Linear Discriminant Analysis

Dr. Jyotismita Chaki

Introduction to LDA

- Linear Discriminant Analysis as its name suggests is a linear model for classification and dimensionality reduction.
- Most commonly used for feature extraction in pattern classification problems.
- First, in 1936 Fisher formulated linear discriminant for two classes, and later on, in 1948 C.R Rao generalized it for multiple classes. LDA projects data from a D dimensional feature space down to a D’ (D>D’) dimensional space in a way to maximize the variability between the classes and reducing the variability within the classes.

Why LDA?

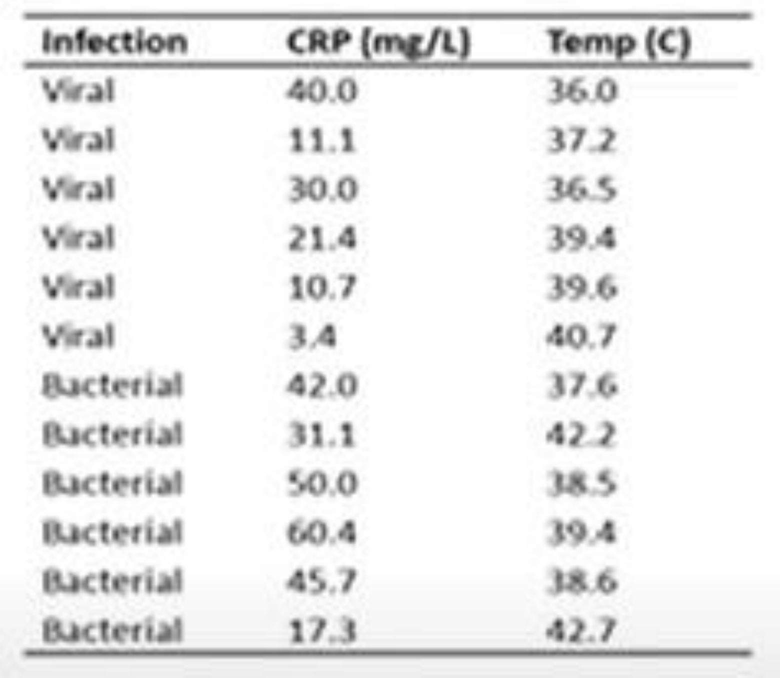
- LDA perform well for data classification.
- LDA can also be used in data preprocessing to reduce the number of features just as PCA which reduces the computing cost significantly.
- LDA is also used in different detection algorithms.
- In Fisherfaces LDA is used to extract useful data from different faces. Coupled with eigenfaces it produces effective results.

Limitations

- Linear decision boundaries may not effectively separate non-linearly separable classes. More flexible boundaries are desired.
- In cases where the number of observations exceeds the number of features, LDA might not perform as desired. This is called *Small Sample Size* (SSS) problem. Regularization is required.

LDA with example

- Separation between two groups: To explain separation let us have a look at the following fictive data set which contains information on 12 patients

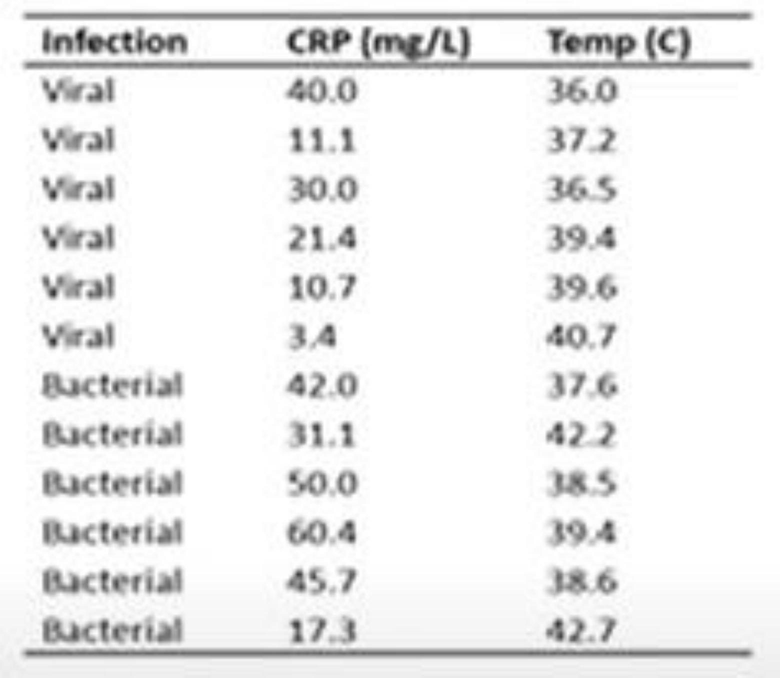


Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

- CRP column shows the concentration of the C reactive protein in blood from the time when the patients entered the hospital.
- Whereas this column shows the body temperature or the same patients at the same time point
- Once the patient 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 bacterial infection.
- After two days at the hospital, first six patients were found to have a viral infection and the last Six patients were confirmed to have bacterial infection

LDA with example

- Separation between two groups: To explain separation let us have a look at the following fictive data set which contains information on 12 patients

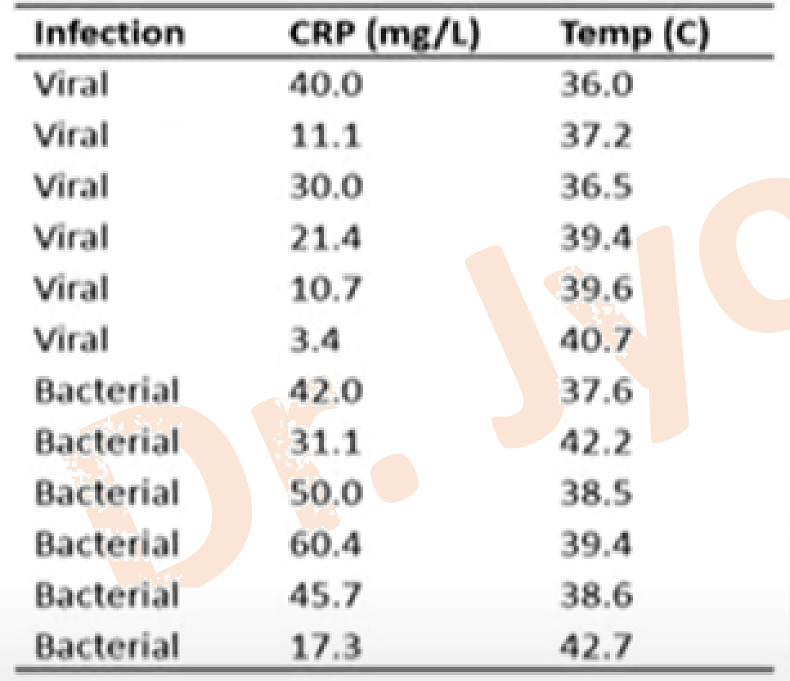


Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

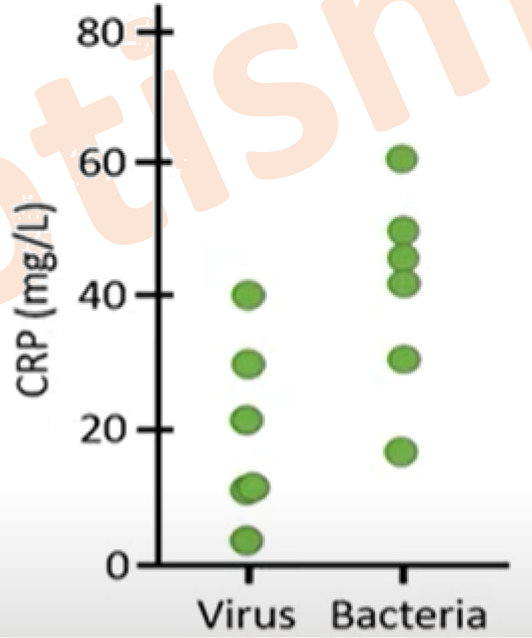
- Since antibiotics are only effective on bacteria, only these patients were treated with antibiotics
- The problem is that we have to wait two days to know if antibiotic treatment is appropriate or not.
- It would be nice if you could use the CRP concentration or the body temperature to tell if patient is a bacterial or viral infection, because the measurement of these variables can be done within just an hour

