no such matrix $\boldsymbol{S}_{\mathrm{b}}$ exists for $\boldsymbol{A}$, then $\boldsymbol{A}$ is called **non-invertible** or **singular**.*

Definition 2 *A real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is called:*

- ***positive definite**, if $\boldsymbol{A} = \boldsymbol{A}^\top$ (symmetric) and $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0$ for all non-zero $\boldsymbol{v} \in \mathbb{R}^n$.*
- ***positive semi-definite**, if $\boldsymbol{A} = \boldsymbol{A}^\top$ (symmetric) and $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \geq 0$ for all non-zero $\boldsymbol{v} \in \mathbb{R}^n$;*

Theorem 1 (Fundamental Subspace Theorem)

$$\mathrm{Row}(A)^\perp = \mathrm{Ker}(A) \tag{6.1}$$

Proof:

1. $\mathrm{Ker}(A) \subseteq \mathrm{Row}(A)^\perp$
2. $\mathrm{Row}(A)^\perp \subseteq \mathrm{Ker}(A)$

1. Proof of $\mathrm{Ker}(A) \subseteq \mathrm{Row}(A)^\perp$:
Let $\boldsymbol{x} \in \mathrm{Ker}(A)$, which means $A\boldsymbol{x} = \boldsymbol{0}$.
For any $\boldsymbol{v} \in \mathrm{Row}(A)$, we can write $\boldsymbol{v}$ as a linear combination of the rows of A:

$$\boldsymbol{v} = c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_m\boldsymbol{a}_m$$

where $\boldsymbol{a}_i$ are the row vectors of A.
Compute the inner product:

$$\boldsymbol{v} \cdot \boldsymbol{x} = (c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_m\boldsymbol{a}_m) \cdot \boldsymbol{x} = c_1(\boldsymbol{a}_1 \cdot \boldsymbol{x}) + \cdots + c_m(\boldsymbol{a}_m \cdot \boldsymbol{x})$$

Since $A\boldsymbol{x} = \boldsymbol{0}$, we have $\boldsymbol{a}_i \cdot \boldsymbol{x} = 0$ for all i, thus:

$$\boldsymbol{v} \cdot \boldsymbol{x} = 0$$

This shows $\boldsymbol{x}$ is orthogonal to every vector in $\mathrm{Row}(A)$, so:

$$\mathrm{Ker}(A) \subseteq \mathrm{Row}(A)^\perp$$

2. Proof of $\mathrm{Row}(A)^\perp \subseteq \mathrm{Ker}(A)$:

Let $\boldsymbol{x} \in \text{Row}(A)^\perp$, meaning $\boldsymbol{x}$ is orthogonal to all vectors in $\text{Row}(A)$.
Since $\text{Row}(A)$ contains all row vectors of A, we have:

$$\boldsymbol{a}_i \cdot \boldsymbol{x} = 0 \quad \text{(for all rows } \boldsymbol{a}_i \text{ of } A)$$

This implies:

$$A\boldsymbol{x} = \boldsymbol{0}$$

Therefore $\boldsymbol{x} \in \text{Ker}(A)$, and we conclude:

$$\text{Row}(A)^\perp \subseteq \text{Ker}(A)$$

From both inclusions, we obtain:

$$\text{Row}(A)^\perp = \text{Ker}(A)$$

Theorem 2 (Double Orthogonal Complement Theorem) *For any subspace $V \subseteq \mathbb{R}^n$,*

$$(V^\perp)^\perp = V \tag{6.2}$$

Proof:

1. $V \subseteq (V^\perp)^\perp$
2. $(V^\perp)^\perp \subseteq V$

1. Proof of $V \subseteq (V^\perp)^\perp$:
Let $\boldsymbol{v} \in V$. By definition of orthogonal complement, for any $\boldsymbol{w} \in V^\perp$ we have $\boldsymbol{v} \cdot \boldsymbol{w} = 0$.
This means $\boldsymbol{v}$ is orthogonal to every vector in $V^\perp$, which by definition implies $\boldsymbol{v} \in (V^\perp)^\perp$. Therefore:

$$V \subseteq (V^\perp)^\perp$$

2. Proof of $(V^\perp)^\perp \subseteq V$:
Let $\boldsymbol{x} \in (V^\perp)^\perp$. We can decompose $\mathbb{R}^n$ as:

$$\mathbb{R}^n = V \oplus V^\perp$$

Thus, $\boldsymbol{x}$ can be uniquely written as:

