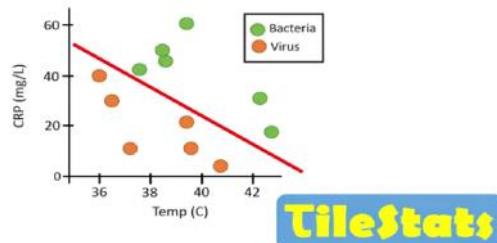


LDA



TileStats

Welcome to this lecture about linear discriminant analysis, a.k.a. LDA.

LDA

- The basics of LDA
- Applications
- LDA vs PCA
- The math

In this lecture, we will have a look at the basics of LDA and some of its applications. We will also discuss the differences between LDA and PCA, and go through the mathematical details of LDA.

PCA vs LDA

PCA vs LDA

LDA is very similar to PCA and I assume that you have watched the lecture about PCA before watching this video. To understand the content of this video, you need to understand the details behind PCA.

PCA vs LDA

Remember that PCA is a method to find the linear combination that accounts for as much variability as possible.

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

For example, let's say that we like to combine the two variables, X1 and X2.

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

PCA will then try to find the optimal values for the weights, alpha 1 and alpha 2,

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

so that the combined variable has maximal variance.

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

$$\alpha_1^2 + \alpha_2^2 = 1$$

Remember that the values we use for the weights are constrained so that the sum of the squared weights is equal to one.

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

$$\alpha_1^2 + \alpha_2^2 = 1$$

Linear discriminant analysis (LDA) is a similar method, which aims to maximize the separation between two or more groups.

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

$$\alpha_1^2 + \alpha_2^2 = 1$$

$$LD = \boxed{\alpha_1} X_1 + \boxed{\alpha_2} X_2$$

LDA then combines these two variables by using optimal values of the weights.

II ▶ 1:30 / 24:26 • Introduction > 🔍 ⌂ ⚙ ⚙

PCA vs LDA

$$PC = \alpha_1 X_1 + \alpha_2 X_2$$

$$\alpha_1^2 + \alpha_2^2 = 1$$

$$\boxed{LD} = \alpha_1 X_1 + \alpha_2 X_2$$

so that the combined variable shows maximal separation between the groups.

Separation

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

So what do we mean by a separation between two or more groups?

Example data

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

To explain separation, let's have a look at the following fictive data set, which contains information on 12 patients.

Example data

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



This column shows the concentration of the c-reactive protein in blood from the time when the patients entered the hospital,

Example data

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



whereas this column shows the body temperature of the same patients at the same time point.

Example data

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



Once the patients had entered the hospital, the presence of bacteria and virus were analyzed from different samples. However, it usually takes several hours or days to determine if a patient has a viral infection or a bacterial infection.

Example data

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



After two days at the hospital, these six patients were found to have a viral infection,

Example data

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



whereas these six patients were confirmed to have a bacterial infection. Since antibiotics are only effective on bacteria, only these patients were treated with antibiotics.

Example data

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



The problem is that we have to wait two days to know if antibiotic treatment is appropriate or not.

Example data

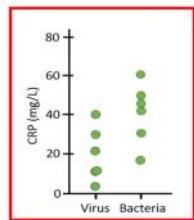
Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



It would be nice if we could use the CRP concentration or the body temperature to tell if a patient has a bacterial or viral infection because the measurement of these variables can be done within just an hour.

Example data

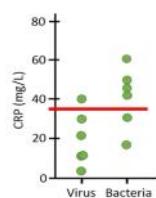
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



If we plot the CRP concentration of the 12 patients, we can see that the ones with a viral infection generally have a lower concentration of CRP compared to the ones who have a bacterial infection.

Separation

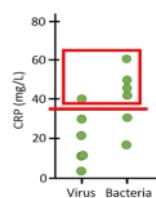
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



Let's say that we would use a cutoff value of 35, to determine if someone has a bacterial or viral infection. However, the problem with this cutoff line is that it cannot separate the groups very well.

Separation

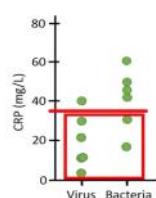
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



We can see that four patients with bacterial infection and one patient with viral infection are above this line,

Separation

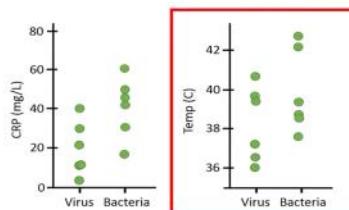
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



whereas five patients with a viral infection, and two patients with bacterial infection are below the line. Using this line, or any other horizontal line, it is impossible to separate the two groups completely.

Separation

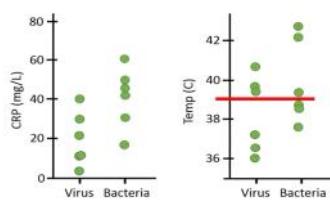
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



Let's check if it is possible to use the body temperature to tell if a patient has a viral or bacterial infection?

Separation

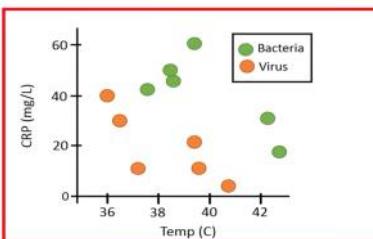
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



As you see, we have the same problem again. There is no line that can separate the two groups of patients.

Separation

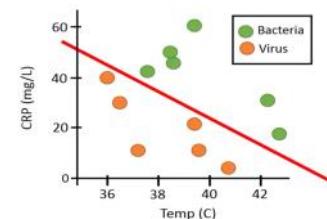
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



However, if we plot the CRP concentration and the body temperature in the same plot and use different colors for patients with a viral infection and bacterial infection,

Separation

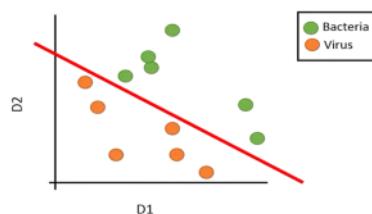
Infection	CRP (mg/L)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



we can see that the following line can separate the two groups completely. By combining the two variables, we can get a better separation between the two groups compared to if we use one of the variables alone.

LDA

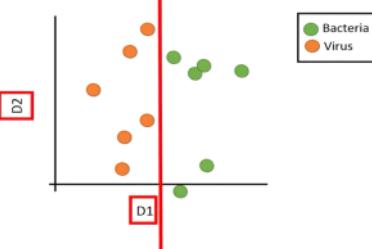
Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



Just as PCA, LDA can be seen as we rotate the data into two new dimensions like this.

LDA

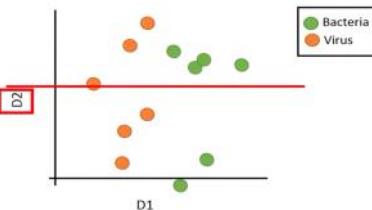
Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



The new dimensions are here called discriminant function one and two.

LDA

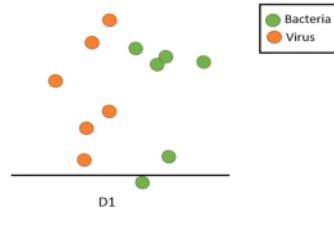
Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



In contrast, we cannot separate the groups based on a horizontal line, which means that the second discriminant function is not useful for separation.

LDA

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



We can therefore delete the second discriminant function and place all data points on just one line.

LDA

Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



We can therefore delete the second discriminant function and place all data points on just one line.

LDA

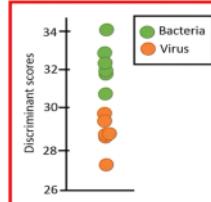
Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



For illustrative purposes, we here rotate the plot so that the data points are plotted vertically.

LDA

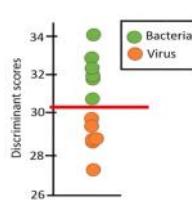
Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



If we use LDA, we can plot the so-called discriminant scores like this.

LDA

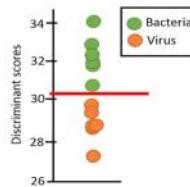
Infection	CRP (mg/l)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



We can now place a line like this that gives us a perfect separation between the two groups. By using LDA, we have combined the variables CRP and body temperature in a way that has maximized the separation between the two groups.

LDA

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

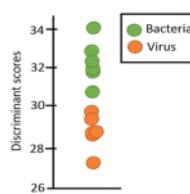


We can now place a line like this that gives us a perfect separation between the two groups. By using LDA, we have combined the variables CRP and body temperature in a way that has maximized the separation between the two groups.

LDA

$$LD = \alpha_1 X_1 + \alpha_2 X_2$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



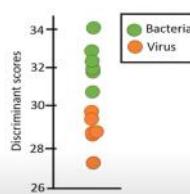
The linear discriminant scores that you see in the plot can be obtained by the following equation.

Linear discriminant analysis (LDA) - simply explained

LDA

$$LD = [0.11] \cdot CRP + [0.70] \cdot Temp$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7



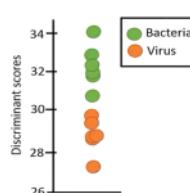
The values of our weights, alpha 1 and alpha 2, were computed to 0.11 and 0.70. These values give us the optimal separation between the groups. We will later see how these weights are calculated.

II ▶ ⏪ 6.23 / 24:26 • Introduction >

LDA

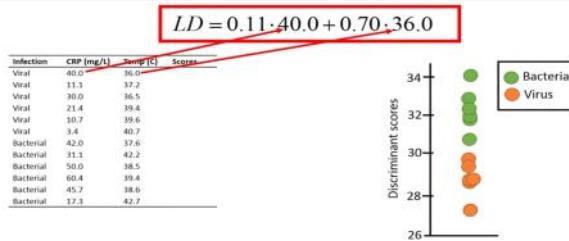
$$LD = 0.11 \cdot CRP + 0.70 \cdot Temp$$

Infection	CRP (mg/L)	Temp (C)	scores
Viral	40.0	36.0	
Viral	11.1	37.2	
Viral	30.0	36.5	
Viral	21.4	39.4	
Viral	10.7	39.6	
Viral	3.4	40.7	
Bacterial	42.0	37.6	
Bacterial	31.1	42.2	
Bacterial	50.0	38.5	
Bacterial	60.4	39.4	
Bacterial	45.7	38.6	
Bacterial	17.3	42.7	



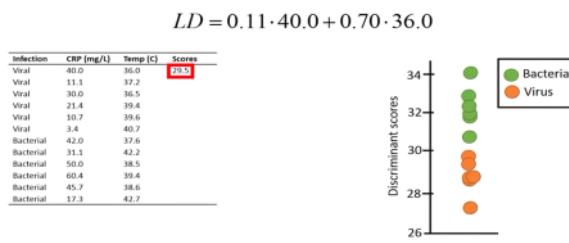
Let's use this equation to calculate the discriminant scores.

