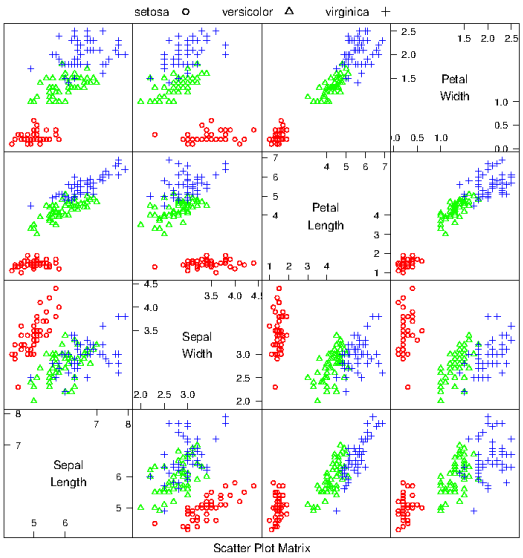
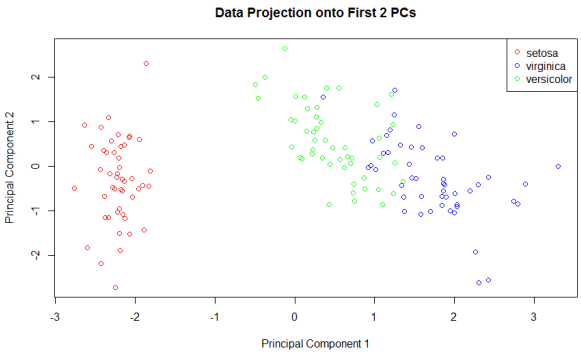
Homework 4
CS498daf, Spring 2016
Cybelle Smith, Varun Somani, David Young

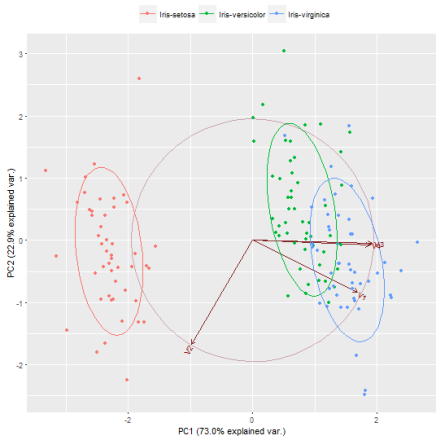
Problem 3.4

a) The plotted scatterplot matrix of this dataset is shown below. Each species is shown with a different marker.

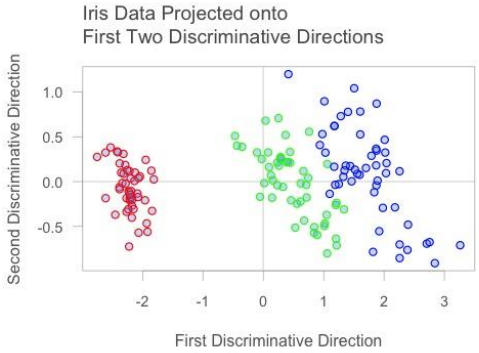


b) The first two principal components of the data were obtained and the data was plotted on those two principal components alone. Again, each species is shown with a different color marker. This process was done using two different packages and plotted twice. Both plots were included for thoroughness. The plot definitely appears to demonstrate greater variance on average than the plots in the scatterplot matrix, and lends itself well to linear classification boundaries. Although there does appear to be more overlap between the virginica and versicolor class spaces than in a few scatterplot matrix cells.



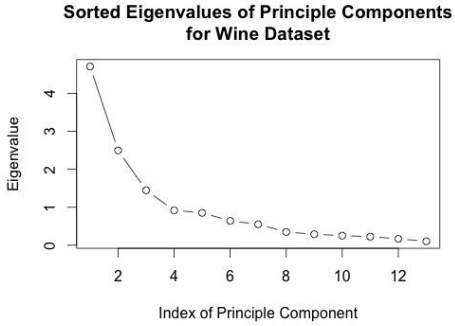


c) Here PLS1 was used to obtain two discriminative directions. The data was then projected on to those directions. The plot does look better. Data in each cluster appears closer to the cluster center, less scattered, with more separation between clusters and fewer outliers.