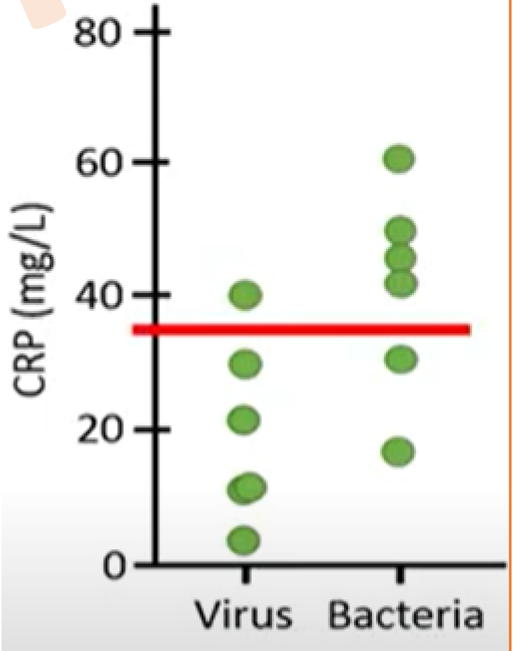
LDA with example

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



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

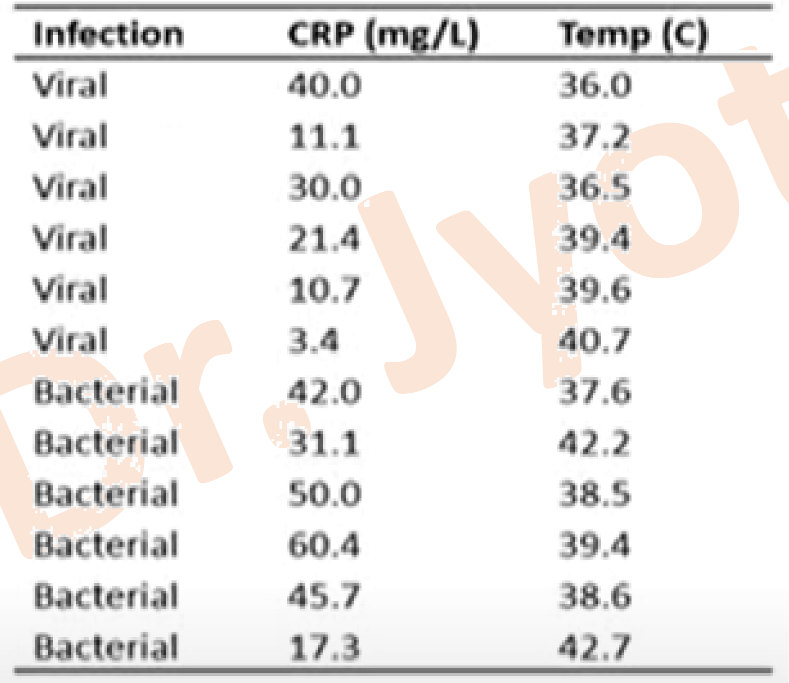




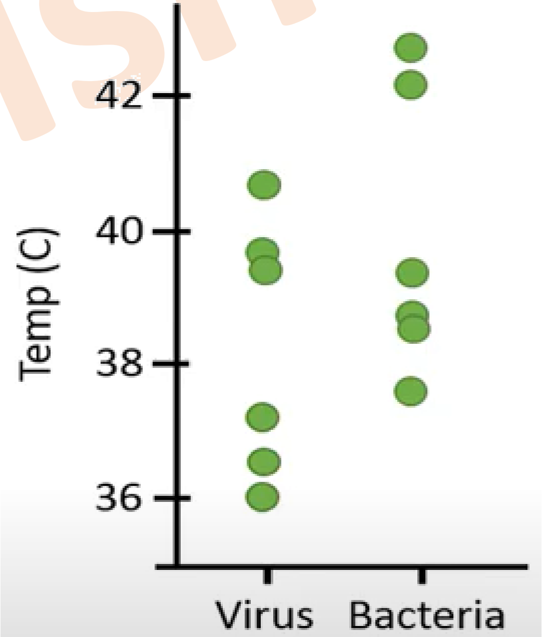
- Let's say that we use a cut-off value of 35 to determine if someone has a bacterial or viral infection.
- However, the problem with these cutoff line is that it cannot separate the groups very well.
- B: 4 and V: 1 patients are above this line
- B: 2 and V: 5 patients are below this line
- Using this line or any other horizontal line, it is impossible to separate the two groups completely

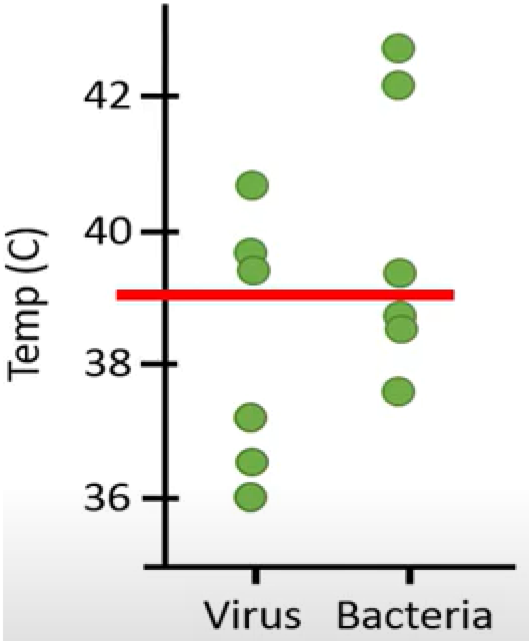
LDA with example

- Lets check, if it is possible to use the body temperature to tell if a patient has a viral bacterial infection.
- As you see, we have the same problem again, but it's no line that can separate the two groups of patients.



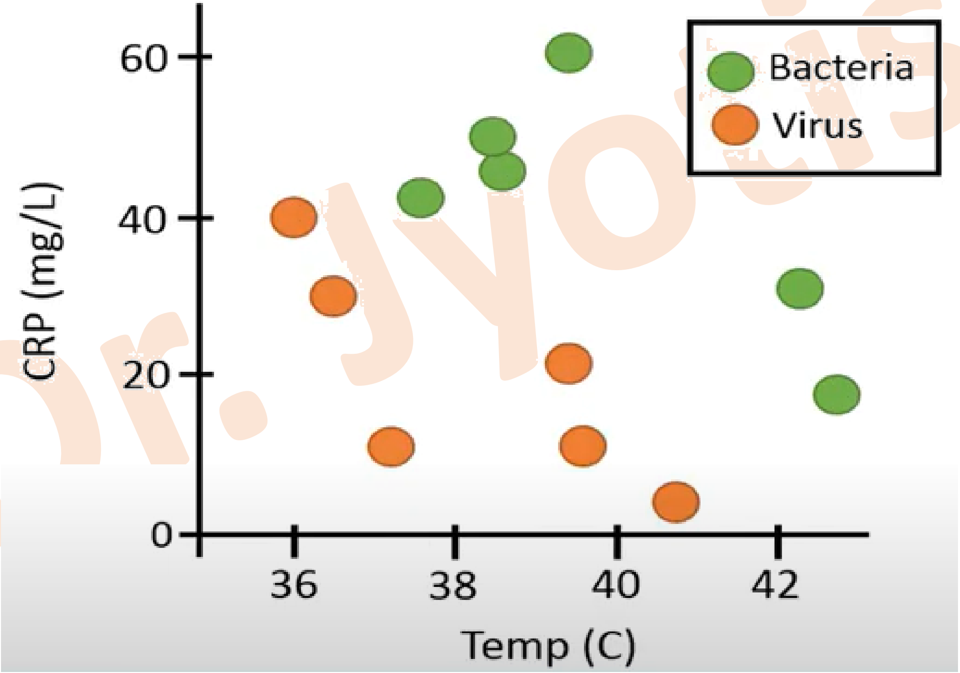
Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

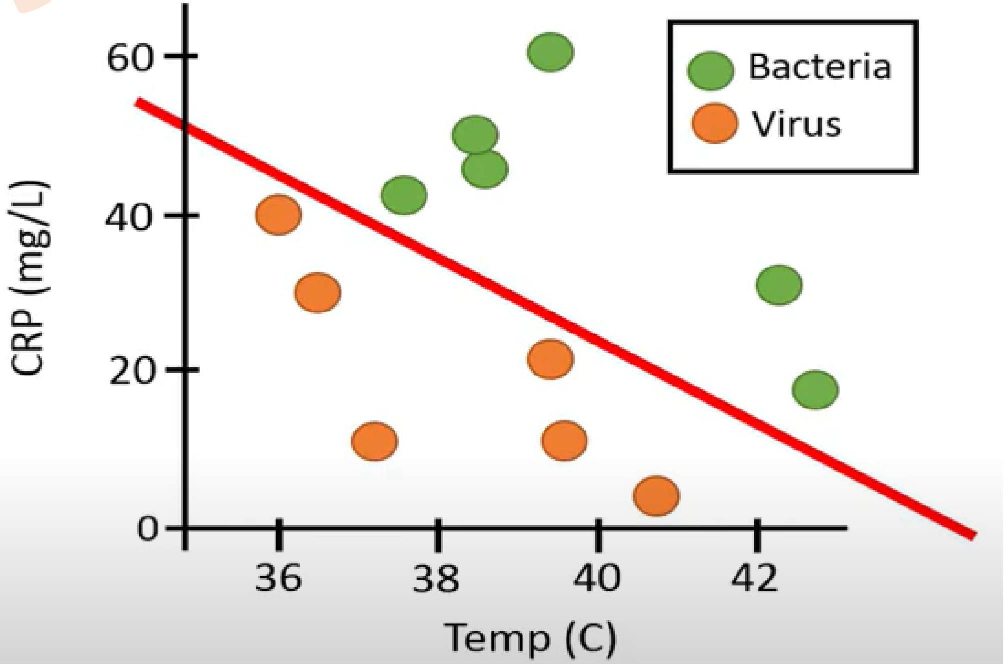




LDA with example

- However, if you 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, we can see that the a line can separate the two groups completely