LDA



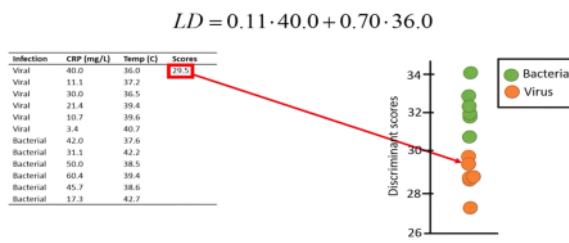
If we plug in the CRP value and the body temperature for the first patient in this equation,

LDA



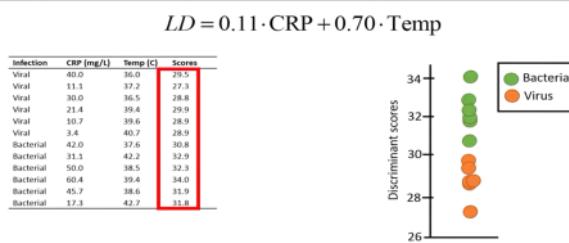
we see that the corresponding discriminant score is about 29.5.

LDA



This score is represented by the following point in the plot.

LDA

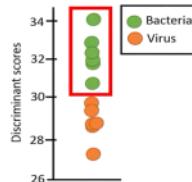


We use the same equation to compute the scores for the other patients as well.

LDA

$$LD = 0.11 \cdot CRP + 0.70 \cdot Temp$$

Infection	CRP (mg/L)	Temp [C]	Score
Viral	40.0	36.0	29.5
Viral	11.1	37.2	27.3
Viral	30.0	36.5	28.8
Viral	21.4	39.4	29.9
Viral	45.7	39.6	28.9
Viral	3.4	40.7	28.9
Bacterial	42.0	37.6	30.8
Bacterial	31.1	42.2	32.9
Bacterial	50.0	38.5	32.3
Bacterial	60.4	39.4	34.0
Bacterial	45.7	38.6	31.9
Bacterial	17.3	42.7	31.8



compared to the ones who have a bacterial infection.

LDA

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Infection	CRP (mg/L)	Temp [C]	Score
Viral	40.0	36.0	29.5
Viral	11.1	37.2	27.3
Viral	30.0	36.5	28.8
Viral	21.4	39.4	29.9
Viral	45.7	39.6	28.9
Viral	3.4	40.7	28.9
Bacterial	42.0	37.6	30.8
Bacterial	31.1	42.2	32.9
Bacterial	50.0	38.5	32.3
Bacterial	60.4	39.4	34.0
Bacterial	45.7	38.6	31.9
Bacterial	17.3	42.7	31.8

Most software tools report the centered discriminant scores, which means that we should subtract the corresponding means from the original values before we calculate the scores.

LDA

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Infection	CRP (mg/L)	Temp [C]	Score
Viral	40.0	36.0	29.5
Viral	11.1	37.2	27.3
Viral	30.0	36.5	28.8
Viral	21.4	39.4	29.9
Viral	45.7	39.6	28.9
Viral	3.4	40.7	28.9
Bacterial	42.0	37.6	30.8
Bacterial	31.1	42.2	32.9
Bacterial	50.0	38.5	32.3
Bacterial	60.4	39.4	34.0
Bacterial	45.7	38.6	31.9
Bacterial	17.3	42.7	31.8

Most software tools report the centered discriminant scores, which means that we should subtract the corresponding means from the original values before we calculate the scores.

LDA

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Infection	CRP (mg/L)	Temp [C]	Score	Cent. score
Viral	40.0	36.0	29.5	-1.1
Viral	11.1	37.2	27.3	-3.3
Viral	30.0	36.5	28.8	-1.8
Viral	21.4	39.4	29.9	-0.7
Viral	45.7	39.6	28.9	-1.7
Viral	3.4	40.7	28.9	1.7
Bacterial	42.0	37.6	30.8	0.2
Bacterial	31.1	42.2	32.9	2.3
Bacterial	50.0	38.5	32.3	1.7
Bacterial	60.4	39.4	34.0	3.5
Bacterial	45.7	38.6	31.9	1.3
Bacterial	17.3	42.7	31.8	1.2

By using this equation, the following centered scores were calculated.

LDA

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Infection	CRP (mg/L)	Temp (C)	Scores	Cent. scores
Viral	40.0	36.0	29.5	-1.1
Viral	11.1	37.2	27.3	-3.3
Viral	30.0	36.5	28.8	-1.8
Viral	21.4	39.4	29.9	-0.7
Viral	34.7	39.6	28.9	-1.7
Viral	3.4	40.7	28.9	1.7
Bacterial	42.0	37.6	30.8	0.2
Bacterial	31.1	42.2	32.9	2.3
Bacterial	50.0	38.5	32.3	1.7
Bacterial	60.4	39.4	34.0	3.5
Bacterial	45.7	38.6	31.9	1.3
Bacterial	17.3	42.7	31.8	1.2

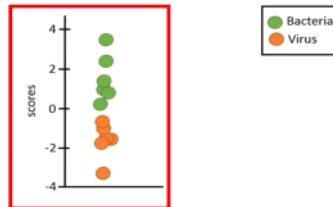


These centered scores can also be obtained by centering our previous scores.

LDA

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Infection	CRP (mg/L)	Temp (C)	Scores	Cent. scores
Viral	40.0	36.0	29.5	-1.1
Viral	11.1	37.2	27.3	-3.3
Viral	30.0	36.5	28.8	-1.8
Viral	21.4	39.4	29.9	-0.7
Viral	34.7	39.6	28.9	-1.7
Viral	3.4	40.7	28.9	1.7
Bacterial	42.0	37.6	30.8	0.2
Bacterial	31.1	42.2	32.9	2.3
Bacterial	50.0	38.5	32.3	1.7
Bacterial	60.4	39.4	34.0	3.5
Bacterial	45.7	38.6	31.9	1.3
Bacterial	17.3	42.7	31.8	1.2



If we plot the centered scores, we see that the scores centre around zero since the mean of these scores is equal to zero.

LDA vs PCA

LDA vs PCA

We will now discuss the main differences between LDA and PCA.

Linear discriminant analysis (LDA) – simply explained



Infection	CRP (mg/L)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	-0.7	1.7
var	1	1

To make a fair comparison between the two methods, we will first standardize our two variables so that they both have a mean of zero and a standard deviation of one.

8:02 / 24:26 • LDA vs PCA >

LDA vs PCA

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	0.7	1.7
var	1	1

$$Z_i = \frac{X_i - \bar{X}}{SD}$$

Remember that we can standardize a variable by the following formula, where we subtract the mean from the values of that variable and divide by its standard deviation.

LDA vs PCA

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	0.7	1.7
var	1	1

$$Z_i = \frac{40.0 - 30.3}{17.8}$$

For example, this first standardized value has been calculated by subtracting the mean of the variable CRP, 30.3, from the CRP value of the first person, which is 40.0, and divide by the standard deviation, which is 17.8 for the CRP variable.

LDA vs PCA

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	0.7	1.7
var	1	1

$$Z_i = \frac{40.0 - 30.3}{17.8}$$

For example, this first standardized value has been calculated by subtracting the mean of the variable CRP, 30.3, from the CRP value of the first person, which is 40.0, and divide by the standard deviation, which is 17.8 for the CRP variable.

LDA vs PCA

$$\begin{aligned} LD1 &= 0.79 \cdot zCRP - 0.61 \cdot zTemp \\ PC1 &= -0.71 \cdot zCRP + 0.71 \cdot zTemp \end{aligned}$$

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	0.7	1.7
var	1	1

By using the standardized data, I have here used a statistical software to compute the following weights from LDA and PCA.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	-0.7	1.7
var	1	1

Note that the variables represent standardized variables with a mean of zero and a standard deviation of one.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	-0.7	1.7
var	1	1

By using a software, the following weights were extracted from the first eigenvector of the covariance matrix when the PCA was computed.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)
Viral	0.5	-1.4
Viral	-1.1	-0.9
Viral	0.0	-1.2
Viral	-0.5	0.2
Viral	-1.1	0.3
Viral	-1.5	0.8
Bacterial	0.7	-0.7
Bacterial	0.0	1.5
Bacterial	1.1	-0.3
Bacterial	1.7	0.2
Bacterial	0.9	-0.2
Bacterial	-0.7	1.7
var	1	1

These weights, from the LDA, were also extracted from the first eigenvector. However, this eigenvector is not an eigenvector of the covariance matrix. Instead, it is an eigenvector of another type of matrix that we will discuss later.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	
			LDA	PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.1	-0.9	0.4	0.1
Viral	0.0	-1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	-0.3	-0.9	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29

By using these equations with the given weights, we can compute the scores based on LDA and PCA.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	
			LDA	PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29

Note that the variance of the scores for the first principal component is 1.29,

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	
			LDA	PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29

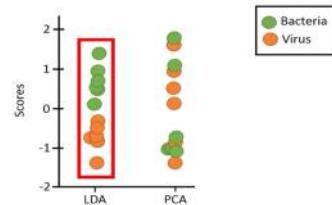
but only 0.71 for the scores based on the first discriminant function. This is no surprise since PCA aims to maximize the variance of the first principal component, whereas LDA instead aims to maximize the separation of the two groups.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	
			LDA	PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



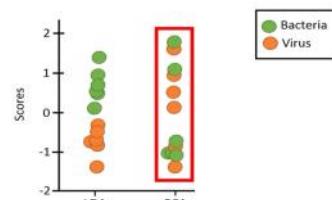
If we plot the scores computed by LDA, we see that the scores for the bacteria group are completely separated from the scores of the virus group.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	
			LDA	PCA
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.8	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



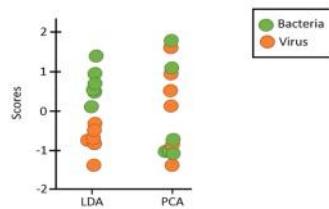
In comparison, the scores computed by the PCA show no clear separation between the groups. Although PCA has captured more variation in the combined variable, it does not separate the groups very well.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	Scores
	LDA	PCA		
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.4	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



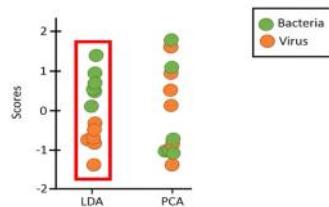
Note that, when we run PCA, we will not use this grouping variable at all, whereas in LDA, we must provide this grouping variable since the method aims to find the maximal separation between these two groups.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	Scores
	LDA	PCA		
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.4	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



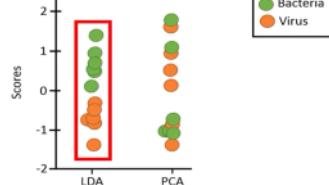
Since PCA will only be based on these two variables, without any information about the groups, it will not be as effective as LDA in separating the groups.

