1. F(1,2) = -4 < 0

F(0,0) = 1 > 0   
these two data points will be assigned different classes as they are on opposing sides of maximum margin classifier’s hyperplane.

1. We can enlarge the feature space. We can do this efficiently by using kernels turning the support vector classifier into a support vector machine. This is done by calculating the inner product of the observations
2. The most optimal within cluster variation (WCV) is to have 0 variation. This is only possible when each observation belongs to its own cluster, which in turn defeats the purpose of classifying things in groups. A better experiment might be to test where the WCV starts to plateau out for different values of K. Through cross-validation to select the best value of K and how to best initialize the classes used during the algorithm.
3. PCA is mainly used for dimensionality reduction. This makes things like visualization and processing much easier. This often comes at a cost of accuracy of the final model.