Ex 4.1

W = Within class Variance - How much each observation Varies from its class mean.

$$W = \sum_K \sum_{g_i \in K} (x_i - \hat{\mu}_K)^T (x_i - \hat{\mu}_K) / (N-K)$$

B = Between Class Variance - How much does each cluster vary from the overall mean, $\hat{\mu} = \sum_K \hat{\pi}_K \hat{\mu}_K$

$$B = \sum_K \hat{\pi}_K (\hat{\mu}_K - \hat{\mu})^T (\hat{\mu}_K - \hat{\mu})$$

Now K centroids in P dimensional input space lie in an affine subspace of dimension $\leq K-1$. If P is much larger than K, this is a considerable drop in dimension.

Also notice, to find the nearest centroid to a point we need only to project it to this affine subspace and find the nearest centroid from that point.

This is the inherent dimensionality reduction in LDA. To determine which dimension subspace ($L \leq K-1$) was optimal for LDA, Fisher defined that to mean the linear combination $Z = a^T X$ such that between class variance is maximised relative to within class variance.

$\Rightarrow$ If $Z = a^T X$ is some linear combination of X then,

$$Var_B(Z) = Var_B(a^T X) = a^T Var_B(X) a = a^T B a$$

and $Var_W(Z) = \ldots = a^T W a$

Thus the Problem amounts to $\max\limits_{a} \dfrac{a^T B a}{a^T W a}$ (4.15)

or equivalently $\max\limits_{a} a^T B a$ subject to $a^T W a = 1$

e.g a constrained maximisation Problem solvable by Lagrange Multipliers.

$\Rightarrow \max\limits_{a} a^T B a$ subject to $a^T W a - 1 = 0$

$$L(a) = a^T B a - \lambda(a^T W a - 1)$$

$$\nabla_a L(a) = \nabla_a (a^T B a) - \lambda \nabla_a (a^T W a) = 0$$

$$\Rightarrow \nabla_a (a^T B a) = \lambda \nabla_a (a^T W a)$$

$$= \quad 2Ba = 2\lambda W a$$

$$\Rightarrow \quad Ba = \lambda W a$$

$\Rightarrow W^{-1} B a = \lambda a$ which is just a standard eigenvalue Problem.

The eigenvector (a) corresponding to the largest eigenvalue (λ) will maximise the expression:

$$\frac{a^T B a}{a^T W a} = \frac{a^T(\lambda W a)}{a^T W a} = \lambda$$