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	Scores
	LDA	PCA		
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.4	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



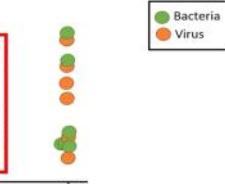
PCA is sometimes referred to as an unsupervised method because it has no information if a data point belongs to a certain group,

LDA vs PCA

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

Infection	CRP (mg/l)	Temp (C)	Scores	Scores
	LDA	PCA		
Viral	0.5	-1.4	-0.5	-1.4
Viral	-1.2	0.9	-1.4	0.1
Viral	0.0	1.2	-0.4	-0.8
Viral	-0.5	0.2	-0.3	0.5
Viral	-1.1	0.3	-0.7	1.0
Viral	-1.5	0.8	-0.7	1.6
Bacterial	0.7	-0.7	0.1	-0.9
Bacterial	0.0	1.5	1.0	1.0
Bacterial	1.1	0.3	0.7	-1.0
Bacterial	1.7	0.2	1.4	-1.1
Bacterial	0.9	-0.2	0.6	-0.8
Bacterial	-0.7	1.7	0.5	1.8
var	1	1	0.71	1.29



whereas LDA is referred to as a supervised method because it has the information about which group a data point belongs to in order to find the best separation between the groups.

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

We will now focus on how we can calculate the weights for the LDA. Remember that the weights for the PCA were extracted from the first eigenvector of the covariance matrix. These weights can be rescaled to loadings.

The weights

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

We will now focus on how we can calculate the weights for the LDA. Remember that the weights for the PCA were extracted from the first eigenvector of the covariance matrix. These weights can be rescaled to loadings.

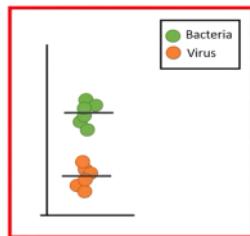
The weights

$$PC1 = -0.71 \cdot zCRP + 0.71 \cdot zTemp$$

$$LD1 = 0.79 \cdot zCRP + 0.61 \cdot zTemp$$

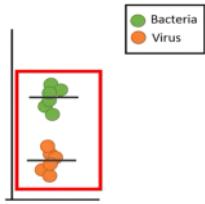
The weights in LDA are also extracted from the first eigenvector, but an eigenvector of a different matrix. These weights are also usually rescaled.

Separation between the groups



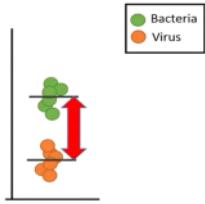
Let's first have a look at what determines how well two groups are separated.

Separation between the groups



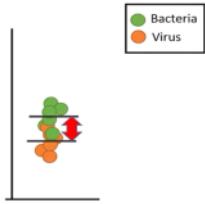
For this data, we can see that the two groups separate very well.

Separation between the groups



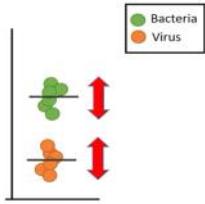
This is due to that the means of the groups are far away from each other.

Separation between the groups



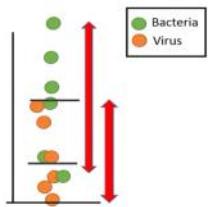
If the means are closer to each other, we can no longer see a clear separation between the two groups.

Separation between the groups



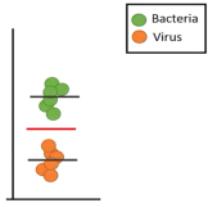
The second reason why these two groups show a good separation is because the spread within the groups is small. The data points are close to the mean values.

Separation between the groups



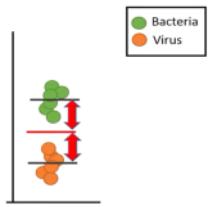
However, if there is a much larger spread of the observations around the means, then there is no longer a clear separation between the groups even though the difference in their means stay the same.

Separation between the groups



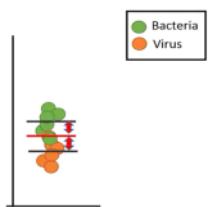
If we calculate a grand mean, which is a mean based on all data points,

Separation between the groups



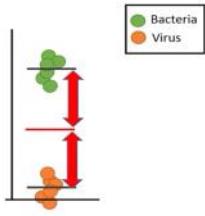
we can think of these distances as how much the group means varies from the grand mean.

Separation between the groups



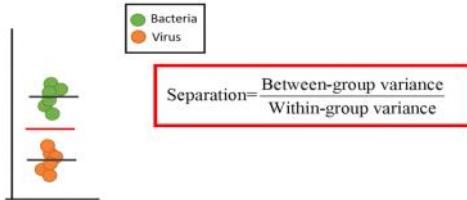
If the means are close, the variation of the group means around the grand mean is small,

Separation between the groups



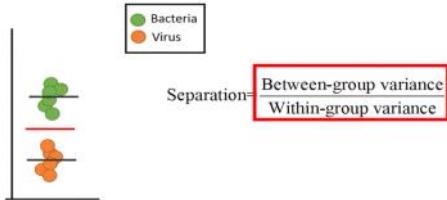
and if the two means are far away from each other, there will be a large variation of the group means around the grand mean.

Separation between the groups



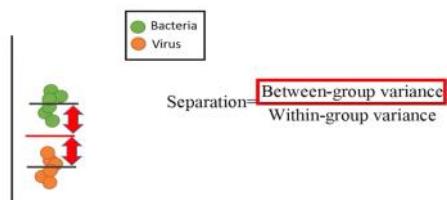
We can conclude that the separation between the groups depends on the variance between the groups, which is here called between-group variance, and the variance within the groups, the within-group variance.

Separation between the groups



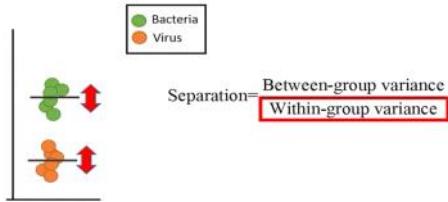
Note that this ratio is similar to the ratio used for ANOVA.

Separation between the groups



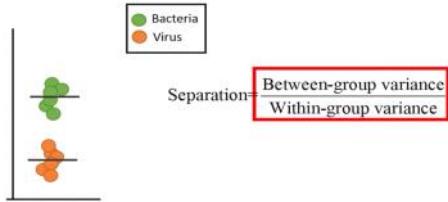
To get a good separation between the groups, we like that the two group means are far away from each other, which means that the between-group variance should be large,

Separation between the groups



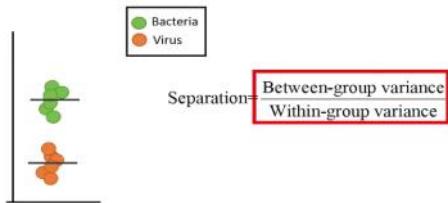
and that the variation within the groups should be small, which will result in a low value of the within-group variance.

Separation between the groups



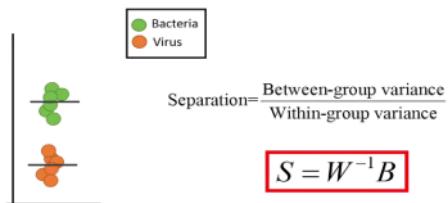
Thus, the ratio of these measures should be as large as possible to get a clear separation between the groups.

Separation between the groups



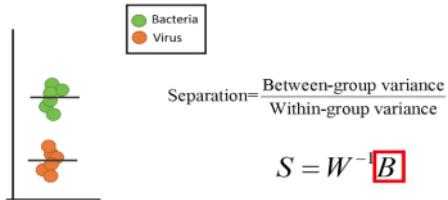
LDA combines variables based on this ratio, where the aim is to transform the data so that the between-group variance is increased and the within-group variance is reduced.

Separation between the groups



In matrix form, the separation can be described like this,

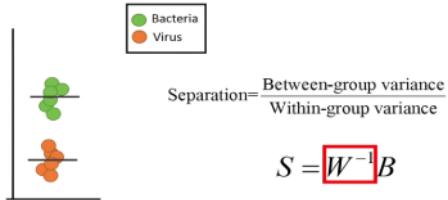
Separation between the groups



$$S = W^{-1} \boxed{B}$$

where B is the between-group covariance matrix,

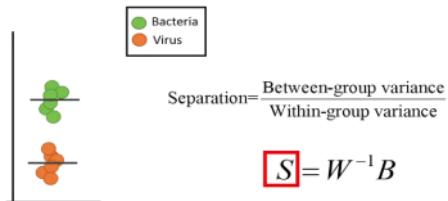
Separation between the groups



$$S = \boxed{W}^{-1} B$$

and W is the pooled within-group covariance matrix.

The math behind LDA



$$\boxed{S} = W^{-1} B$$

Matrix S is our target matrix. We will compute the eigenvectors of this matrix in order to get our weights.

The math behind LDA

$$\boxed{S} = W^{-1} B$$

To calculate matrix S , we first need to calculate the pooled within-group covariance matrix and the between-group covariance matrix.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

To show the calculations, we will use our original example data.

Pooled within-group covariance matrix

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

To calculate the pooled within-group covariance matrix, we first calculate the covariance matrix for the virus group,

Pooled within-group covariance matrix

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

and then for the bacteria group. Note that the values in these two matrices have been rounded.

Pooled within-group covariance matrix

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

Since the sample sizes are equal between the two groups, the pooled within-group covariance matrix is simply the mean of these two covariance matrices.

$$\text{cov}_{\text{virus}} = \begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix}$$

$$\text{cov}_{\text{Bacteria}} = \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}$$

$$W = \frac{\begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix} + \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}}{2} = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

Pooled within-group covariance matrix

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$\text{cov}_{\text{virus}} = \begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix}$$

$$W = \frac{\begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix} + \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}}{2} = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$\text{cov}_{\text{Bacteria}} = \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}$$

$$W^* = \frac{(n_1 - 1)\text{cov}(A) + (n_2 - 1)\text{cov}(B)}{n_1 + n_2 - 2}$$

If the two groups have an unequal sample size, we should use the following equation to compute the weighted average of the two covariance matrices.