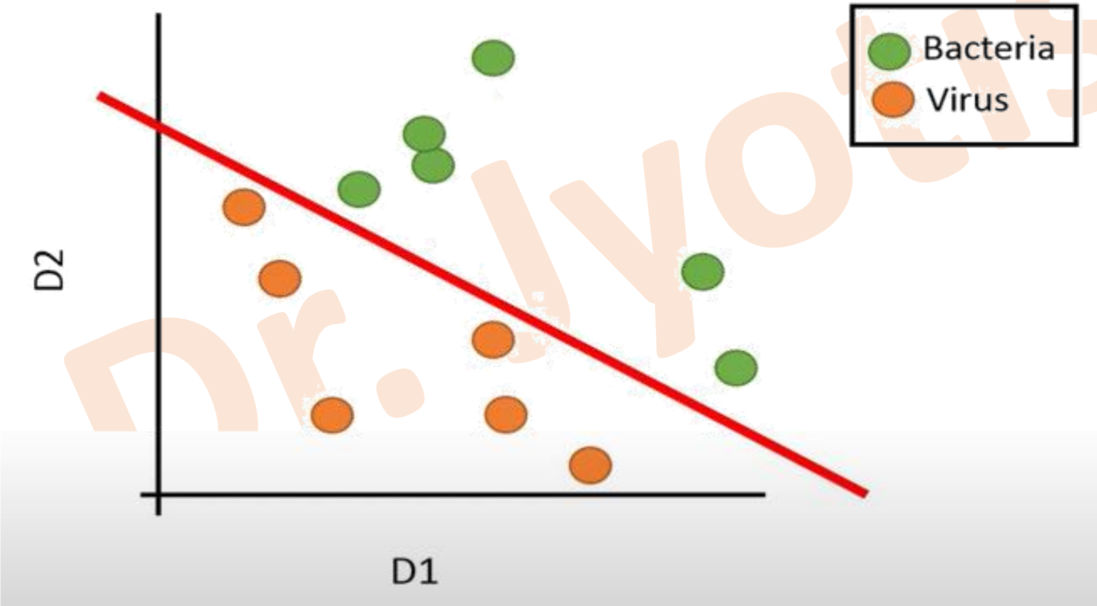


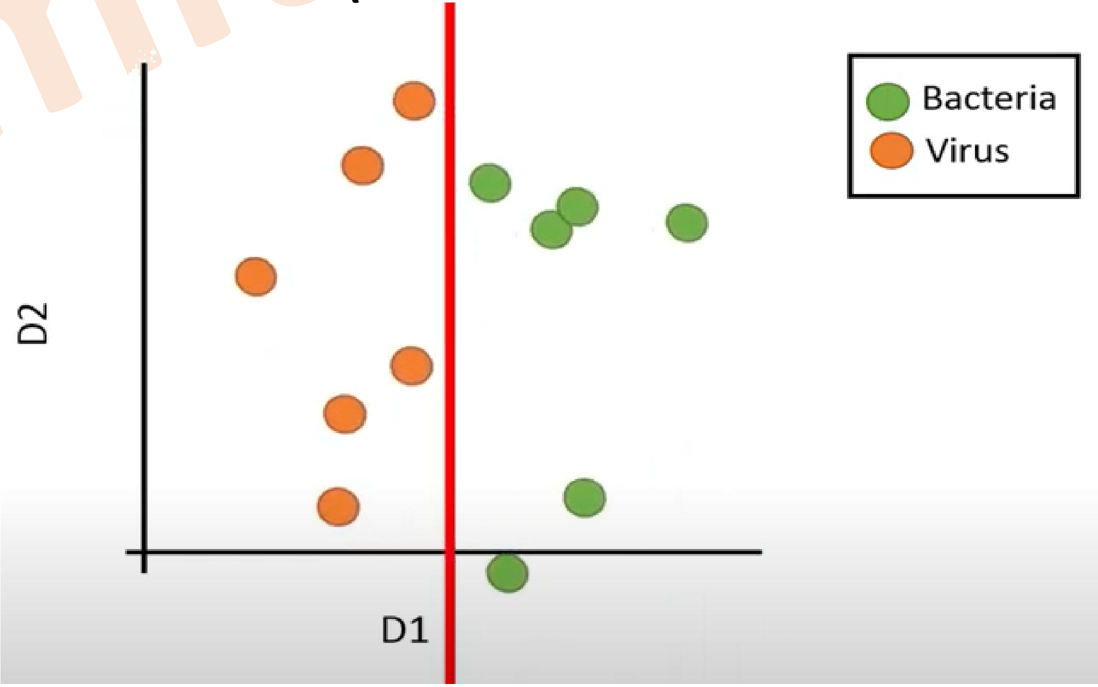


LDA with example

- By combining the two variables, we can get the better separation between the two groups compared with the use one of the variables alone.

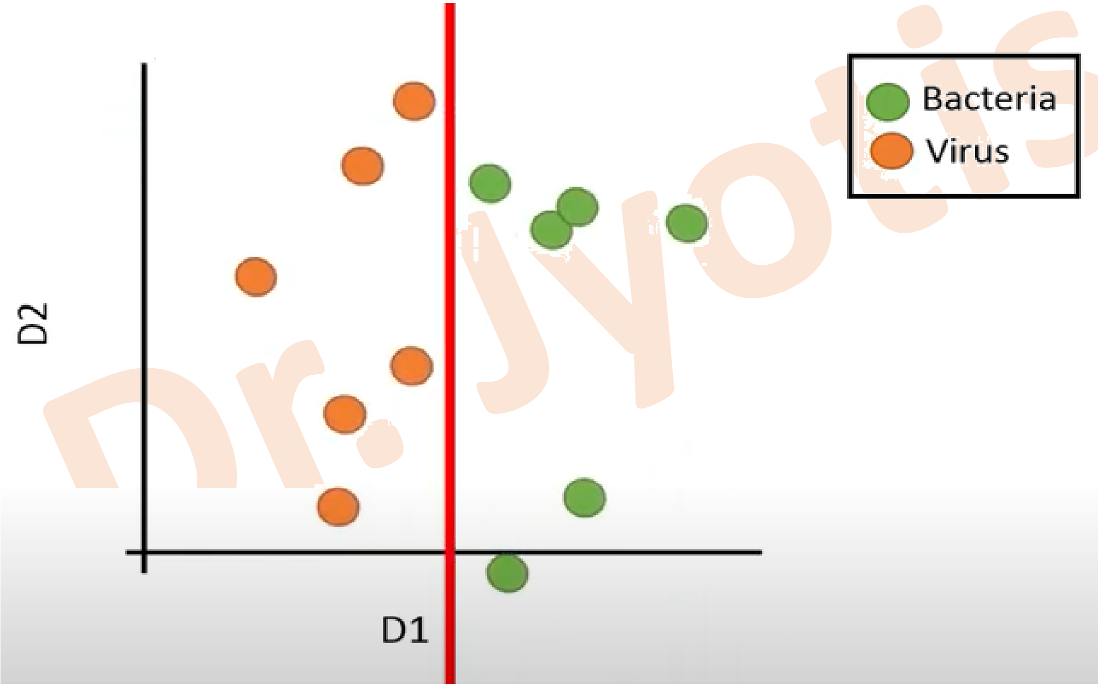
LDA can be seen if we rotate the data into two new dimensions (Discriminant function one and two) (Right figure).

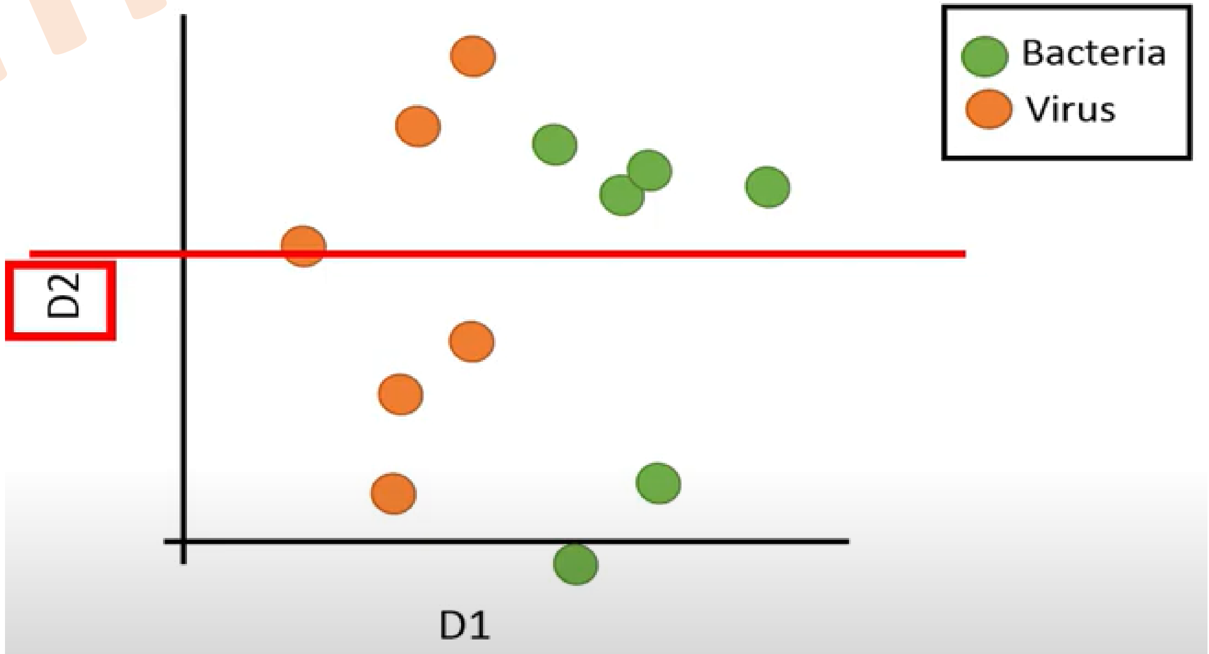




LDA with example

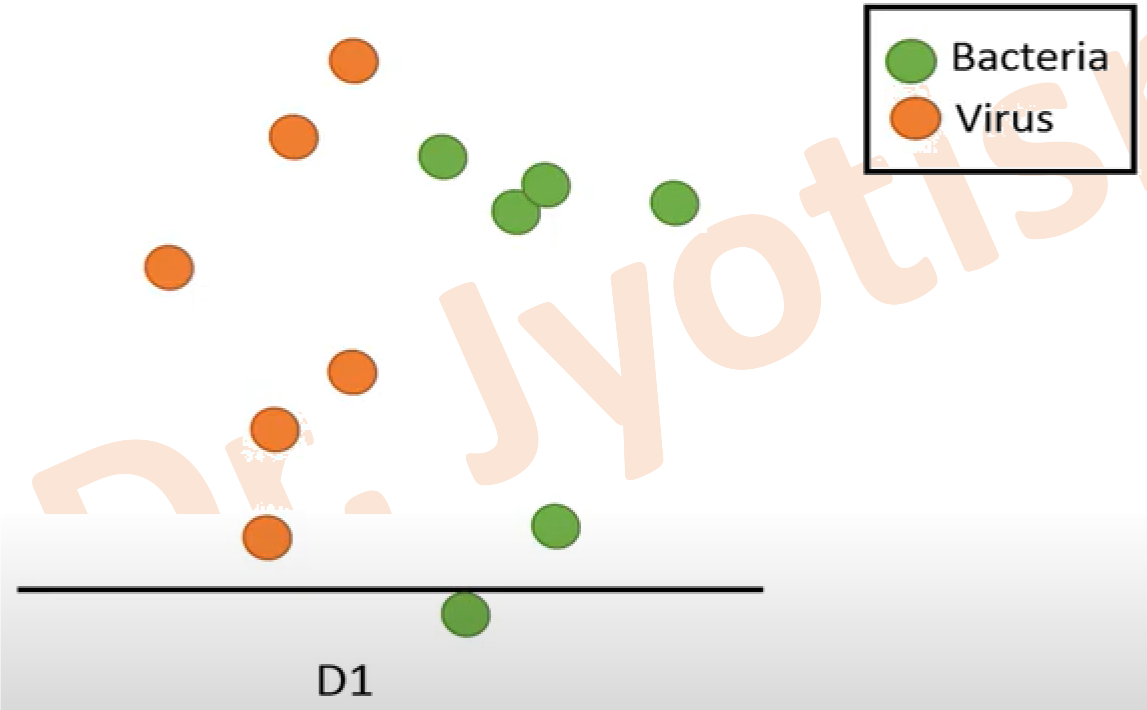
- We can see that the first discriminant function can be used to separate the two groups perfectly.
- In contrast, we cannot separate the groups based on the horizontal; which means that the second discriminant function line is not useful for separation