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{v}^\perp \quad \text{where } \boldsymbol{v} \in V,\ \boldsymbol{v}^\perp \in V^\perp$$

Since $\boldsymbol{x} \in (V^\perp)^\perp$, it must be orthogonal to all vectors in $V^\perp$. In particular:

$$\boldsymbol{x} \cdot \boldsymbol{v}^\perp = (\boldsymbol{v} + \boldsymbol{v}^\perp) \cdot \boldsymbol{v}^\perp = \boldsymbol{v} \cdot \boldsymbol{v}^\perp + \boldsymbol{v}^\perp \cdot \boldsymbol{v}^\perp = 0 + \|\boldsymbol{v}^\perp\|^2 = 0$$

This implies $\|\boldsymbol{v}^\perp\|^2 = 0$, so $\boldsymbol{v}^\perp = \boldsymbol{0}$. Therefore $\boldsymbol{x} = \boldsymbol{v} \in V$, proving:

$$(V^\perp)^\perp \subseteq V$$

Combining both inclusions, we conclude:

$$(V^\perp)^\perp = V$$

Theorem 3 (Cholesky Decomposition) *If a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive semi-definite, then there exists a lower triangular matrix $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$.*

Proof: Ignored.

Theorem 4 (Generalized Fritz John Necessary Conditions) *Consider the following mathematical programming problem:*

$$\boxed{\begin{aligned} &\text{(F)} \quad \underset{\boldsymbol{x}}{Minimize} \quad f(x) && (6.3)\\ &\qquad Subject\ to \quad g_i(x) \le 0, \quad i \in \{1,2,\dots,m\} && (6.4)\\ &\qquad\qquad\quad\ \ h_j(x) = 0, \quad j \in \{1,2,\dots,n\} && (6.5)\end{aligned}}$$

where $f(x)$, $g_i(x)$, and $h_j(x)$ are functions defined on the n-dimensional Euclidean space $\mathbb{R}^n$ and have continuous first partial derivatives on $\mathbb{R}^n$.
If this problem has minimum, and $\bar{x}$ is an optimal solution, then there exist vectors

$$\bar{\beta} \in \mathbb{R}, \quad \bar{\boldsymbol{u}} = (\bar{u}_1, \dots, \bar{u}_m)^\top \in \mathbb{R}^m, \quad \bar{\boldsymbol{v}} = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_k)^\top \in \mathbb{R}^k$$

such that

$$\bar{\beta}\nabla\theta(\bar{x}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{j=1}^{k} \bar{v}_j \nabla h_j(\bar{x}) = 0 \tag{6.6}$$

$$\sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}) = 0 \tag{6.7}$$

$$\bar{\beta} \ge 0 \tag{6.8}$$

$$\bar{u}_i \ge 0 \qquad \forall i \in \{1, \cdots, m\} \tag{6.9}$$

$$\begin{bmatrix} \bar{\beta} \\ \bar{\boldsymbol{u}} \\ \bar{\boldsymbol{v}} \end{bmatrix} \neq 0 \tag{6.10}$$

Proof: `https://pages.cs.wisc.edu/~olvi/oldpapers/MFCQ.pdf`

Proposition 1 *Consider the following two mathematical programming problem:*

$$\boxed{\begin{aligned} &\text{(P)} \quad \underset{\boldsymbol{x}}{Minimize} \quad f(x) && (6.11)\\ &\qquad Subject\ to \quad g_i(x) \le 0, \quad i \in \{1,2,\dots,m\} && (6.12)\\ &\qquad\qquad\quad\ \ h_j(x) = 0, \quad j \in \{1,2,\dots,n\} && (6.13)\end{aligned}}$$

