625.664 Computational Statistics

Problem Set 13

Associated Reading: Handout: Chapter 8 from Gentle

Complete the problems either by hand or using the computer and upload your final document to the Blackboard course site. All final submittals are to be in PDF form. Please document any code used to solve the problems and include it with your submission.

1. For univariate data $x_1, x_2, \ldots,$ quick insights can be obtained by a "4-plot" that consists of the following four plots

    - plot of x_i versus i to see if there is any trend in the way the data are ordered;
    - plot of x_{i+1} versus x_i to see if there are systematic lags;
    - histogram;
    - a q-q plot of the data versus the normal distribution.

    (a) Posted on the course Blackboard site is the data Univariate.txt. Perform a "4-plot" on the data and discuss the information gained from these preliminary plots.
    (b) Give the Broken-Line ECDF and Mountain plot of the data. What do these plots tell you?

2. Generate a sample of size 200 of pseudorandom numbers from a mixture of two univariate normal distributions. Let the population consist of 80% from a $N(0, 1)$ distribution and 20% from a $N(3, 1)$ distribution. Plot the density of this mixture. Notice that it is bimodal. Now plot a histogram of the data using 9 bins. Is it bimodal? Choose a few different numbers of bins and plot histograms.

3. Generate a sample of pseudorandom numbers from a normal $(0, 1)$ distribution and produce a quantile plot of the sample against a normal $(0, 1)$ distribution, similar to Q-Q plot presented in Lecture 13B. Do the tails of the sample seem light? (How can you tell?) If they do not, generate another sample and plot it. Does this erratic tail behavior indicate problems with the random number generator? Why might you expect often (more than 50% of the time) to see samples with light tails?

4. Posted on the course Blackboard site is the data MultiforGraph.txt. Use any of the techniques discussed in class to obtain information about this data set. What can you say about each of the variables individually? Do you see any dependencies among the variables? You may want to try techniques such as sorting the data or looking at subsets.

5. Posted on the course Blackboard site is the data Bears.txt a subset of a data set described in Reader's Digest (April, 1979) and Sports Afield, (September, 1981). The data set consists of several measurements for bears that were captured, measured, and released. The variables in the data set are:

    1. estimated age in months

2. gender (1=male, 2=female)
3. length of head in inches
4. width of head in inches
5. girth of the neck in inches
6. body length in inches
7. girth of the chest in inches
8. weight in pounds
9. name

Apply the following graphical techniques to the multivariate random variable composed of variables 3 to 7. For each technique, describe its usefulness in analyzing this data set. (Think about finding maximum values, minimum values, grouping the data, correlations, etc.)
(a) Stars
(b) Chernoff faces
(c) Parallel Coordinates

****

A note about commands in R. Feel free to use the built in R commands below to assist you in creating your plots. I have listed the required package after each of the commands. Do not use the built in qq command.
(i) splom - lattice
(ii) parallel - lattice
(iii) faces - TeachingDemos
(iv) stars - graphics
You may need to load the packages - this can be done with the command "library(packagename)". Also, if there are any packages that you do not have, you can obtain them from the Project-R webstite at "http://cran.r-project.org/web/packages/". Simply select the required package, download the appropriate file, and place the folder in your R library folder. I did not have the TeachingDemos package in my initial load of R.