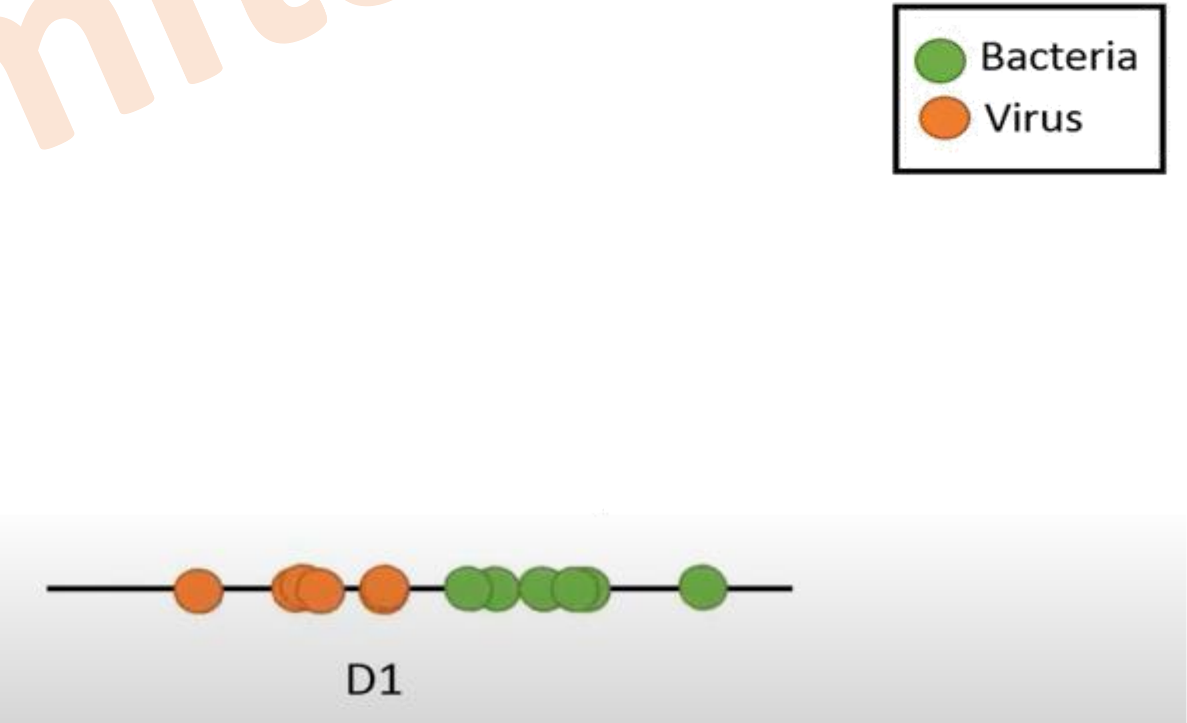




LDA with example

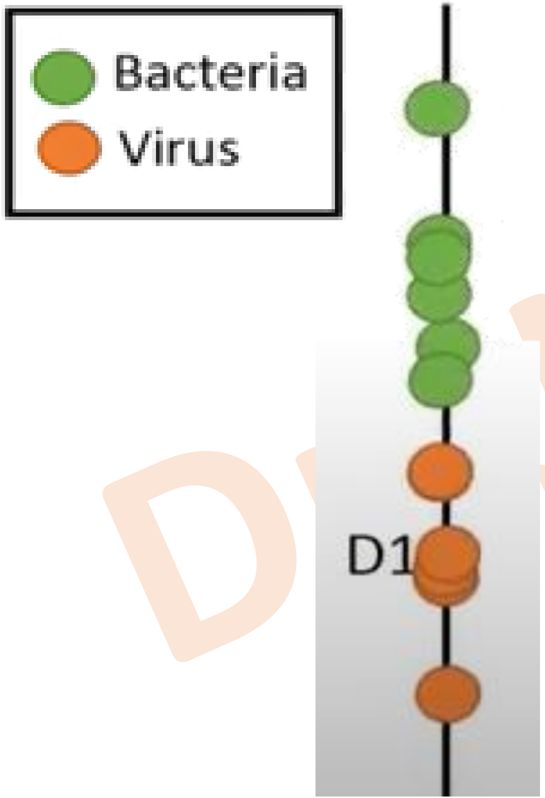
- We can therefore delete the second discriminant function and place all data points on just one line (right figure)



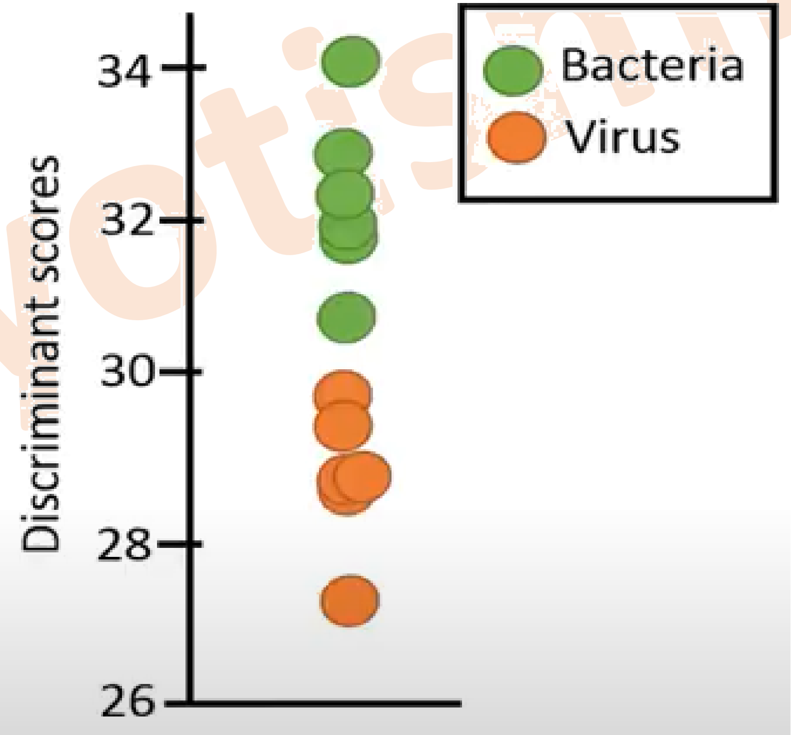


LDA with example

For illustrative purposes, we irritate the plot so that the data points are plotted vertically