Pooled within-group covariance matrix

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

Let's put the pooled within-group covariance matrix up here.

Between-group covariance matrix

$$S = W^{-1}\boxed{B}$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

We will now try to calculate the between-group covariance matrix.

Between-group covariance matrix

$$S = W^{-1}B$$

$$\boxed{T = B + W}$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	45.7	38.6
Bacterial	17.3	42.7

A simple way to calculate the between-group covariance matrix is to first calculate the total covariance because

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

the between-group covariance is equal to the total covariance minus the pooled within-group covariance.

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

This is the total covariance matrix, which shows the variance and covariance of the two variables when we do not separate the groups.

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

The total covariance matrix, is therefore the covariance matrix based on all data points.

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

This matrix is the same covariance matrix that we would use for computing the PCA.

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

$$B = \boxed{\begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}} - \boxed{\begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}} = \boxed{\begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}}$$

We can now calculate the between-group covariance matrix,

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

$$B = \boxed{\begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}} - \boxed{\begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}} = \boxed{\begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}}$$

by subtracting the pooled within-group covariance matrix,

Between-group covariance matrix

$$S = W^{-1}B$$

$$T = B + W$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$B = T - W$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

$$B = \boxed{\begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}} - \boxed{\begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}} = \boxed{\begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}}$$

from the total covariance matrix.

Between-group covariance matrix

$$S = W^{-1}B$$

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

$$B = \boxed{\begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}}$$

Let's put the between-group covariance matrix here.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$W^{-1} = \boxed{\begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}}$$

$$B = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

Next, we calculate the inverse of the pooled within-group covariance matrix.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$W^{-1} = \boxed{\begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}}$$

$$B = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

We are now ready to calculate matrix S.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$W^{-1} = \boxed{\begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}}$$

$$B = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

$$\boxed{S} = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix} \cdot \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix} = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

We calculate matrix S by

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$W^{-1} = \boxed{\begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}}$$

$$B = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

$$S = \boxed{\begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}} \cdot \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix} = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

multiplying the inverse of the pooled within-group covariance matrix,

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

$$W^{-1} = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}$$

$$B = \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix} \begin{bmatrix} 108.9 & 11.5 \\ 11.5 & 0.32 \end{bmatrix} = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

by the between-group covariance matrix.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

Once we have calculated matrix S, the rest of the calculations are similar to those used in PCA. However, instead of calculating the eigenvalues and eigenvectors of the covariance matrix, we will calculate these on the following matrix.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

By using a software, the following eigenvectors were computed.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

This is the first eigenvector of matrix S,

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

whereas this is the second eigenvector.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

$$\text{LD1} = 0.150 \cdot \text{CRP} + 0.989 \cdot \text{Temp}$$

We then combine the two variables, the CRP and the body temperature, by using the weights that are provided by the first eigenvector.

The math behind LDA

$$S = W^{-1}B$$

Infection	CRP (mg/l)	Temp (C)
Viral	60.0	36.0
Viral	13.1	37.2
Viral	30.0	36.5
Viral	23.4	39.4
Viral	10.7	39.9
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	33.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$S = \begin{bmatrix} 2.05 & 0.16 \\ 14.15 & 0.96 \end{bmatrix}$$

$$\text{Eigenvectors} = \begin{bmatrix} 0.150 & -0.074 \\ 0.989 & 0.997 \end{bmatrix}$$

$$\text{LD1} = 0.150 \cdot \text{CRP} + 0.989 \cdot \text{Temp}$$

By using this equation, with the given weights, we can calculate the discriminant scores that will show the best linear separation between the groups.

The math behind LDA

$$\text{LD1} = 0.150 \cdot \text{CRP} + 0.989 \cdot \text{Temp}$$

Infection	CRP (mg/l)	Temp (C)	Scores
Viral	60.0	36.0	
Viral	13.1	37.2	
Viral	30.0	36.5	
Viral	23.4	39.4	
Viral	10.7	39.9	
Viral	3.4	40.7	
Bacterial	42.0	37.6	
Bacterial	33.1	42.2	
Bacterial	50.0	38.5	
Bacterial	60.4	39.4	
Bacterial	45.7	38.6	
Bacterial	17.3	42.7	

We move the equation up here and use it to calculate the discriminant scores based on the CRP concentration and the body temperature of each patient.

Calculate the scores

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP (mg/l)	Temp (C)	Score
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

These are the calculated unstandardized scores.

Calculate the scores

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP (mg/l)	Temp (C)	Score
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

We will now rescale these weights so that they correspond to the weights or loadings that are commonly shown by most statistical software tools.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP (mg/l)	Temp (C)	Score
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

In LDA, the weights are usually rescaled so that the pooled group variance of the scores is equal to one.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP (mg/l)	Temp (C)	Score
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

Let's calculate the variance of the scores for the virus group.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.0
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

$$\text{var}(\text{scores})_{\text{Viral}} = 1.60$$

The variance of these scores is about 1.60,

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.0
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

$$\text{var}(\text{scores})_{\text{Viral}} = 1.60$$

$$\text{var}(\text{scores})_{\text{Bacterial}} = 2.37$$

and the variance of the scores for the bacteria group is 2.37.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.0
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

$$\text{var}(\text{scores})_{\text{Viral}} = 1.60$$

$$\text{var}(\text{scores})_{\text{pooled}} = 1.989$$

$$\text{var}(\text{scores})_{\text{Bacterial}} = 2.37$$

The average of these variances is 1.989, which is the pooled variance since the two groups have an equal sample size. If the two groups have an unequal sample size, we need to calculate a weighted mean instead.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores
Viral	60.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.0
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.8	45.0
Bacterial	17.3	42.7	44.8

$$\sqrt{1.989}$$

$$\text{var}(\text{scores})_{\text{pooled}} = 1.989$$

If we now divide the weights by the square root of the pooled variance, which corresponds to the pooled standard deviation,

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores
Viral	40.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.9
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.6	45.0
Bacterial	17.3	42.7	44.8

$$LD1 = 0.11 \cdot CRP + 0.70 \cdot Temp$$

we would get the following types of weights, or loadings, that are usually presented by most statistical software tools.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores	Weights
Viral	40.0	36.0	41.6	0.150
Viral	11.1	37.2	38.4	0.989
Viral	30.0	36.5	40.6	
Viral	21.4	39.4	42.2	
Viral	10.7	39.6	40.9	
Viral	3.4	40.7	40.8	
Bacterial	42.0	37.6	43.5	0.11
Bacterial	31.1	42.2	46.4	0.70
Bacterial	50.0	38.5	45.5	
Bacterial	60.4	39.4	48.0	
Bacterial	45.7	38.6	45.0	
Bacterial	17.3	42.7	44.8	

$$LD1 = 0.11 \cdot CRP + 0.70 \cdot Temp$$

By using this discriminant function with these rescaled weights, we can compute the following scores. These scores are the same as the ones in the first example in this video.

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores	Weights
Viral	40.0	36.0	41.6	0.150
Viral	11.1	37.2	38.4	0.989
Viral	30.0	36.5	40.6	
Viral	21.4	39.4	42.2	
Viral	10.7	39.6	40.9	
Viral	3.4	40.7	40.8	
Bacterial	42.0	37.6	43.5	0.11
Bacterial	31.1	42.2	46.4	0.70
Bacterial	50.0	38.5	45.5	
Bacterial	60.4	39.4	48.0	
Bacterial	45.7	38.6	45.0	
Bacterial	17.3	42.7	44.8	

$$LD1 = 0.11 \cdot CRP + 0.70 \cdot Temp$$

If we now calculate the variance of the scores for each group,

Normalize the weights

$$LD1 = 0.150 \cdot CRP + 0.989 \cdot Temp$$

Infection	CRP [mg/l]	Temp [C]	Scores	Weights
Viral	40.0	36.0	41.6	0.150
Viral	11.1	37.2	38.4	0.989
Viral	30.0	36.5	40.6	
Viral	21.4	39.4	42.2	
Viral	10.7	39.6	40.9	
Viral	3.4	40.7	40.8	
Bacterial	42.0	37.6	43.5	0.11
Bacterial	31.1	42.2	46.4	0.70
Bacterial	50.0	38.5	45.5	
Bacterial	60.4	39.4	48.0	
Bacterial	45.7	38.6	45.0	
Bacterial	17.3	42.7	44.8	

$$LD1 = 0.11 \cdot CRP + 0.70 \cdot Temp$$

and pool those, we see that this pooled variance is now equal to one.

Centered scores

Infection	CRP [mg/l]	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.1	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.9	-1.7
Viral	3.4	40.7	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	0.2
Bacterial	31.1	42.2	46.6	32.9	2.3
Bacterial	50.0	38.5	45.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.9	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

Note that most statistical software tools will report the centered discriminant scores,

Centered scores

Infection	CRP [mg/l]	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.1	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.9	-1.7
Viral	3.4	40.7	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	0.2
Bacterial	31.1	42.2	46.6	32.9	2.3
Bacterial	50.0	38.5	45.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.9	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

which can be calculated based on the following equation,

Centered scores

Infection	CRP [mg/l]	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.1	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.9	-1.7
Viral	3.4	40.7	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	0.2
Bacterial	31.1	42.2	46.6	32.9	2.3
Bacterial	50.0	38.5	45.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.9	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

where we first subtract the means from the original values.

Centered scores

Infection	CRP [mg/l]	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.1	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.9	-1.7
Viral	3.4	40.7	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	0.2
Bacterial	31.1	42.2	46.6	32.9	2.3
Bacterial	50.0	38.5	45.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.9	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

For example, to calculate this value,

Centered scores

Infection	CRP (mg/l)	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.3	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.2	-1.7
Viral	3.4	38.4	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	2.3
Bacterial	31.1	42.2	46.6	32.9	1.7
Bacterial	50.0	38.5	49.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.0	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

we plug in the two measurements for this person,

Centered scores

Infection	CRP (mg/l)	Temp (C)	Scores	Scores	Cent. scores
Viral	40.0	36.0	41.6	29.5	-1.1
Viral	11.3	37.2	38.4	27.3	-3.3
Viral	30.0	36.5	40.6	28.8	-1.8
Viral	21.4	39.4	42.2	29.9	-0.7
Viral	10.7	39.6	40.8	30.2	-1.7
Viral	3.4	38.4	40.8	28.9	-1.7
Bacterial	42.0	37.6	43.5	30.8	2.3
Bacterial	31.1	42.2	46.6	32.9	1.7
Bacterial	50.0	38.5	49.5	32.3	1.7
Bacterial	60.4	39.4	48.0	34.0	3.5
Bacterial	45.7	38.6	45.0	31.0	1.3
Bacterial	17.3	42.7	44.8	31.8	1.2

$$LD = 0.11 \cdot (CRP - \bar{CRP}) + 0.70 \cdot (Temp - \bar{Temp})$$

and the means of the two variables. By doing this, the score is calculated to about minus 1.1.