$$\boxed{\begin{aligned} &\text{(D)} \quad \underset{\boldsymbol{x},\beta,\boldsymbol{u},\boldsymbol{v}}{Minimize} \quad f(\boldsymbol{x}) && (6.14)\\ &\qquad Subject\ to \quad \beta \cdot \nabla\theta(\boldsymbol{x}) + \sum_{i=1}^{m} [\boldsymbol{u}]_i \cdot \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{k} [\boldsymbol{v}]_j \cdot \nabla h_j(\boldsymbol{x}) = 0 && (6.15)\\ &\qquad\qquad\quad\ \ \sum_{i=1}^{m} [\boldsymbol{u}]_i \cdot g_i(\boldsymbol{x}) = 0 && (6.16)\\ &\qquad\qquad\quad\ \ g_i(\boldsymbol{x}) \le 0, \quad i \in \{1,2,\dots,m\} && (6.17)\\ &\qquad\qquad\quad\ \ h_j(\boldsymbol{x}) = 0, \quad j \in \{1,2,\dots,n\} && (6.18)\\ &\qquad\qquad\quad\ \ \beta \ge 0 && (6.19)\\ &\qquad\qquad\quad\ \ [\boldsymbol{u}]_i \ge 0, \quad i \in \{1,2,\dots,m\} && (6.20)\\ &\qquad\qquad\quad\ \ \begin{bmatrix} \beta \\ \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} \neq 0 && (6.21)\end{aligned}}$$

where $f(\boldsymbol{x})$, $g_i(\boldsymbol{x})$, and $h_j(\boldsymbol{x})$ are functions defined on the n-dimensional Euclidean space $\mathbb{R}^n$ and have continuous first partial derivatives on $\mathbb{R}^n$.
Suppose that (P) has optimal solutions:

- *$\Rightarrow$: If $\boldsymbol{x}^*$ is an optimal solution of (P), then there $\exists\ \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ such that $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is an optimal solution of (D).*
- *$\Leftarrow$: If $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is an optimal solution of (D), then $\boldsymbol{x}^*$ is an optimal solution of (P).*

Proof ($\Rightarrow$):
If $\boldsymbol{x}^*$ is an optimal solution of (P), then by **Theorem 4** there exists $\beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ such that $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ satisfying (6.6), (6.7), (6.8), (6.9), (6.10).
Since $\boldsymbol{x}^*$ is an optimal solution of (P), it must be feasible as well, which means that $\boldsymbol{x}^*$ should satisfy (6.12), (6.13).
Combine these we know that: If $\boldsymbol{x}^*$ is an optimal solution of (P), then there exists $\beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ such that $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is the solution of (6.6), (6.7), (6.8), (6.9), (6.10), (6.12), (6.13) (which is equivalent to (6.15), (6.16), (6.19),(6.20),(6.21),(6.17),(6.18)). Therefore, we know that there exists $\beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*$ such that $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is a feasible solution of (D).
Denote all feasible solutions of (D) to be $(\boldsymbol{x}, \beta, \boldsymbol{u}, \boldsymbol{v})$, since the (6.17), and (6.18) hold (which are equivalent to (6.12),(6.13)), $\boldsymbol{x}$ is a feasible solution of (P) as well.
Now, $\boldsymbol{x}$ is a feasible solution of (P), $\boldsymbol{x}^*$ is an optimal solution of (P), which implies that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$. This means that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ holds for all feasible solutions in (D).
Therefore, $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is an optimal solution of (D) because $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ holds for all feasible solutions in (D) and $(\boldsymbol{x}^*, \beta^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ is a feasible solution of (D).
Proof ($\Leftarrow$):
By **Theorem 4** and feasibility, we know that if a solution of (P) is optimal, it must satisfy (6.6),(6.7),(6.8),(6.9),(6.10),(6.12),(6.13). Then the optimal solutions are the solutions with minimal objective values and satisfying (6.6),(6.7),(6.8),(6.9),(6.10),(6.12),(6.13).
Therefore, the optimal solution of (D) has minimal objective values among those who exactly satisfying (6.6),(6.7),(6.8),(6.9),(6.10),(6.12),(6.13) since these are equivalent to (6.15) $\sim$ (6.21).
This completes the proof.
To help the readers understand the relationships between these solutions, we draw a Venn Chart in Figure 6.1, where:

- All: All points in $\mathbb{R}^n$
- Feasible: All feasible solutions of (P).
- Stationary: All feasible solutions of (P) which satisfying (6.6),(6.7),(6.8),(6.9),(6.10).
- Optimal: The optimal solutions of (P).

