If u is an eigenvector of $S_W^{-1}S_B$, i.e.

$$S_W^{-1} S_B u = \lambda u$$

then
$$u = \frac{1}{\lambda} S_W^{-1} S_B u$$

$$\overset{\text{def of } S_B}{=} \frac{1}{\lambda} S_W^{-1} (\mu^{(1)}-\mu^{(2)})\underbrace{\left((\mu^{(1)}-\mu^{(2)})^T u\right)}_{\text{scalar}}$$

$$= \underbrace{\frac{(\mu^{(1)}-\mu^{(2)})^T u}{\lambda}}_{\text{scalar}} S_W^{-1} (\mu^{(1)}-\mu^{(2)})$$

Thus we can take
$$u = \frac{S_W^{-1}(\mu^{(1)}-\mu^{(2)})}{\|S_W^{-1}(\mu^{(1)}-\mu^{(2)})\|}.$$

Remark: Comparison between PCA and LDA

(1) Both are dimension reduction techniques

(2) PCA maximizes projected data variability, while LDA maximizes projected data separability.

(3) PCA is unsupervised, while LDA is supervised.

4.3.3 Algorithm (LDA for binary classification)

Input: Sample pts with labels $(x^{(1)}, Y_1)$, $\dots$ $(x^{(n)}, Y_n)$ where $Y_i = \pm 1$

Output: 1D direction u with maximal projected data separability.

Step 1: Compute the means and sample covariance matrices of the two classes

$$\mu^{(1)} = \frac{1}{n_1} \sum_{\{i: Y_i = 1\}} x^{(i)} \qquad \mu^{(2)} = \frac{1}{n_2} \sum_{\{i: Y_i = -1\}} x^{(i)}$$

$$C^{(1)} = \frac{1}{n_1} \sum_{\{i: Y_i = 1\}} (x^{(i)} - \mu^{(1)})(x^{(i)} - \mu^{(1)})^T$$

$$C^{(2)} = \frac{1}{n_2} \sum_{\{i : y_i = -1\}} (x^{(i)} - \mu^{(2)})(x^{(i)} - \mu^{(2)})^T$$

where n_1 = # of sample pts with $y_i = 1$

n_2 = - - - - - $y_i = -1$

Step 2: Compute the between-class and within-class scatter matrix S_B and S_W.

$$S_B = (\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})^T$$

$$S_W = n_1 C^{(1)} + n_2 C^{(2)}$$

Step 3: u is an eigenvector associated to the largest eigenvalue of $S_W^{-1} S_B$. Alternatively, u can be taken as

$$u = \frac{S_W^{-1}(\mu^{(1)} - \mu^{(2)})}{\| S_W^{-1}(\mu^{(1)} - \mu^{(2)}) \|}$$

Reference: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Hastie - Tibshirani - Friedman, Section 4.3.