Standardized discriminant function coefficients

Standardized discriminant function coefficients

Finally, we will have a look at the so-called standardized discriminant function coefficients, which are commonly used to analyze how much each variable contributes to the separation.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.3	37.2	-0.7	-0.6
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.3	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

One way to calculate these coefficients is to first compute the LDA on standardized data.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.5	-0.1	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	11.1	37.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)}}$$

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

To standardize a variable, we usually subtract its mean from the original values,

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.5	-0.1	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	11.1	37.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)}}$$

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

and divide by the square root of the variance, which is the standard deviation of that variable.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.5	-0.1	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	11.1	37.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)}}$$

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

In contrast, for LDA we divide by the square root of the pooled variance of the groups for that variable.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.5	-0.1	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	11.1	37.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)}}$$

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

If we use this equation, we will get the following standardized variables.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-3.8	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	32.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

As an example, we will here look at how these standardized values have been calculated.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-3.8	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	32.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$\bar{X} = 30.3$$

We first need to calculate the mean of the CRP variable, which is 30.3.

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-3.8	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	32.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{Viral}} = 188.3$$

The variance of the CRP for the virus group is 188.3,

Standardized discriminant function coefficients

Infection	CRP (mg/l)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-3.8	-0.8
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	32.2	-0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.3	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{Viral}} = 188.3$$

$$\text{var}(X)_{\text{Bacterial}} = 228.0$$

and the variance of the bacteria group is 228.0.

Standardized discriminant function coefficients

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{Viral}} = 188.3$$

$$\text{var}(X)_{\text{pooled}} = 208.1$$

$$\text{var}(X)_{\text{Bacterial}} = 228.0$$

We can now pool these two variances, and since the sample sizes are equal between the groups, the pooled variance is simply the mean of these two variances. We see that the pooled variance is 208.1

Standardized discriminant function coefficients

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$Z = \frac{X - 30.3}{\sqrt{208.1}}$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{Viral}} = 188.3$$

$$\text{var}(X)_{\text{pooled}} = 208.1$$

$$\text{var}(X)_{\text{Bacterial}} = 228.0$$

Let's plug in the mean and the pooled variance.

Standardized discriminant function coefficients

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$Z = \frac{40 - 30.3}{\sqrt{208.1}} = 0.7$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{viral}} = 188.3$$

$$\text{var}(X)_{\text{pooled}} = 208.1$$

$$\text{var}(X)_{\text{Bacterial}} = 228.0$$

As an example, we will here calculate the standardized CRP value for person number one.

Standardized discriminant function coefficients

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

$$Z = \frac{X - \bar{X}}{\sqrt{\text{var}(X)_{\text{pooled}}}}$$

$$Z = \frac{40 - 30.3}{\sqrt{208.1}} = 0.7$$

$$\bar{X} = 30.3$$

$$\text{var}(X)_{\text{viral}} = 188.3$$

$$\text{var}(X)_{\text{pooled}} = 208.1$$

$$\text{var}(X)_{\text{Bacterial}} = 228.0$$

We see that the calculation results in a standardized value of about 0.7.

Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

We now simply use these values to compute the scores of the first discriminant function.

Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

By using a software with these standardized values, we will get the following weights,

Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

$$LD1 = 1.53 \cdot zCRP + 1.41 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

and if we rescale the weights in the same way as we have seen previously, we will get the following standardized discriminant function coefficients, which are commonly reported by most statistical software tools.

Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

$$LD1 = 1.53 \cdot zCRP + 1.41 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

These coefficients are usually used to determine how much each variable contributes to the separation. If a variable is associated with a relatively high weight, we can interpret this as the variable contributes more to the separation.

Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

$$LD1 = 1.53 \cdot zCRP + 1.41 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

In other words, if a variable is associated with a relative high weight, that variable is better to separate the groups compared to the other variables.

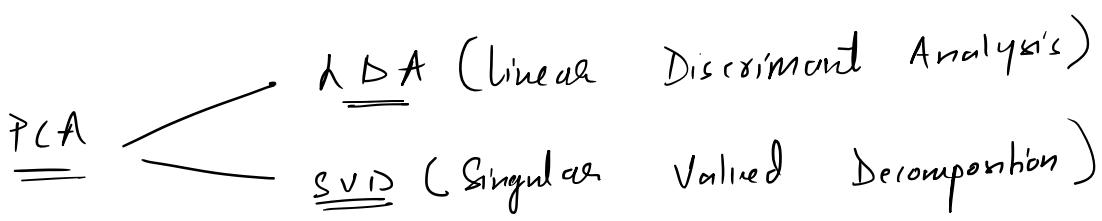
Standardized discriminant function coefficients

$$LD1 = 0.73 \cdot zCRP + 0.68 \cdot zTemp$$

$$LD1 = 1.53 \cdot zCRP + 1.41 \cdot zTemp$$

Infection	CRP (mg/L)	Temp (C)	Z CRP	Z Temp
Viral	40.0	36.0	0.7	-1.5
Viral	11.1	37.2	-1.3	-0.9
Viral	30.0	36.5	0.0	-1.3
Viral	21.4	39.4	-0.6	0.2
Viral	10.7	39.6	-1.4	0.3
Viral	3.4	40.7	-1.9	0.8
Bacterial	42.0	37.6	0.8	-0.7
Bacterial	31.1	42.2	0.1	1.6
Bacterial	50.0	38.5	1.4	-0.3
Bacterial	60.4	39.4	2.1	0.2
Bacterial	45.7	38.6	1.1	-0.2
Bacterial	17.3	42.7	-0.9	1.8

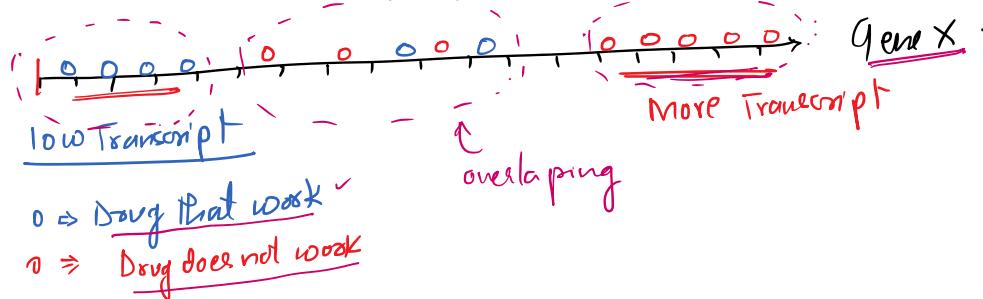
In our example, the values of the two coefficients are quite similar, which means that the variables CRP and body temperature contribute to about the same extent to separate the virus group from the bacteria group.



LDA [Linear Discriminant Analysis].

- * Suppose \Rightarrow We got a cancer drug ✓
 - \hookrightarrow It works great for some people ✓
 - \hookrightarrow But it makes it worse for other people ✓
- * How do we decide whom to give the drug? ?
- * May be gene study of patients will be of some help.

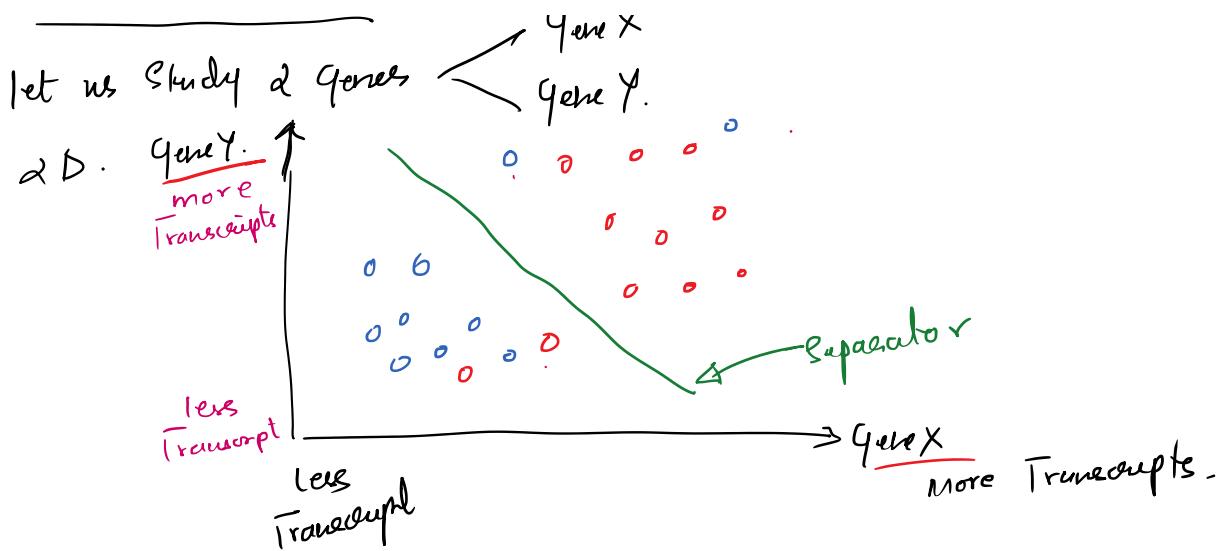
(1) Consider only one gene \Rightarrow Gene X. [I-D]



- * From above we can see In cases with More Transcript drug is not working.
- * Whereas In cases of low Transcript the drug tends to work.
- * Now there are few exceptions (there are overlaps).
- * Summary \Rightarrow Gene X does OK job but it has few overlaps.

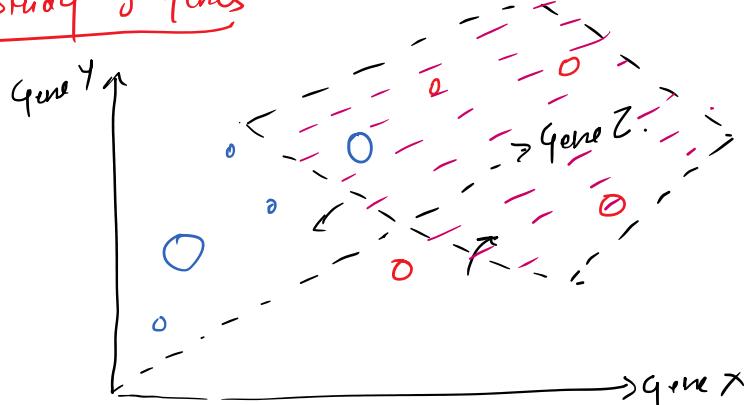
* Can we do better?

