DSC 424
Homework 5

Submit your answers to these in a single PDF file with all relevant explanations and graphs included.

1) **(20 points)** Download the "kellog.dat" data file which contains data on 22 cereals from Kellog. Each cereal has 9 metric values that measure various aspects of the cereal. We are not given the meanings of these variables, but in spite of this use the data to classify the cereals:

    a. Read the data into a data.frame in R. Note that the data file has two extra rows, you can ignore these with the "skip=2" parameter in read.table, or you can manually delete them. Also, you will want to put the cereal names in the row.names with "row.names=1" which indicates to use the first column as the row names.

    b. Compute the distance matrix with "dist". Just treat the ordinal and binary categorical variables as metric variables (this is actually ok here because they are either interval variables), or binary variables encoded as [0, 1].

    c. Run multidimensional scaling on the distance matrix with the "isoMDS" command from the MASS library. This computes MDS and provides a bit more and as its output, providing both an array of "$points" to plot and a stress value. Plot the points from c) and report the stress value. How faithfully does the plot reproduce the distances in the data according to the stress value (remember the stress value from R is actually multiplied by 100 so it is a percentage)?

    d. How many clusters or groups does the data fall into? Can you identify some distinct groupings? Interpret at least two of the groupings of cereals based on their names in the data file.

    e. Run an agglomerative hierarchical clustering on the dataset and plot the result as a dendogram.

    f. At a level of 3 clusters in the dendogram, use the cutree(h, k=3) command to evaluate the clusters and then replot the MDS using these categories to color the data. Interpret the results.

    g. **(Extra Credit, 3 points)** Give a practical interpretation for at least one of the two dimensions in the MDS.

2) **Problem #2 (Canonical Correlation Analysis – 10 points):** Water, soil, and mosquito fish samples were collected at *n* = 165 sites/stations in the marshes of southern Florida. The following water variables were measured:

MEHGSWB	Methyl Mercury in surface water, ng/L
TURB	in situ surface water turbidity
DOCSWD	Dissolved Organic Carbon in surface water, mg/L
SRPRSWFB	Soluble Reactive Phosphorus in surface water,mg/L or ug/L

THGFSFC Total Mercury in mosquitofish (*Gambusia affinis*), average of 7 individuals, ug/kg

In addition, the following soil variables were measured:

THGSDFC Total Mercury in soil, ng/g
TCSDFB Total Carbon in soil, %
TPRSDFB Total Phosphorus in soil, ug/g

Perform a canonical correlation analysis, describing the relationships between the soil and water variables using the data[1] found in data_marsh_cleaned.csv.

1) Answer the following questions regarding the canonical correlations. (Note that a, b and c can all be done directly from the output of canonical correlation)

    a. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.

    b. Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value.

    c. Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value.

    d. Present the three canonical correlations and list any conclusions that you can draw.

2. Answer the following questions regarding the canonical variates.

    a. Give the formulae for the first canonical variate for the soil and water variables.

    b. Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables and use these to interpret the variates (do this as best as you can. Even with a lack of domain knowledge you should be able to describe the relationship in more general terms given the variables involved and the correlations.)

[1] http://www.epa.gov/region4/sesd/reports/epa904r07001.html