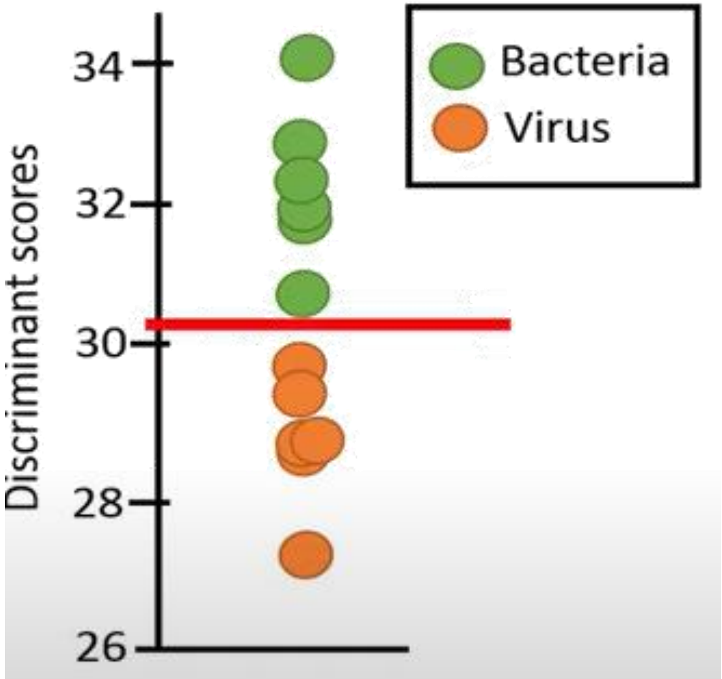


If you use LDA, we can plot the so called discriminant scores



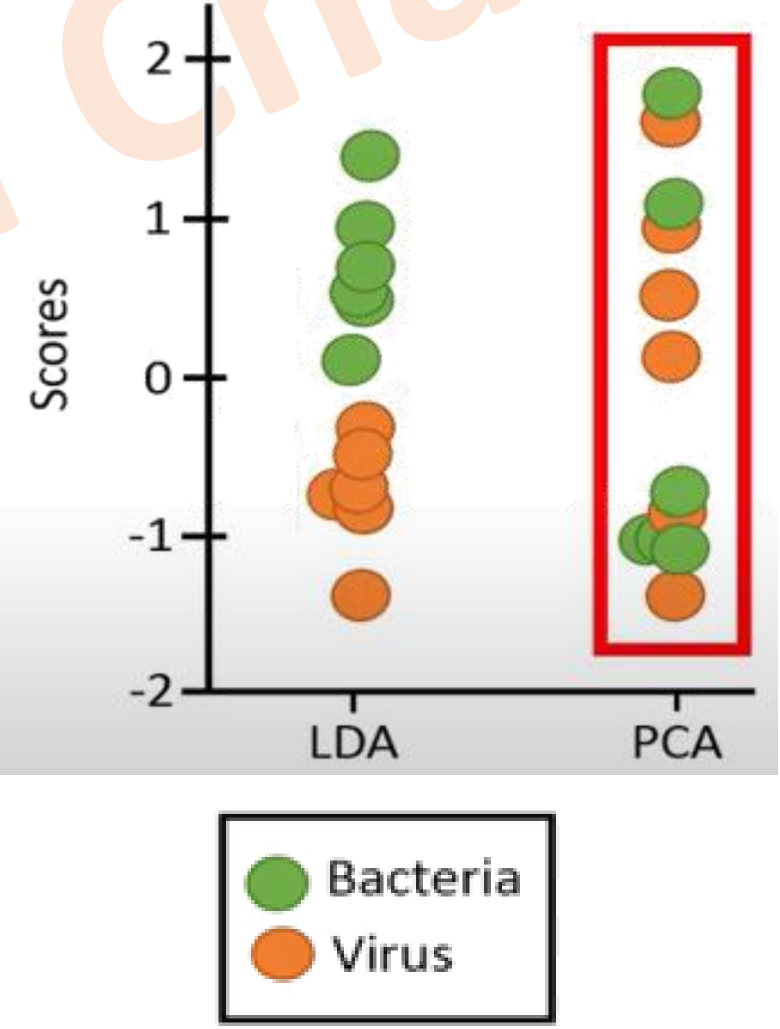
We can now place a line that gives us a perfect separation between the two groups.

By using LDA we have combined the variable CRP and body temperature in a way that has maximized the separation between the two groups

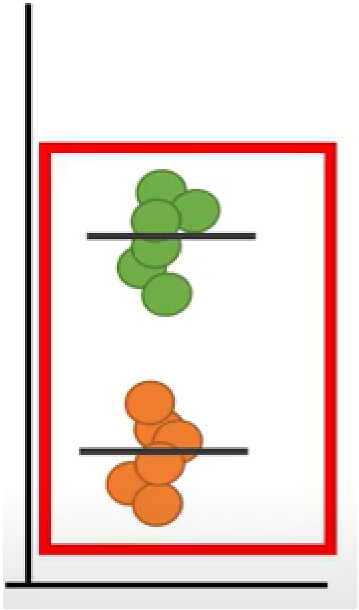


LDA with example

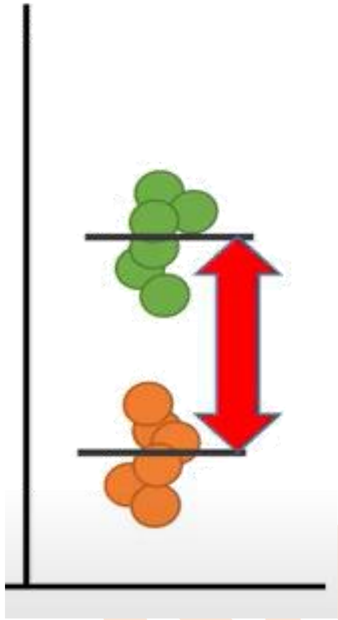
- Difference between PCA and LDA:
  - PCA aims to maximize the variance of the first principal component
  - LDA maximize the separation of the two groups
  - To calculate LDA separation variables or groups (e.g., virus / bacteria) are needed, but for PCA it is not mandatory,
  - PCA sometimes referred to as unsupervised method as it don’t have the information of groups. LDA is referred to as a supervised method because testing information about which group the data bone belongs to.
- We can see that the LDA scores for the bacteria group are completely separated from the scores for the virus group.
- 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.
- PCA is not very effective while separating the groups.



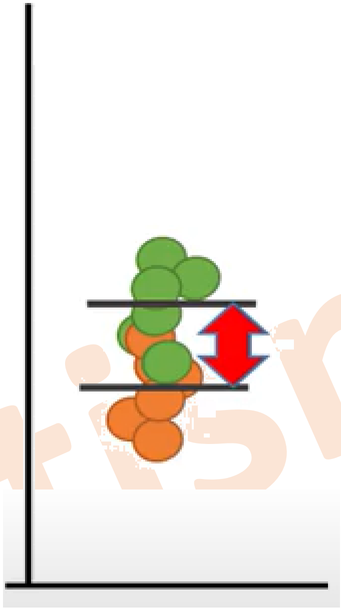
LDA with example



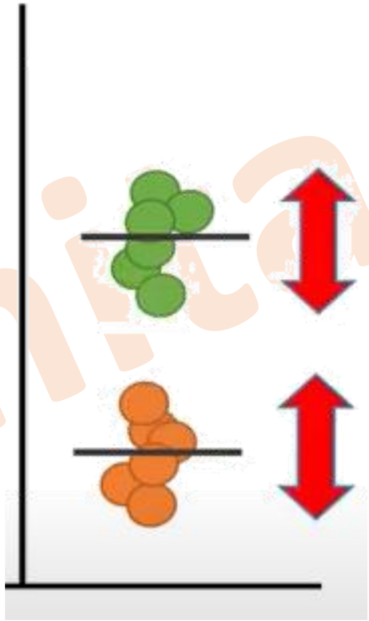
For this data we can see that the two groups separate very well.



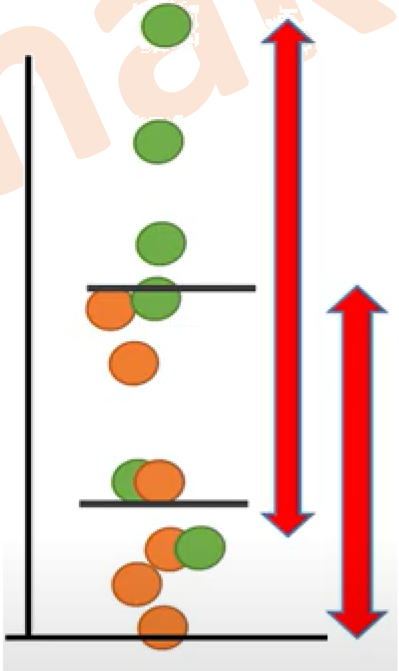
Means of the two groups are far away from each other



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

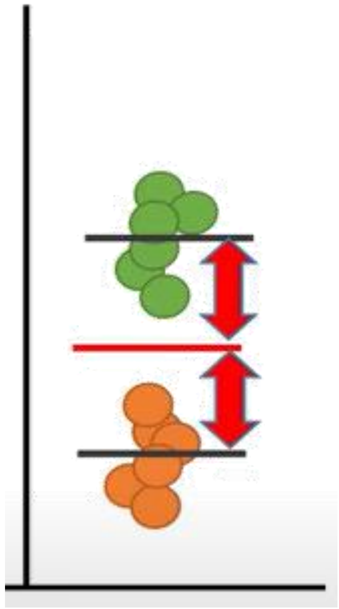


The second reason why these two groups show a good separation is because the spread within the groups is small. The data points are close to the mean values

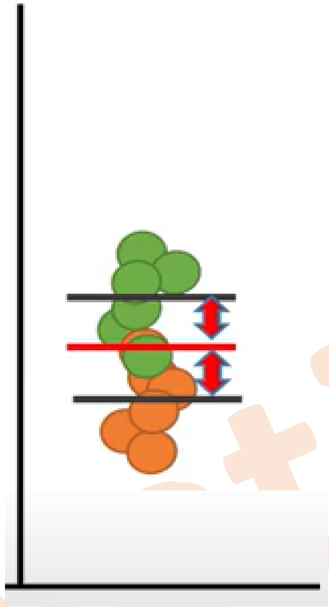


If there is a much larger spread of the observations around the means, then there is no longer clear separation between the groups, even though the difference in their means stay the same

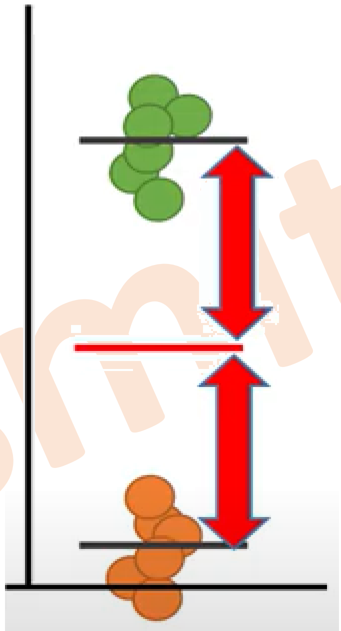
LDA with example



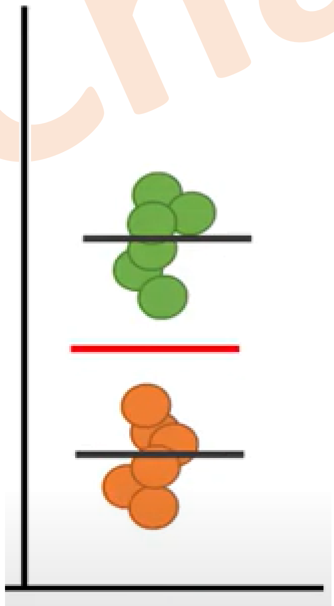
If we calculate the Grand mean, a mean based on all data points, we can think of these distances how much varies from the grand mean



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



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



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 **within group variance**

LDA with example

$$\text{Separation}=\frac{\text{Between-group variance}}{\text{Within-group variance}}$$

- To get a good separation between the groups, the two group means should be far away from each other, which means that it **between group variance should be large**.
- The variation within the group should be smaller, which could result in a **low value of the within group variance.**
- That's the ratio of these measures should be as large as possible to get a clear separation between the groups**.**
- LDA combines variables based on this ratio with the aim to transform the data so that between group variance is increased and within group variance is reduced.

