Statistics 5525: Homework 4

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

Read *Nonlinear Dimensionality Reduction by Locally Linear Embedding*, Sam T. Roweisand and Lawrence K. Saul, *Science* (2000).

Done.

Problem 2

Read *Probabilistic Principal Component Analysis*, Michael E. Tipping and Christopher M. Bishop (1999).

Done.

Problem 3

Obtain the "cereal" data set from the website. The matrix $X' =< x'_1, \ldots, x_2 2' >$ indicates the features (Calories Protein Fat Sodium Fiber Carbo Sugars Shelf Potass Vitamins) over each of 22 cereals (names indicated at the bottom of the file).

Part a

Perform 2-D Classical MDS on the data set. That is, find lower dimensional coordinates ($z's$ in 2-D), such that the found $z's$ minimize the stress function:

$$\sqrt{\sum_{i<j}(||z_i - z_j||_2 - ||x_i - x_j||_2)^2}. \tag{1}$$

(Note: in Matlab, this is accomplished either via the cmdscale function, or the mdscale function, using Euclidean norms and the "strain" metric).

Here I performed 2-D Classical MDS on the cereal data set with the R package. Before doing the MDS, I rescaled the columns of the data set so that each variable has its mean as zero and standard deviation as one. the MDS mapped the data in ten dimensional feature space into a two dimensional space where the value in expression (1) is minimized. The data in the reduced space is given in Figure 1.

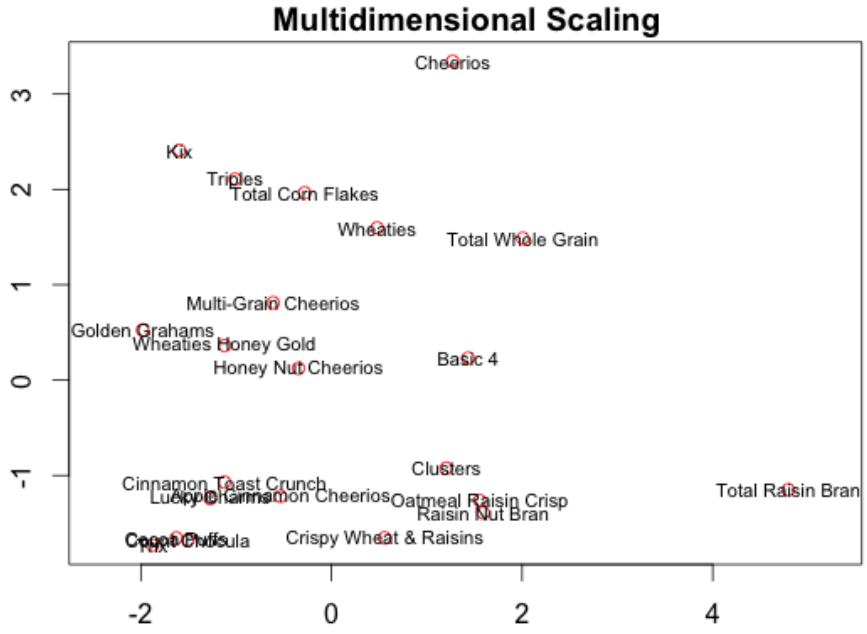

Figure 1: The result of feature space dimension reduction with Multi-dimensional Scaling (MDS).

Part b

Perform a 2-D PCA projection of the data.

Here I coded the algorithm of Principal Component Analysis. Here's the detailed procedures that I have implemented: first, I scaled the columns of the data; second calculate the covariance matrix with the scaled data; third, find the eigenvectors V that diagonalize the covariance matrix; forth, collect the first two columns of V (the two-column matrix is named V_{tranc}); fifth, transform X into a low dimensional space with XV_{tranc}. The data in the reduced space is given in Figure 2.

Part c

Verify that the *relative* distances between those found in the 2-D projections in parts a & b are the same. Make a conclusion.

