Homework 6 – Due 12 am CST, 2 April 2012

The total points on this assignment is 100. The dataset is posted on the class website. All problems are from the book and are worth the points indicated in the parenthesis. Datasets are also all available on the class website.

1. Problem 8.10 [15 points]
2. Problem 8.22 [10 points]

3. The United States Postal Service has had a long-term project to automating the recognition of handwritten digits for zip codes. The file available in the usual places has data on different numbers of specimens for each digit. Each observation is in the form of a 256-dimensional vector of pixel intensities. These form a 16x16 image of pixel intensities for each letter. The objective is to distinguish one digit from another.

   (a) We will see whether the digits are distinguishable. To do so, we will first prepare the dataset by rooting out those pixels (coordinates) which do not contribute to categorization. Do so, using appropriate univariate but simultaneous methods, at the 5% level. For the remainder of this question, we will be focused on the 100 most significant coordinates (in terms of \( p \)-value for the above test). [10 points]
      i. Evaluate whether the variance-covariance matrices are all equal across all digits. [5 points]
      ii. If these are not equal, we will assume that they are the true values of the individual group dispersion matrices. Derive a likelihood ratio test statistic for testing differences in mean effects across different digits. [5 points]
      iii. Use a nonparametric bootstrap method to provide a \( p \)-value for the above test. [10 points]

   (b) We will now use principal components to reduce dimensionality of the original dataset. Use the principal components and determine the number of components needed to explain at least 80% of the total variation in the data, at the 5% level of significance. [10 points]
      i. Display the components (using color or characters for each digit) using appropriate methods. Compare with the displays obtained using the reduced dataset. [5 points]
      ii. Repeat (i), (ii) and (iii) above on the principal components of the dataset. (It is helpful for you to write functions that you can then call repeatedly for answering these problems.) [30 points]