MSiA 421, Data Mining
PCA and Dimensionality Reduction
Due: Tuesday, March 13 by 11:59pm

1. (27 points) Variation of JWHT 10.11. The data set `gene.csv` consists for 40 tissue samples with measurements of 1000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group. Read the data into R. Note that the first thousand variables are gene measurements, and the 1001th variable (`sick`) equals 1 if the subject is diseased and 0 if healthy.

   (a) (2 points) Run PCA on the data. What fraction of variation is accounted for by the first two components?

   (b) (3 points) Generate a scatterplot of the second component against the first. Encode `sick` using different colors and/or plotting symbols. Submit your scatterplot and briefly discuss what the plot reveals. (For you to think about but not turn in: if you were to build a supervised learning model such as a tree or logistic regression predicting `sick`, what are the virtues of using the principal components as predictors as opposed to the original 1000 X variables?)

2. (Variation on LRU 11.2.1) Let `X=matrix(c(1:4, 1,4,9,16), nrow=4)`.

   (a) Find $X^\mathsf{T}X$ and XX^T. Hint: `t(X)` gives the transpose of X and the `%*%` operator does matrix multiplication.

   (b) Find the eigenvalues and vectors of $X^\mathsf{T}X$.

   (c) Find the eigenvalues and vectors of XX^T.

   (d) Comment on the eigenvalues in parts (b) and (c).

3. (variation of LRU exercise 11.3.1) Consider the following matrix:

```
X = matrix(c(1,3,5,0,1,   2,4,4,2,3,   3,5,3,4,5), nrow=5)
```

   (a) Compute $X^\mathsf{T}X$ and XX^T.

   (b) Find the eigenvalues and eigenvectors of the matrices of part (a). You may use the `eigen` function.

   (c) Confirm that $X^\mathsf{T}X = \Gamma\Lambda\Gamma^\mathsf{T}$ by multiplying the matrices found in part (b). Submit R code/output

   (d) Find the SVD of X and relate it to the results from part (b). The result from the `svd` function should match the results from (b). Follow the definition of SVD given in class, where there are exactly r (the rank of X) singular values.

   (e) Confirm that $X = ULV^\mathsf{T}$ by multiplying the matrices found in part (d).

(f) Set the smaller singular value to 0 and compute the one-dimensional approximation to X, called $\hat{X}$ or `Xhat`.

(g) Compute sum of squared values in $\hat{X}$. How does the sum relate to the singular values?

(h) Compute $X - \hat{X}$, square the values and add them up. How does the sum (of squared "errors") relate to the singular values?

(i) How much of the "energy" of the original singular values is retained by the approximation? Note, by "energy" they mean the squared singular values.

4. Fisherman on 28 lakes in Wisconsin were asked to report the time they spent fishing and how many of each type of fish they caught. Their responses were converted to a catch rate per hour for x_1 = Bluegill, x_2 = Black crappie, x_3 = Smallmouth bass, x_4 = Largemouth bass, x_5 = Walleye, x_6 = Northern pike. The correlation matrix is below based on a sample of about 120 fisherman. Fish caught by the same fisherman live alongside of each other, so the data should provide some evidence on how the fish group. The first four fish belong to the centrarchids, the most plentiful family. Walleye are the most popular fish to eat.

```
R = matrix(c(
  1, .4919, .2636, .4653, -.2277, .0652,
  .4919, 1, .3127, .3506, -.1917, .2045,
  .2635, .3127, 1, .4108, .0647, .2493,
  .4653, .3506, .4108, 1, -.2249, .2293,
  -.2277, -.1917, .0647, -.2249, 1, -.2144,
  .0652, .2045, .2493, .2293, -.2144, 1),
  byrow=T, ncol=6)
```

(a) Comment on the pattern of correlation within the centrarchid family. Does the walleye appear to group with other fish?

(b) Perform a PCA using only x_1–x_4. Interpret the first two loading vectors.

(c) Perform a PCA on all six variables. Generate a scree plot by plotting the eigenvalues against 1–6.

(d) What fraction of variation is accounted for by the first two PCs in part (c)?

(e) Interpret the first loading vector from part (c).