Let us Study of Genes $\begin{cases} \text{Gene X} \\ \text{Gene Y} \end{cases}$



It is better than studying only one gene but still there are overlapping.

Let us Study 3 Genes



In case of 3 genes we will need to represent the information in 3-D.

→ Here the separator is a plane.

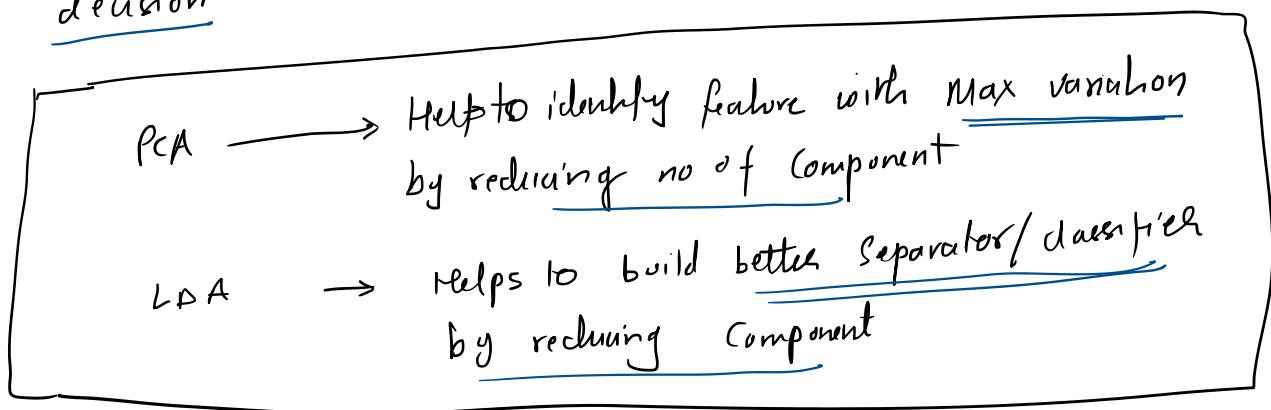
Suppose we want to study 4 genes $\xrightarrow{\text{Need}} \text{4-D Represent} \rightarrow \text{Can't draw 4D graph}$.

* Here more the Dimension, More will be Complication in graph representation.

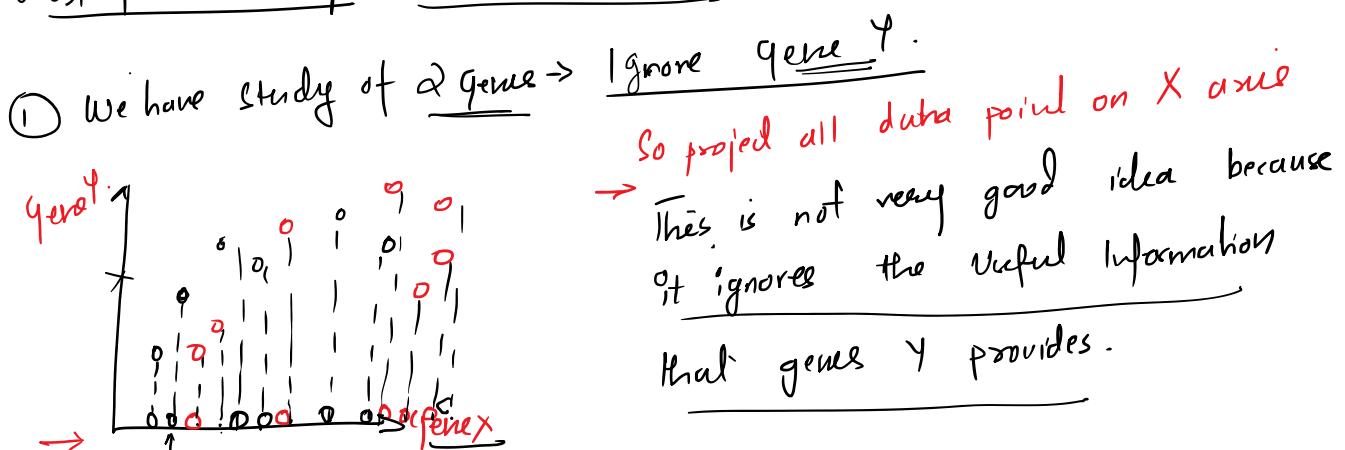
+ PCA reduces dimensions by focusing on the genes with most variance (More the variance, More the information).

- * Normally PCA is useful in plotting data with lot of dimension (or genes) onto simple XY plane \rightarrow
- * Here we are not interested much in identifying genes with most variance
Instead (LDA) we are interested in Maximizing Separability betn the two groups so we can make best classification decision.

Note



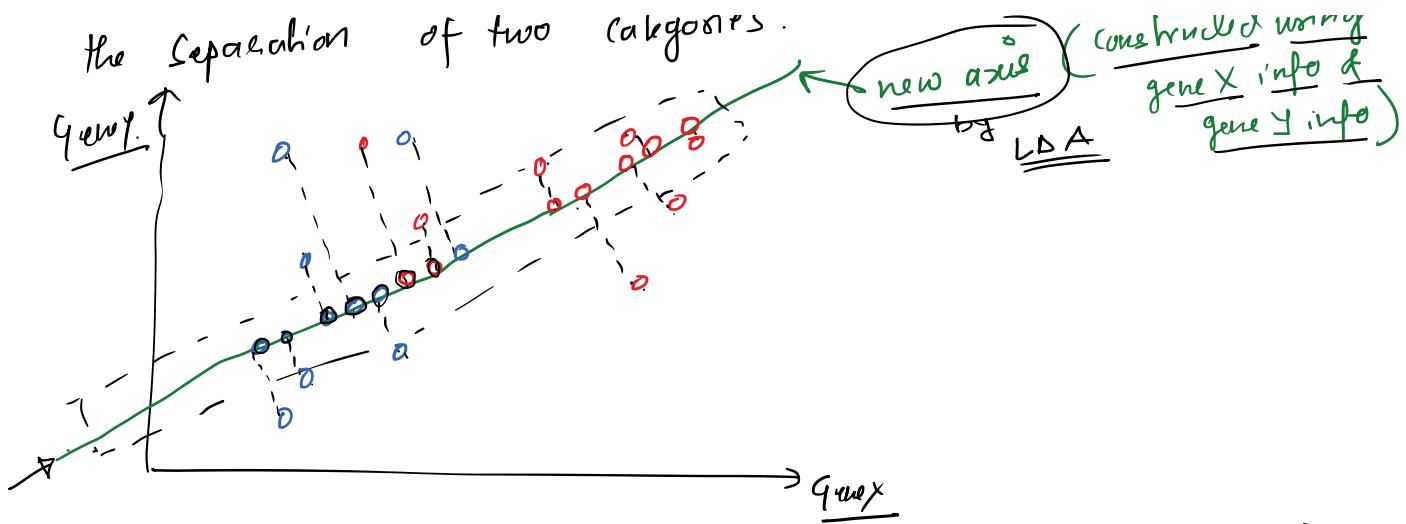
- * Worst possible ways to Reduce Dimension \Rightarrow



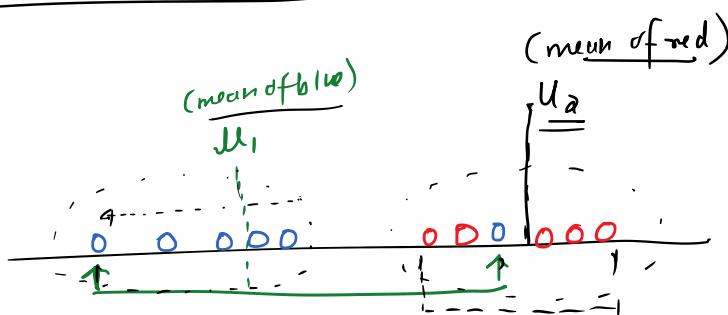
- * LDA provides better way

\rightarrow LDA reduces an $\alpha \beta$ graph to 1-D graph.
 \rightarrow LDA uses both the genes to create a new axis and project the data on new axis in a way to maximize the separation of two categories.





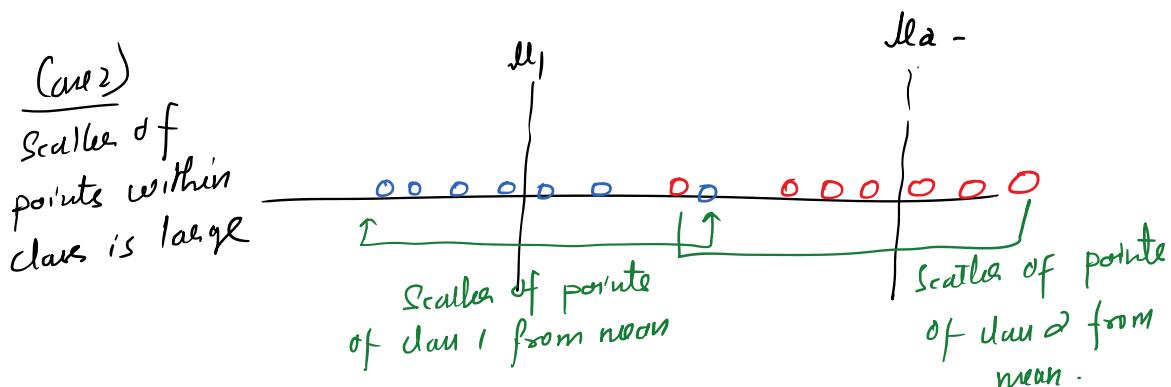
How LDA creates a New Axis's \rightarrow \underline{m}_1 and \underline{m}_2 are mean of respective class.



Case 1) means of two classes are near.
→ the points in each class are widely scattered

The new axis is created according to two criteria
(considered simultaneously)

- ① Maximize the distance between mean of two classes.)
- ② Minimize the Variation (LDA calls it scatter) and is represented by S^2 for each category.

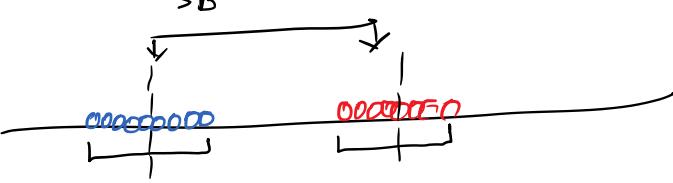


$S_B \rightarrow$ Between class scatter.