LDA with example

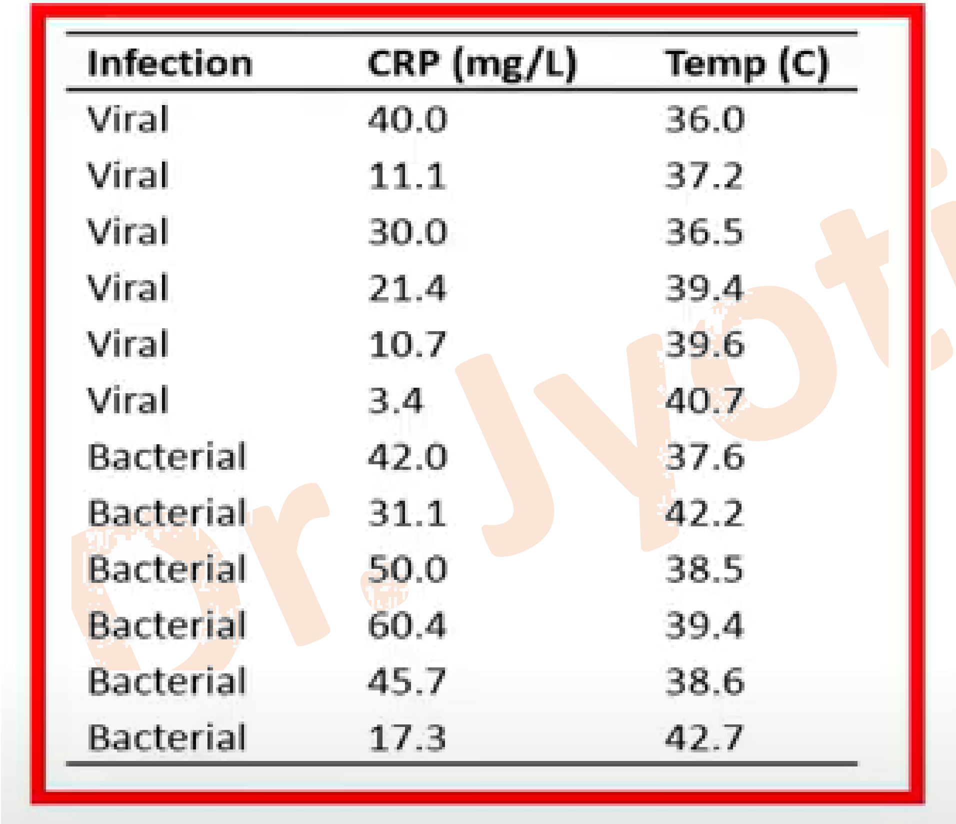
- The separation can be described like this:

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

- B is the between group covariance matrix and W is the pooled within group covariance matrix.
- We compute the eigenvectors of this matrix in order to get our weights.
- To calculate matrix S, the first need to calculate the pooled within group covariance matrix and the between group covariance matrix.

LDA with example

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



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 calculate the pooled within group covariance matrix, we first calculate the covariance matrix with the virus group and then for the bacteria group. The values in these two matrices have been rounded.

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

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

LDA with example

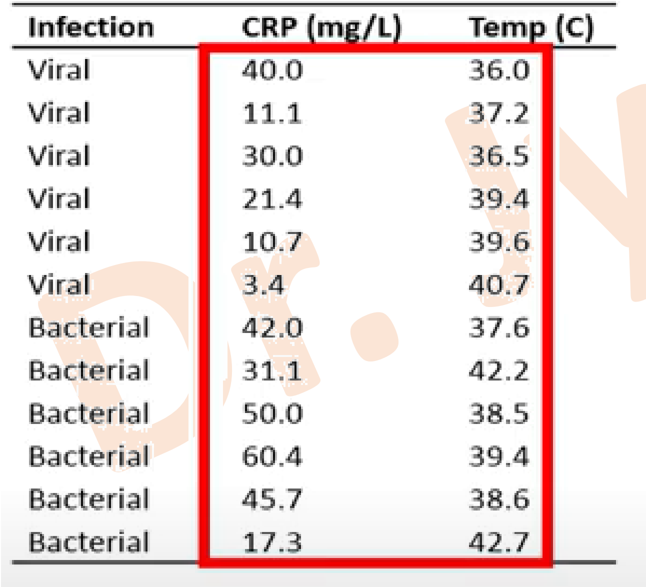
- Since the sample size are equal between the two groups, the pool within group covariance matrix is simply the mean of these two covariance matrices

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

$$W = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix} \qquad W^{-1} = \begin{bmatrix} 0.012 & 0.066 \\ 0.066 & 0.609 \end{bmatrix}$$

LDA with example

- We now try to calculate the between group covariance matrix.
- A simple way to calculate between group covariance matrix is the first calculate the total covariance because between group covariance is equal to the total covariance minus the pooled within group covariance.



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

$$B = T - W$$

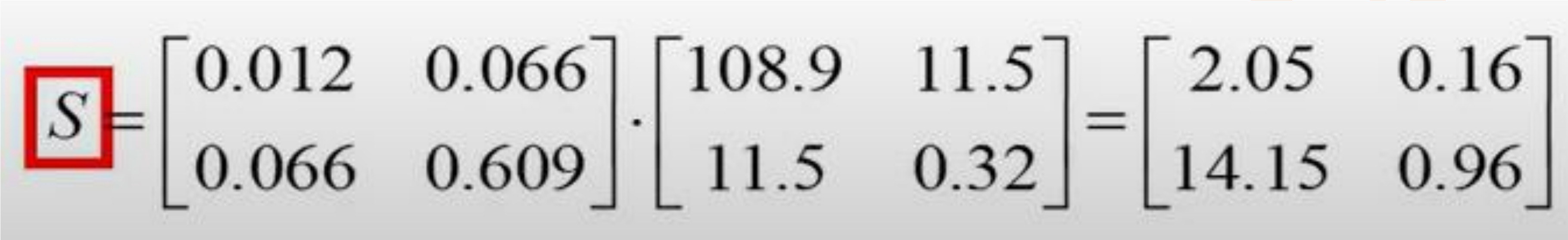
$$T = \begin{bmatrix} 317.1 & -11.0 \\ -11.0 & 4.4 \end{bmatrix}$$

The total covariance matrix shows the variance and the covariance of the two variables when we did not separate the groups. This the covariance matrix is calculated based on all data points.

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

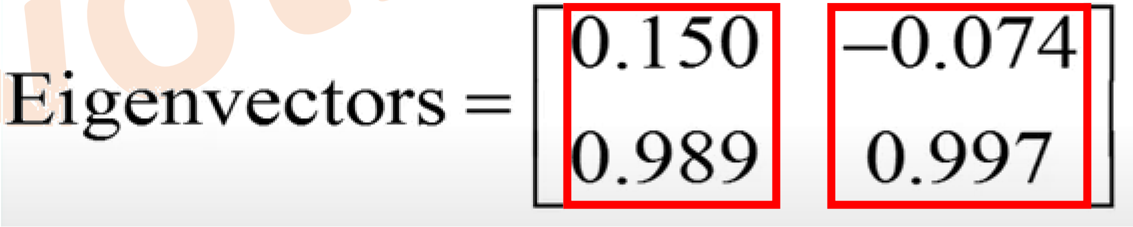
LDA with example

- We calculate matrix S



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

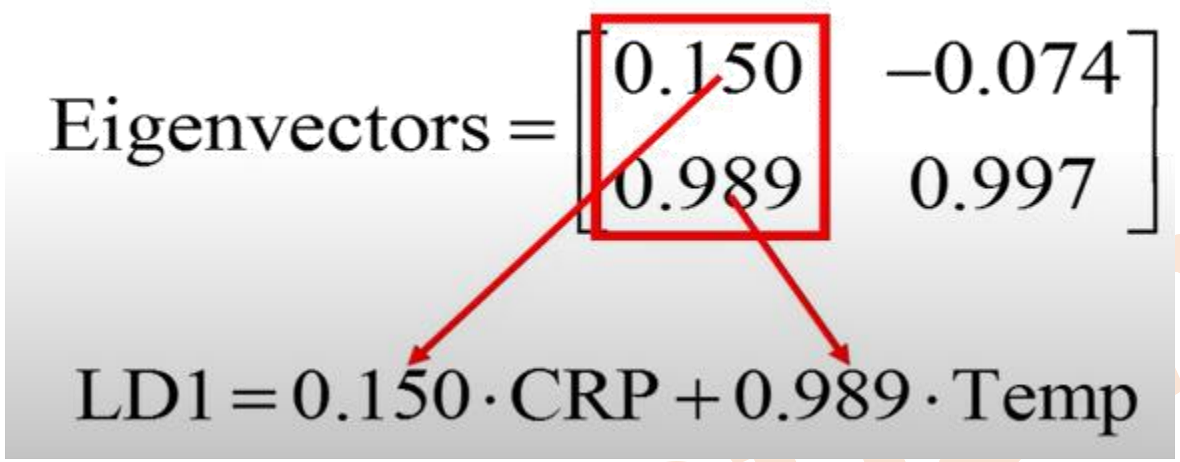
- Eigenvectors of S:



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

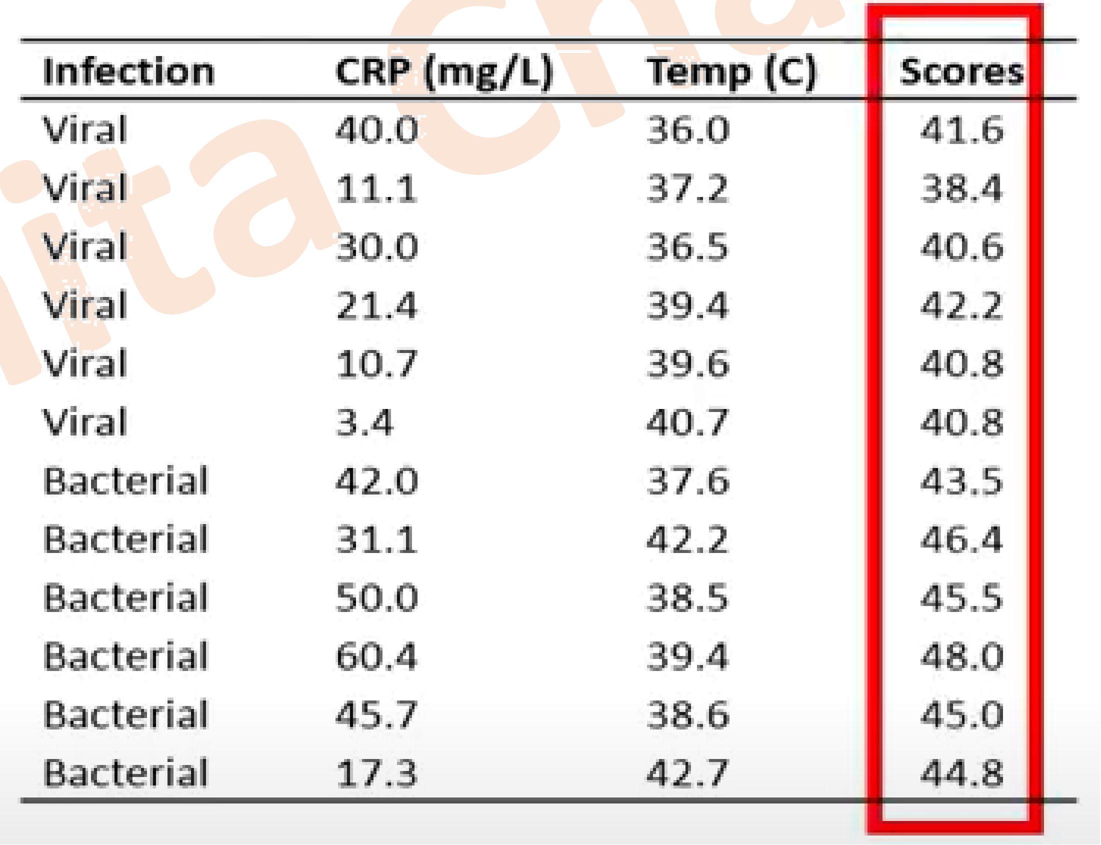
1st EV 2nd EV

LDA with example



$$\text{Eigenvectors} = \left[\begin{array}{cc} \boxed{\begin{matrix} 0.150 \\ 0.989 \end{matrix}} & \begin{matrix} -0.074 \\ 0.997 \end{matrix} \end{array}\right]$$

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

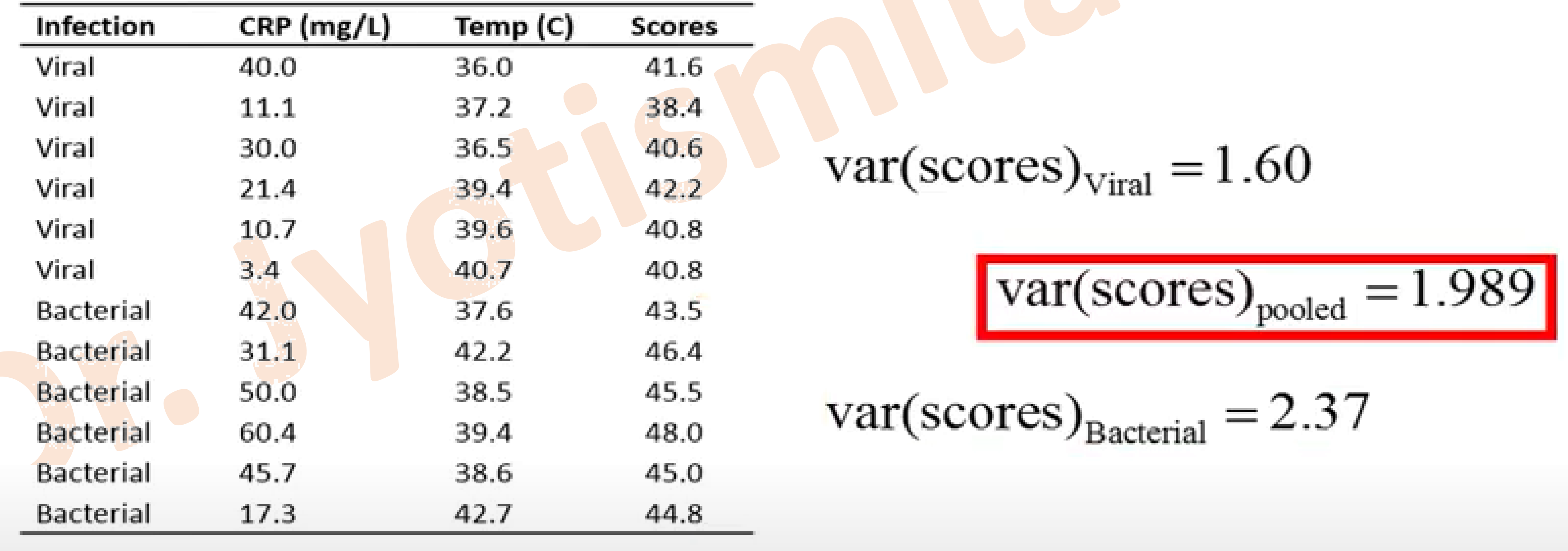


Infection	CRP (mg/L)	Temp (C)	Scores
Viral	40.0	36.0	41.6
Viral	11.1	37.2	38.4
Viral	30.0	36.5	40.6
Viral	21.4	39.4	42.2
Viral	10.7	39.6	40.8
Viral	3.4	40.7	40.8
Bacterial	42.0	37.6	43.5
Bacterial	31.1	42.2	46.4
Bacterial	50.0	38.5	45.5
Bacterial	60.4	39.4	48.0
Bacterial	45.7	38.6	45.0
Bacterial	17.3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

unstandardized scores

LDA with example

- In LDA, the weights are usually rescaled so that the pooled group variance of the scores is equal to 1



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	40.8
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	38.6	45.0
Bacterial	17.3	42.7	44.8

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

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

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

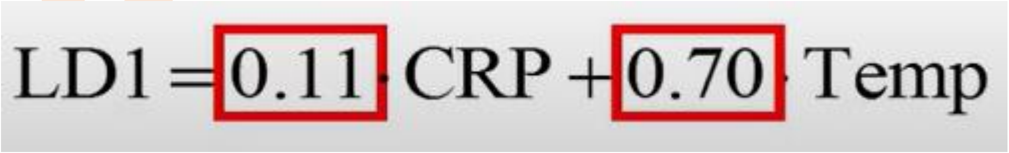
LDA with example

- If we now to divide the weights by the square root of the pooled variance, which corresponds to pull standard deviation, we get the following types of weights or loadings that are usually presented by most statistical software tools

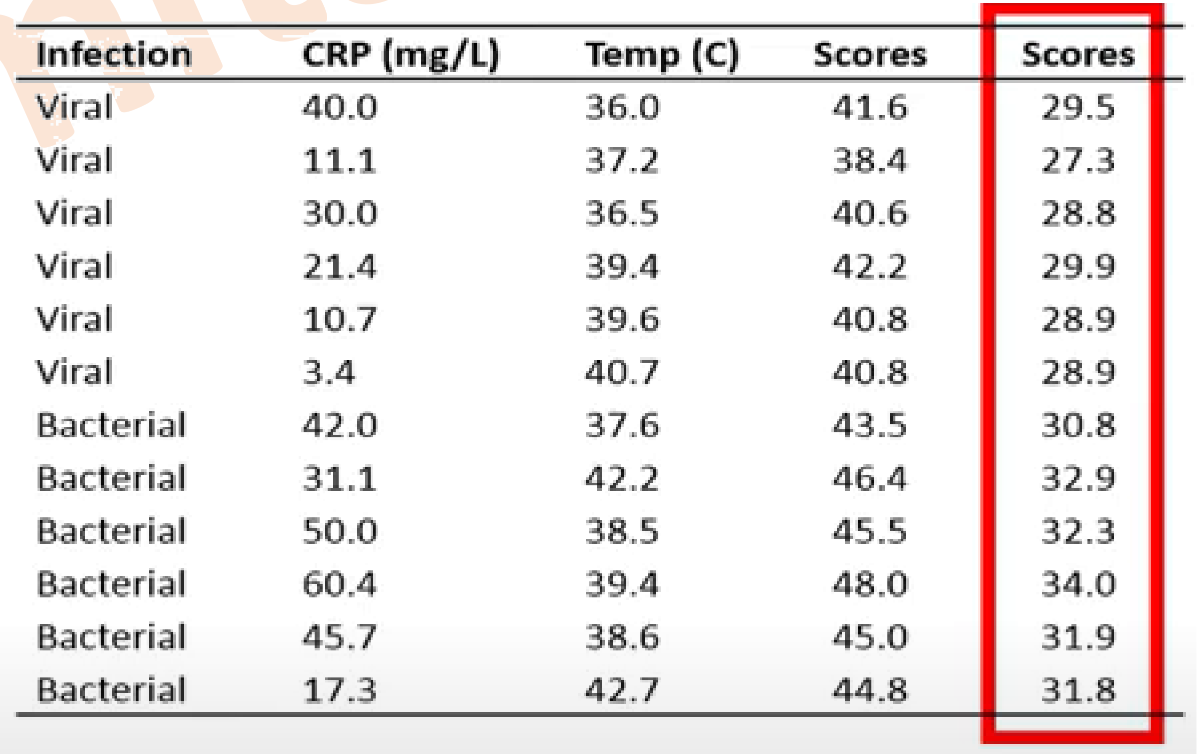
$$\mathrm{LD1} = 0.150 \cdot \mathrm{CRP} + 0.989 \cdot \mathrm{Temp}$$