Comparing the data points in Figure 1 and Figure 2, we can find the relative distances between the data points in Figure 1 and Figure 2 are exactly the same. More careful observation tell us the Figure 2 is just a mirror image of Figure 1

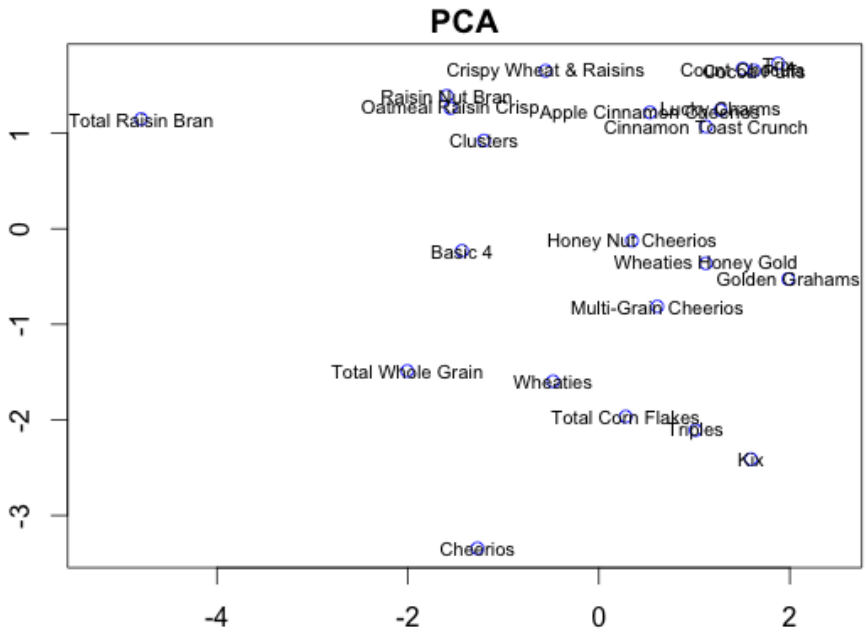

Figure 2: The result of feature space dimension reduction with Principal Component Analysis (PCA).

against the diagonal line from top left corner to the bottom right corner. Figure 3 is a mirror image of Figure 2, and you can see now Figure 3 is exactly the same as Figure 1. This means the classical MDS is equivalent to PCA.

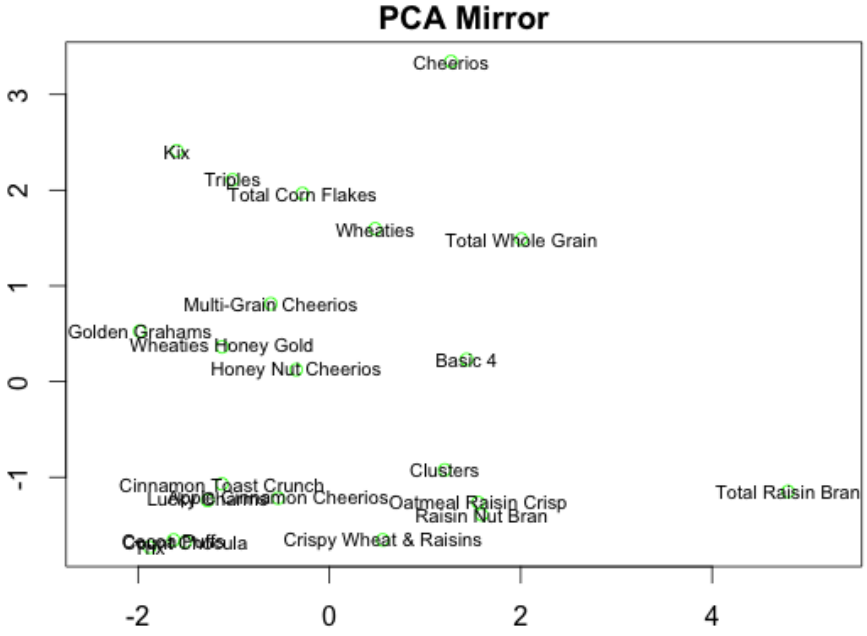

Figure 3: A mirror image of the result of feature space dimension reduction with Principal Component Analysis (PCA).