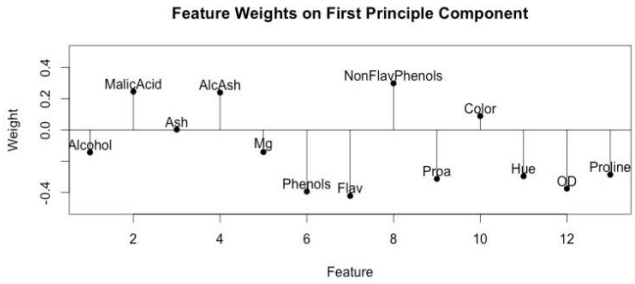


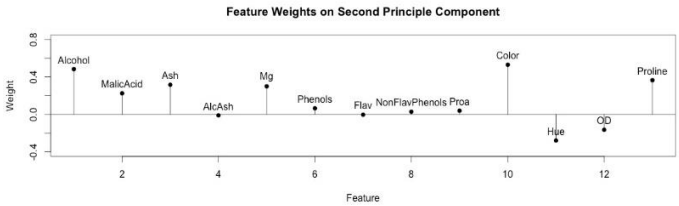
Problem 3.5

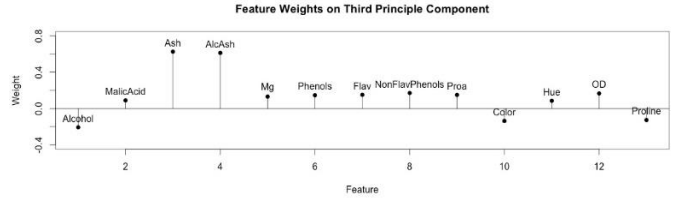
a) The eigenvalues of the covariance matrix were plotted in sorted order. The knee of the sorted eigenvalue plot would indicate that 3 principal components are enough to well represent the data.



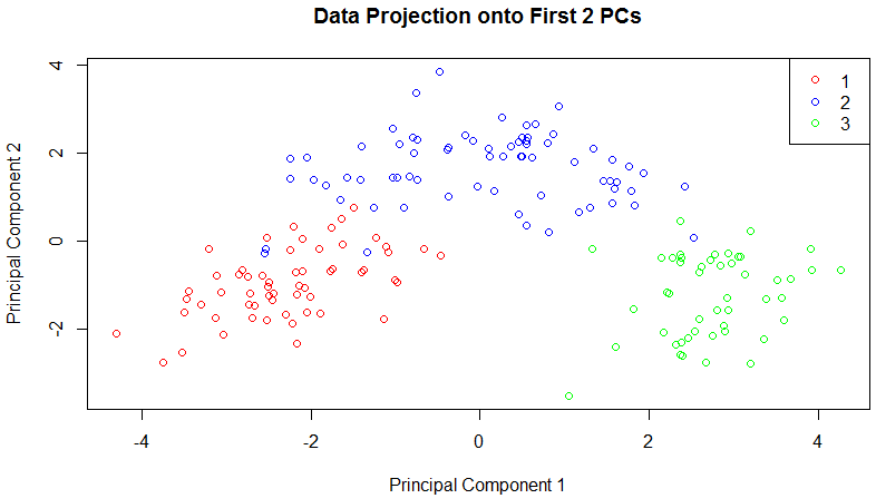
b) A stem plot was constructed for each of the first 3 principal components. The first principal component seems to be influenced/weighted by more components in total than the second and third principal components. In the second and third principal components, features 5-8, 10 and 11 seem to contribute less.





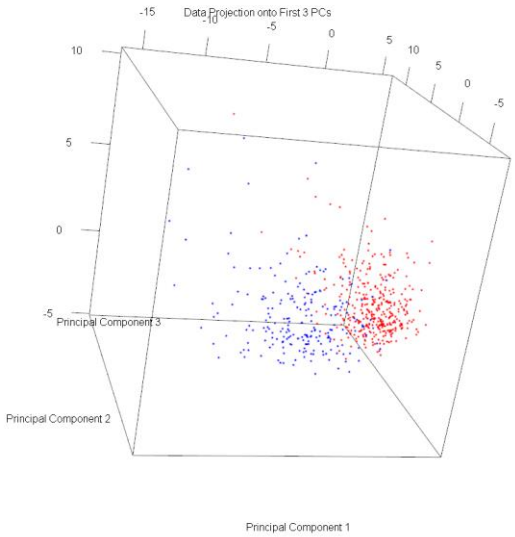


c) The first two principal components were computed for this dataset, and the data was projected onto those components. A scatter plot of this two dimensional dataset was generated.



Problem 3.7

a) The dataset was plotted on the first three principal components, using different markers for benign and malignant cases. It appears that a plane somewhere along the first principal component (somewhat parallel to the plane formed by PC2 and PC3) would best serve as a classification boundary between the two classes.



b) Here PLS1 was used to obtain three discriminative directions. The data was then projected onto those directions. The data looks rather similar but with slightly better separation and not as much overlap.