$$\sqrt{1.989}$$

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

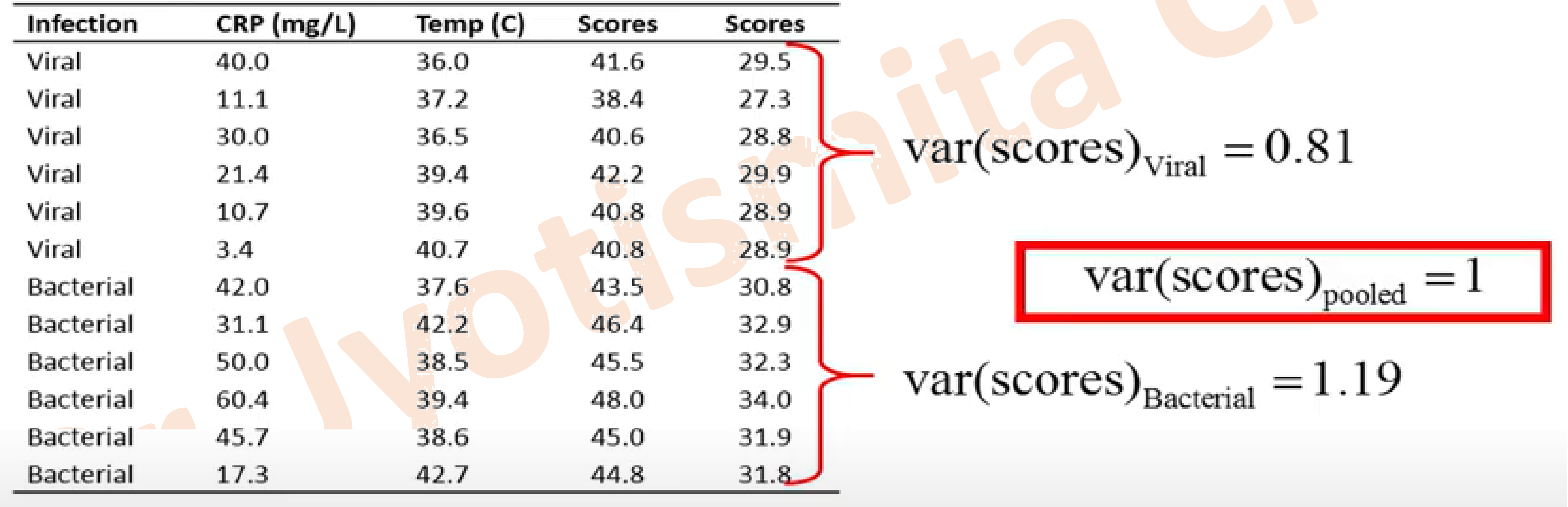


$$\mathrm{LD1} = \boxed{0.11} \cdot \mathrm{CRP} + \boxed{0.70} \cdot \mathrm{Temp}$$



Infection	CRP (mg/L)	Temp (C)	Scores	Scores
Viral	40.0	36.0	41.6	29.5
Viral	11.1	37.2	38.4	27.3
Viral	30.0	36.5	40.6	28.8
Viral	21.4	39.4	42.2	29.9
Viral	10.7	39.6	40.8	28.9
Viral	3.4	40.7	40.8	28.9
Bacterial	42.0	37.6	43.5	30.8
Bacterial	31.1	42.2	46.4	32.9
Bacterial	50.0	38.5	45.5	32.3
Bacterial	60.4	39.4	48.0	34.0
Bacterial	45.7	38.6	45.0	31.9
Bacterial	17.3	42.7	44.8	31.8

LDA with example



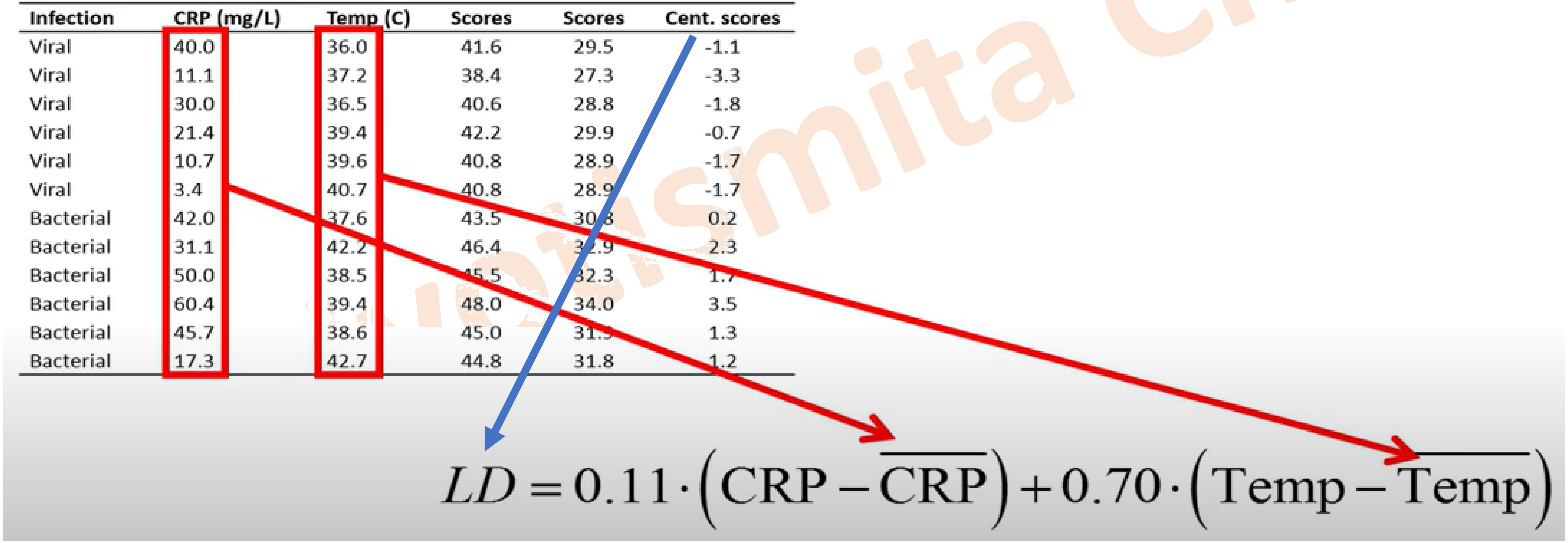
Infection	CRP (mg/L)	Temp (C)	Scores	Scores
Viral	40.0	36.0	41.6	29.5
Viral	11.1	37.2	38.4	27.3
Viral	30.0	36.5	40.6	28.8
Viral	21.4	39.4	42.2	29.9
Viral	10.7	39.6	40.8	28.9
Viral	3.4	40.7	40.8	28.9
Bacterial	42.0	37.6	43.5	30.8
Bacterial	31.1	42.2	46.4	32.9
Bacterial	50.0	38.5	45.5	32.3
Bacterial	60.4	39.4	48.0	34.0
Bacterial	45.7	38.6	45.0	31.9
Bacterial	17.3	42.7	44.8	31.8

$$\text{var}(\text{scores})_{\text{Viral}} = 0.81$$

$$\boxed{\text{var}(\text{scores})_{\text{pooled}} = 1}$$

$$\text{var}(\text{scores})_{\text{Bacterial}} = 1.19$$

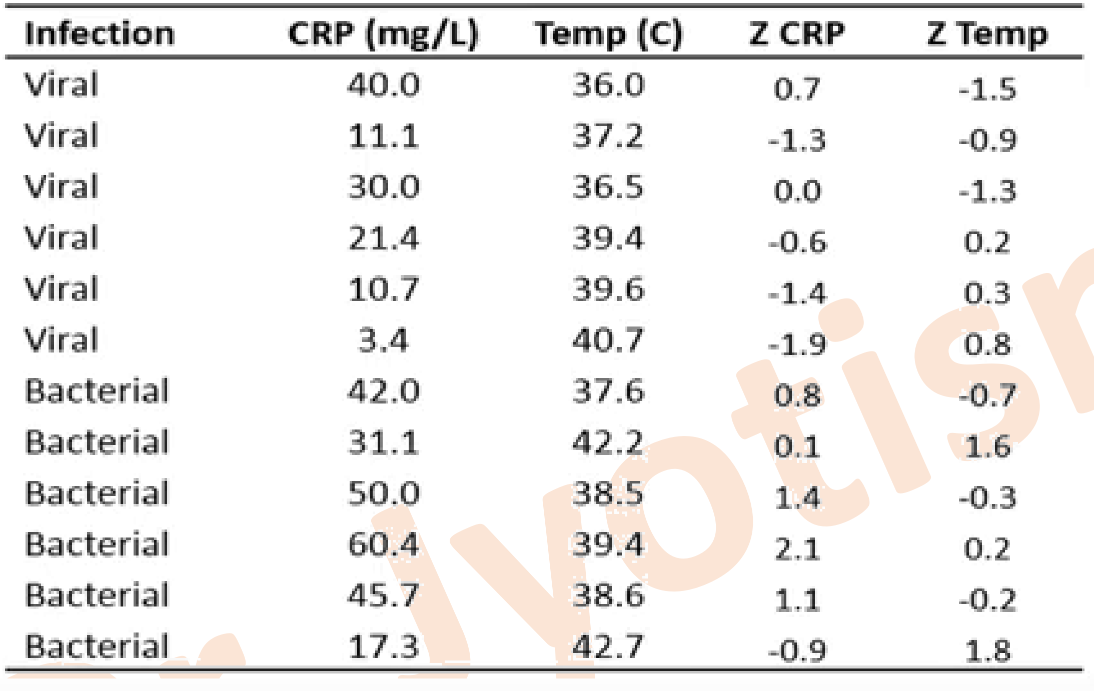
LDA with example



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	28.9	-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	46.4	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

$$LD = 0.11 \cdot \left(\mathrm{CRP} - \overline{\mathrm{CRP}}\right) + 0.70 \cdot \left(\mathrm{Temp} - \overline{\mathrm{Temp}}\right)$$

LDA with example



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

Standardize the calculation using Z value

$$Z = \frac{X - \overline{X}}{\sqrt{\mathrm{var}(X)_{\mathrm{pooled}}}}$$

For Z CRP:

$$\overline{X} = 30.3$$

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

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

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

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

$$\mathrm{LD1} = \boxed{1.53} \cdot \mathrm{zCRP} + \boxed{1.41} \cdot \mathrm{zTemp}$$

These coefficients are usually used to determine how much each variable contributes to the separation