Proposition 2 *Let $\boldsymbol{A} \in \mathbb{F}^{m\times p}$, $\boldsymbol{C} \in \mathbb{F}^{q\times r}$, and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n \in \mathbb{F}^{p\times q}$. Then the following identity holds:*

$$\boldsymbol{A}\left(\sum_{i=1}^{n} \boldsymbol{B}_i\right)\boldsymbol{C} = \sum_{i=1}^{n} \boldsymbol{A}\boldsymbol{B}_i\boldsymbol{C}$$

Proof:
We begin by expanding the left-hand side of the equation:

$$\boldsymbol{A}\left(\sum_{i=1}^{n} \boldsymbol{B}_i\right)\boldsymbol{C} = \boldsymbol{A}\left(\boldsymbol{B}_1 + \boldsymbol{B}_2 + \cdots + \boldsymbol{B}_n\right)\boldsymbol{C}$$

Using the distributive property of matrix multiplication over addition, we can distribute $\boldsymbol{A}$ and $\boldsymbol{C}$ over the sum:

$$= \boldsymbol{A}\boldsymbol{B}_1\boldsymbol{C} + \boldsymbol{A}\boldsymbol{B}_2\boldsymbol{C} + \cdots + \boldsymbol{A}\boldsymbol{B}_n\boldsymbol{C}$$

Thus, we obtain the right-hand side of the equation:

$$= \sum_{i=1}^{n} \boldsymbol{A}\boldsymbol{B}_i\boldsymbol{C}$$

This completes the proof.