⇒ points of points

(con 3) Scatter of points
within class is small.

$S_B \rightarrow$ Between class scatter.



$\underline{S_w}$ Scatter of points is small within class.
 $\underline{S_B}$ Between class scatter.

Goal $\rightarrow \sqrt{S_B} \Rightarrow$ Maximize } Must be focused
 $\sqrt{S_w} \Rightarrow$ Minimize } on both
 simultaneously.

Mathematical Representn \Rightarrow
$$\frac{(m_1 - m_2)^2}{S_{w1}^2 + S_{w2}^2} = \frac{\text{Square of diff bet two mean}}{\text{Sum of scattered within each category/class}}$$

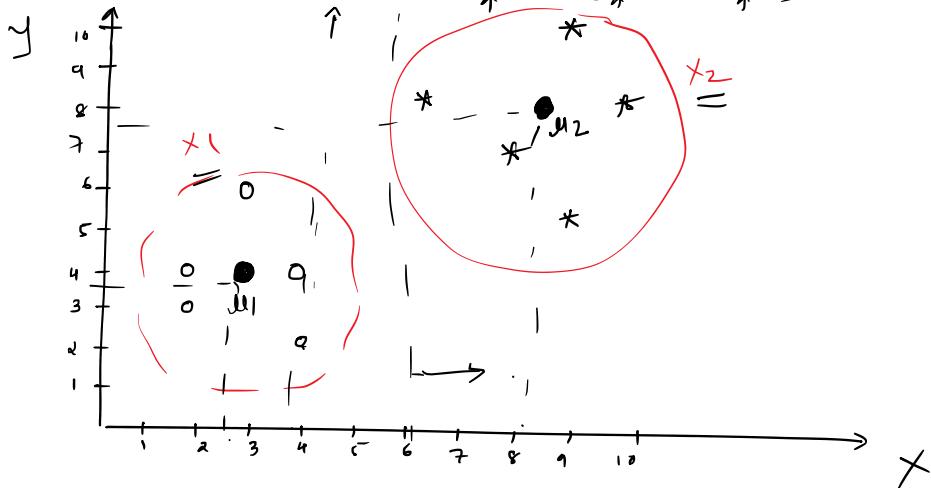
Ideally
$$\frac{(m_1 - m_2)^2}{S_{w1}^2 + S_{w2}^2} \Rightarrow$$
 Large ✓

$$S_{w1}^2 + S_{w2}^2 \Rightarrow$$
 Small ✓

Consider \Rightarrow Compute the (linear Discriminant projection) for the following two Dimensional Dataset -

$$\text{Class 1} \rightarrow \underline{\underline{X}_1} = (x_1, y_1) = \{ (4, 2), (2, 4), (2, 3), (3, 6), (4, 4) \} \Rightarrow \circ$$

$$\text{Class 2} \rightarrow \underline{\underline{X}_2} = (x_2, y_2) = \{ (9, 10), (6, 8), (9, 5), (8, 7), (10, 8) \} \Rightarrow *$$



Step 1 Find (two means)

$$\begin{aligned} \underline{\underline{\mu}}_1 &= \frac{1}{N_1} \sum x \in \underline{\underline{X}}_1 \\ &= \frac{1}{5} \left[\left(\frac{4}{2} \right) + \left(\frac{2}{4} \right) + \left(\frac{2}{3} \right) + \left(\frac{3}{6} \right) + \left(\frac{4}{4} \right) \right] = \frac{1}{5} \begin{pmatrix} 15 \\ 19 \end{pmatrix} \\ &= \underline{\underline{\begin{pmatrix} 3 \\ 3.8 \end{pmatrix}}} \end{aligned}$$

$$\begin{aligned} \underline{\underline{\mu}}_2 &= \frac{1}{N_2} \sum x \in \underline{\underline{X}}_2 \\ &= \frac{1}{5} \left[\left(\frac{9}{10} \right) + \left(\frac{6}{8} \right) + \left(\frac{9}{5} \right) + \left(\frac{8}{7} \right) + \left(\frac{10}{8} \right) \right] = \frac{1}{5} \begin{pmatrix} 42 \\ 38 \end{pmatrix} \\ &= \underline{\underline{\begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}}} \end{aligned}$$

Step 2 Covariance Matrix of Both class.

Covariance Matrix of Class 1. ($\underline{\underline{X}}_1$)

$$\underline{\underline{S}}_1 = \frac{1}{N-1} \sum_{x \in \underline{\underline{X}}_1} (x - \underline{\underline{\mu}}_1)(x - \underline{\underline{\mu}}_1)^T$$

$$\begin{aligned}
 S_1 &= \frac{1}{N-1} \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^T \\
 &= \frac{1}{4} \left[\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 \right. \\
 &\quad \left. + \left[\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 4 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 3.8 \end{pmatrix} \right]^2 \right] \\
 &= \begin{bmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{bmatrix}_{//}
 \end{aligned}$$

Covariance Matrix of Second class

$$\begin{aligned}
 S_2 &= \frac{1}{N-1} \sum_{x \in X_2} (x - \mu_2)(x - \mu_2)^T \\
 &= \frac{1}{4} \left[\left[\begin{pmatrix} 9 \\ 10 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 6 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 9 \\ 5 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^2 \right. \\
 &\quad \left. + \left[\begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 10 \\ 8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^2 \right] \\
 &= \begin{bmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{bmatrix}
 \end{aligned}$$

Step 3 Within Class Scatter Matrix (Minimize)

$$\begin{aligned}
 S_W &= S_1 + S_2 \\
 &= \begin{bmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{bmatrix} + \begin{bmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{bmatrix} \\
 &= \begin{bmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{bmatrix}_{//}
 \end{aligned}$$

Step 4 Between Class Scatter Matrix (Maximize).

$$\begin{aligned}
 S_B &= (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \\
 &= \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right] \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^T
 \end{aligned}$$

$$= \begin{bmatrix} -5.4 \\ -3.8 \end{bmatrix} \begin{bmatrix} -5.4 & -3.8 \end{bmatrix}$$

$$S_B = \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix}$$

Step 5 The LDA projection is then obtained as solution of the generalized Eigen Value Problem

$$S_w^{-1} S_B w = \lambda w$$

$$\Rightarrow |S_w^{-1} S_B - \lambda I| = 0$$

$$\Rightarrow \left| \left(\begin{array}{cc} 3.3 & -0.3 \\ -0.3 & 5.5 \end{array} \right)^{-1} \left(\begin{array}{cc} 29.16 & 20.52 \\ 20.52 & 14.44 \end{array} \right) - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\Rightarrow \left| \left(\begin{array}{cc} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{array} \right) \left(\begin{array}{cc} 29.16 & 20.52 \\ 20.52 & 14.44 \end{array} \right) - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$= \left| \begin{pmatrix} 9.2213 - \lambda & 6.489 \\ 4.2339 & 2.9794 - \lambda \end{pmatrix} \right| \quad \text{--- } I$$

$$\Rightarrow (9.2213 - \lambda)(2.9794 - \lambda) - 6.489 \times 4.2339 = 0$$

$$\lambda^2 - 12.2007\lambda = 0 \Rightarrow \lambda(\lambda - 12.2007) = 0$$

$$\lambda \Rightarrow \boxed{\begin{array}{l} \lambda_1 = 0 \\ \lambda_2 = 12.2007 \end{array}} \quad \text{2 Eigen Values.}$$

Now By Substituting $\lambda = \lambda_1$ in I we get

$$\boxed{\text{Eigen Vector } I = w_1 = \begin{pmatrix} -0.5755 \\ 0.8178 \end{pmatrix}}$$

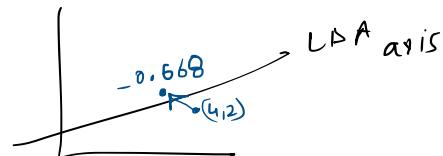
for Eigen

Eigen vector $\underline{w}_1 = \underline{\underline{w}}_1 = \begin{pmatrix} 0.8178 \\ 0.9088 \end{pmatrix}$

By Substituting $\lambda = \lambda_1$ in I we get

Eigen Vector $\underline{w}_2 = \underline{\underline{w}}_2 = \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix}$

Step 6 Final $y = \underline{w}^T \underline{x}$
 Data \uparrow Input Data
 Projection Vector



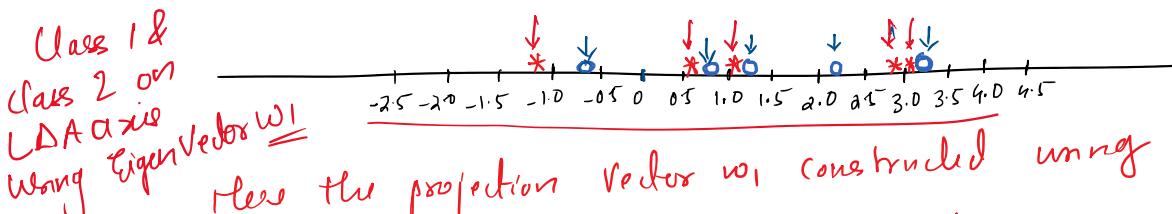
For Projection Vector \underline{w}_1

Final Data of $\underline{x}_1 = \underline{w}_1^T \underline{x}_1 = \begin{bmatrix} -0.5755 & 0.8178 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 3 \\ 6 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 4 \\ 1 \end{bmatrix}$

Blue $= \begin{bmatrix} -0.668 \\ 0.334 \end{bmatrix} = \begin{bmatrix} 1.3022 \\ 0.9688 \end{bmatrix}$

Final Data of $\underline{x}_2 = \underline{w}_1^T \underline{x}_2 = \begin{bmatrix} -0.5755 & 0.8178 \end{bmatrix} \begin{bmatrix} 9 \\ 6 \end{bmatrix} = \begin{bmatrix} 3.0894 \\ 0.7874 \end{bmatrix}$

Red $= \begin{bmatrix} 3.0894 \\ 0.7874 \end{bmatrix}$



Here the projection vector \underline{w}_1 constructed wrong.
Smaller Eigen value leads to bad separability.
 $(\lambda_1=0)$

For Projection Vector \underline{w}_2

$$\text{final Data } X_1 = w_2^T X_1$$

$$= \begin{bmatrix} 0.9088 & 0.4173 \end{bmatrix} \begin{bmatrix} (1) \\ (2) \\ (3) \\ (4) \\ (5) \end{bmatrix}$$

Class 1 on LDA using w_2 vector

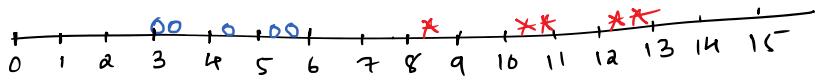
$$= \begin{bmatrix} 4.4698 & 3.4868 & 3.0645 & 5.2302 & 5.3044 \end{bmatrix}$$

$$\text{Final Data } X_2 = w_2^T X_2$$

$$= \begin{bmatrix} 0.9088 & 0.4173 \end{bmatrix} \begin{bmatrix} (9) \\ (8) \\ (5) \\ (7) \\ (6) \end{bmatrix}$$

Class 2 on LDA using w_2 vector

$$\rightarrow = \begin{bmatrix} 12.3522 & 8.7912 & 10.2657 & 10.1951 & 12.4264 \end{bmatrix}$$



Here the projection vector corresponding to larger eigen value λ_2 leads to good separability.