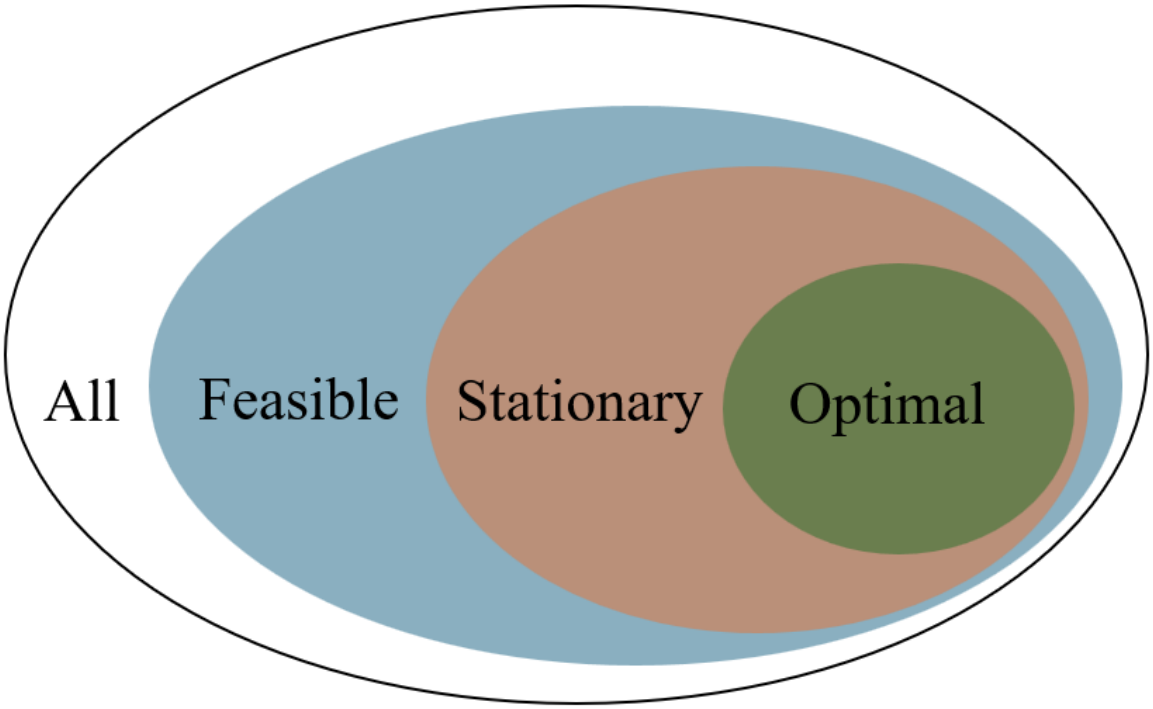

Figure 6.1: Venn Chart for Generalized Fritz John Condition

Proposition 3 *If a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric, positive semi-definite, and of rank one, then there exists a vector $\boldsymbol{u} \in \mathbb{R}^n$ such that $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{u}^\top$.*

Proof:
Since $\boldsymbol{A}$ is symmetric and positive semi-definite, there exists a matrix $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ such that

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top, \tag{6.22}$$

where $\boldsymbol{L}$ is lower triangular with non-negative diagonal entries (by the Cholesky decomposition **Theorem 3**).
Given that $\text{rank}(\boldsymbol{A}) = 1$, it follows that $\boldsymbol{A}$ has rank one, so $\boldsymbol{L}$ must also have rank one. Thus, $\boldsymbol{L}$ has exactly one nonzero column, say the first column, and all other columns are zero vectors. Denote this nonzero column by $\boldsymbol{u} \in \mathbb{R}^n$. Then $\boldsymbol{L}$ can be written as

$$\boldsymbol{L} = \boldsymbol{u}\boldsymbol{e}_1^\top, \tag{6.23}$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^n$ is the first standard basis vector.
Substituting back, we have

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top = (\boldsymbol{u}\boldsymbol{e}_1^\top)(\boldsymbol{e}_1\boldsymbol{u}^\top) = \boldsymbol{u}(\boldsymbol{e}_1^\top\boldsymbol{e}_1)\boldsymbol{u}^\top = \boldsymbol{u}\boldsymbol{u}^\top, \tag{6.24}$$

as desired.

Proposition 4 *Let $\boldsymbol{A}$ to be a positive semi-definite matrix:*

- $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$ *if and only if* $\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} = 0$
- $\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$ *if and only if* $\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} > 0$

Proof of $(\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} = 0)$:
If $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$, then $\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{0} = 0$, which completes the proof.

Proof of ($\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = 0 \Rightarrow \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$):
$\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = 0$. Since $\boldsymbol{A}$ is positive semi-definite, there exists $\boldsymbol{L}$ such that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ (**Theorem 3**):

$$\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = 0 \tag{6.25}$$
$$\boldsymbol{v}^\top \boldsymbol{L}\boldsymbol{L}^\top \boldsymbol{v} = 0 \tag{6.26}$$
$$(\boldsymbol{L}^\top \boldsymbol{v})^\top (\boldsymbol{L}^T \boldsymbol{v}) = 0 \tag{6.27}$$
$$(\boldsymbol{L}^\top \boldsymbol{v})^2 = 0 \tag{6.28}$$
$$\boldsymbol{L}^\top \boldsymbol{v} = \boldsymbol{0} \tag{6.29}$$

Continue:

$$\boldsymbol{A}\boldsymbol{w} = \boldsymbol{L}\boldsymbol{L}^\top \boldsymbol{w} = \boldsymbol{L}(\boldsymbol{L}^\top \boldsymbol{w}) = \boldsymbol{L} \cdot \boldsymbol{0} = \boldsymbol{0} \tag{6.30}$$

which completes the proof.

Proof of ($\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0} \Rightarrow \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0$) :
Since we have proved that if $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = 0$, then $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$. Therefore, when $\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$, $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \neq 0$.
Since $\boldsymbol{A}$ is positive semi-definite, by **Definition 2**' we have $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \geq 0$. Combine $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \neq 0$ and $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \geq 0$ we have $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0$, which completes the proof.

Proof of ($\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0 \Rightarrow \boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$):
Since we have proved that $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$, $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = 0$. If $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0$, we know that $\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$, which completes the proof.