This is LDA \rightarrow when the 2D data points reduced to 1-D data points

\rightarrow and also using projection vector were able to predict a good separator of classes.

SVD [Singular Value Decomposition]

- * We normally use 2D matrix to represent Data Values where column represents features and row represents samples data points
- * Matrix computation with all the values in matrix sometime become redundant or computationally expensive.
- * We need to represent matrix in a form such that the most important part of matrix which is needed for further computation could be extracted easily.
- * This can be done by SVD

SVD Theorem \Rightarrow A rectangular matrix A_{mn} can be decomposed it into product of 3 matrices.

$$A_{mn} = U \sum_{m \times m}^{m \times m} \Sigma_{m \times n} V^T_{n \times n}$$

$$\frac{m \times m}{\boxed{m \times n}} \quad \frac{m \times n}{\boxed{m \times n}} \quad \frac{n \times n}{\boxed{m \times n}}$$

where $U_{m \times m} \Rightarrow$ Orthogonal matrix

$\Sigma_{m \times n} \Rightarrow$ Diagonal matrix of Eigen Values

$V^T_{n \times n} \Rightarrow$ Transpose of orthogonal matrix $\underline{\underline{V}_{n \times n}}$

Columns of U are the orthonormal Eigen Vectors of $A A^T$

& columns of V are the orthonormal Eigen Vectors of $A A^T$ & Σ is diagonal matrix.

The elements of Σ are the square roots of Eigen values of

U & V in decreasing order.

Ex Find the SVD of $\underline{A} = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}_{2 \times 3}$

Note we have to Decompose the matrix A into $\underline{U} \underline{\Sigma} \underline{V^T}$.

To find \underline{U}
 $\underline{\Sigma}$
 $\underline{V^T}$.

Solution \Rightarrow To find $\underline{U} := \underline{A} \underline{A}^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$

for Eigen Values $\Rightarrow \begin{vmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{vmatrix} = 0 \quad \text{--- (I)}$

$$\Rightarrow \lambda^2 - 22\lambda + 120 = 0$$

$$\boxed{\lambda_1 = 10} \quad \boxed{\lambda_2 = 12}$$

For Eigen Vector $\Rightarrow \begin{bmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

(one) \Rightarrow Substitute $\lambda = \lambda_1 = 10$

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\frac{x_1}{1} = -\frac{x_2}{1} = 1$$

$$\therefore x_1 = 1 \quad x_2 = -1$$

Eigen Vector $x_1 = \underline{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}$

(contd) Substitute $\lambda = \lambda_2 = 10$

$$\begin{bmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Eigen Vector $X_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

* In U the Eigen Vector generated by larger Eigen value will be the first column.

$$\therefore U = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

↑ ↑
 Eigen vector Eigen vector
 of $\lambda=12$ of $\lambda=10$

Now we need to Normalize the matrix \Rightarrow divide by length of respective Vector.

*
$$U = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

↑ ↑
 length of length of Vector
 Vector $\lambda=12$ $\lambda=10$

$$\sqrt{1^2+1^2} = \sqrt{2}$$

$$\sqrt{1^2+(-1)^2} = \frac{1}{\sqrt{2}}$$

Slop 2 To find V

$$A^T A = \begin{bmatrix} 3 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

These are Eigen Values

$$\lambda^3 - 22\lambda^2 + 132\lambda = 0$$

$$\begin{array}{l} \therefore \lambda_1 = 0 \\ \lambda_2 = 10 \\ \lambda_3 = 12 \end{array}$$

We need to find Eigen Vectors for the 3 Eigen values.

(case 1) Eigen Vector for $\lambda_1=0$

$$\Rightarrow \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

By Cramers Rule

$$\frac{x_1}{|10 \ 0 \ 2|} = \frac{x_2}{|0 \ 10 \ 4|} = \frac{x_3}{|10 \ 0 \ 10|}$$

$$= \frac{x_1}{-20} = -\frac{x_2}{40} = \frac{x_3}{10} = \frac{-1}{20}$$

$$x_1 = 1 \quad x_2 = 2 \quad x_3 = -5$$

$$\text{Eigen Vector } X_1 = \begin{bmatrix} 1 \\ 2 \\ -5 \end{bmatrix}$$

(case 2) Eigen Vector for $\lambda = 10$

$$\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 4 \\ 2 & 4 & -8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

By Cramers Rule

$$\frac{x_1}{|0 \ 0 \ 2|} = -\frac{x_2}{|0 \ 0 \ 4|} = \frac{x_3}{|2 \ 4 \ -8|} \Rightarrow \frac{x_1}{-16} = -\frac{x_2}{-8} = \frac{x_3}{0} = \frac{-1}{8}$$

$$\therefore x_1 = 2 \quad x_2 = -1 \quad x_3 = 0$$

$$\text{Eigen Vector } x_2 = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$$

Case 3) Eigen Vector for $\lambda_3 = 12$

$$\begin{bmatrix} -2 & 0 & 2 \\ 0 & -2 & 4 \\ 2 & 4 & -10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

By Cramer's Rule

$$\frac{x_1}{\begin{vmatrix} 0 & 2 \\ -2 & 4 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} -2 & 2 \\ 0 & 4 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} -2 & 0 \\ 0 & -2 \end{vmatrix}} \Rightarrow \frac{x_1}{4} = \frac{-x_2}{-8} = \frac{x_3}{4} = \frac{1}{4}$$

$$\therefore x_1 = 1 \quad x_2 = -2 \quad x_3 = 1$$

$$\text{Eigen Vector } x_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

$$\text{Now } V = \begin{bmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & 5 \end{bmatrix}$$

↑ ↑ ↑
 Eigen Vector for Eigen Vector for Eigen Vector
 $\lambda = 12$ $\lambda = 10$ for $\lambda = 0$.

Now Normalizing the matrix \rightarrow Divide by length of Vector.

$$V = \begin{bmatrix} \sqrt{56} & 2/\sqrt{56} & 1/\sqrt{30} \\ 2/\sqrt{56} & -1/\sqrt{5} & 2/\sqrt{30} \\ 1/\sqrt{56} & 0 & -5/\sqrt{30} \end{bmatrix}$$

↑ ↑ ↑
 length length length
 of Vector of Vector of Vector

$$= \begin{bmatrix} \sqrt{11}/\sqrt{56} & 2/\sqrt{56} & 1/\sqrt{30} \end{bmatrix}$$

$$V^T = \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ 2/\sqrt{5} & -1/\sqrt{5} & 0 \\ 1/\sqrt{30} & 2/\sqrt{30} & -5/\sqrt{30} \end{bmatrix}$$

Step 3 To find Σ (or D)

$$\Sigma = \begin{vmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & \sqrt{6} \end{vmatrix}_{3 \times 3}$$

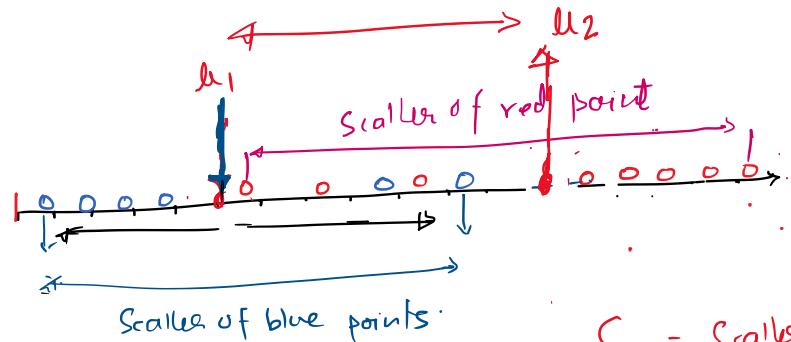
\uparrow
diag matrix \Rightarrow the diag elements are square root of eigen values in decreasing order.

$$A = U \Sigma V^T$$

$$A = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & \sqrt{6} \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ 2/\sqrt{5} & -1/\sqrt{5} & 0 \\ 1/\sqrt{30} & 2/\sqrt{30} & -5/\sqrt{30} \end{bmatrix}$$

$$=$$

- Consider
2 groups
* Group of Blue points
* Group of Red points



u_1 = Mean of blue point

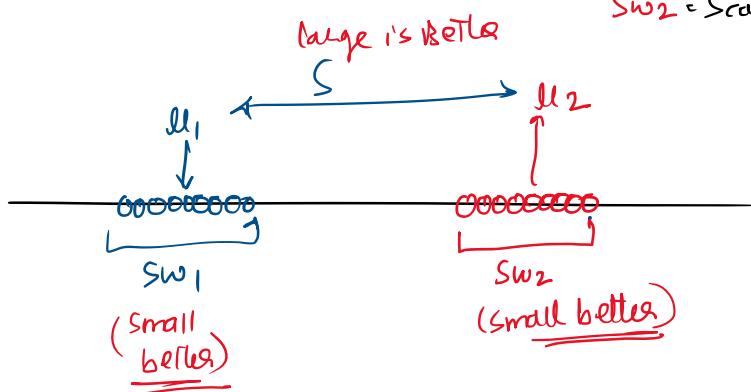
u_2 = mean of red point

S_B = Scatter between two groups
= Should be large

S_W → Scatter within a group.
= Should be small

S_{W1} = Scatter within group 1
(Blue points)

S_{W2} = Scatter within group 2
(Red points)



Note

In math Any Square Matrix A can be represented as
(Decomposed)

3 component matrix

$$A = U D U^T$$

↓ ↓ ↓
 Eigen Vector Transpose of Eigen Vector Diagonal Matrix of Eigen Value

This makes majority of operations on matrix to be easy as operation on Diagonal matrix it's easy.

as operation on Diagonal matrix is easy .

In Real life \rightarrow Database \Rightarrow Matrix \Rightarrow Rectangular Matrix