

# Homework 2

## Statistics W4240: Data Mining

### Columbia University

#### Due Tuesday, September 30 (All Sections)

For your .R submission, submit a file for question 2 labeled `hw02_q2.R`. The write up should be saved as a .pdf and under 4MB. Submit all files through Courseworks:

**Problem 0. (5 Points Extra Credit)** Fill out the questionnaire on Google Forms about class conditions.

[https://docs.google.com/a/columbia.edu/forms/d/1d0TDoE-70mg\\_2D4ukHJf9mukBBT4ocGrBczLzY5ql-E/viewform?usp=send\\_form](https://docs.google.com/a/columbia.edu/forms/d/1d0TDoE-70mg_2D4ukHJf9mukBBT4ocGrBczLzY5ql-E/viewform?usp=send_form)

**Problem 1. (35 Points)** In this problem we will manually go through all of the steps for PCA. Basic computations like finding the eigenvalues for a matrix may be done using R.

- (2 Points) Load `hw02.q1.p1.fall14.csv`. Find the column means and the row means for the data. What do these values tell us about this data set?
- (3 Points) Center the data and find the empirical covariance matrix,  $\hat{\Sigma}$ . This should be a 5-by-5 matrix. What do the diagonal values of the covariance matrix tell us about this data set? What do the off diagonal elements tell us about this data set?
- (5 Points) Give the eigenvalues and associated eigenvectors of  $\hat{\Sigma}$ . Why does this matrix have the same left eigenvectors as right eigenvectors?

$$\mathbf{x}_{left}^T \hat{\Sigma} = \lambda \mathbf{x}_{left}^T, \quad \hat{\Sigma} \mathbf{x}_{right} = \lambda \mathbf{x}_{right}$$

- (5 Points) Give all of the loadings and all of the scores for the data.
- (5 Points) Plot the proportion of variance captured against the number of components included. How many components should we include and why?
- (5 Points) Load `hw02.q1.p2.fall14.csv`. This has 5 new observations in the original coordinates. Give their scores.
- (5 Points) Now use only the first two scores to represent the observations from the previous part. What are the coordinates of the projections in the original space,  $\mathbf{x}'$ ? What is their Euclidean distance from the original data points?
- (5 Points) Define the error of a point as

$$d(\mathbf{x}', \mathbf{x}) = \mathbf{x}' - \mathbf{x},$$

which is a 5-dimensional vector of errors. In what direction is  $d(\mathbf{x}', \mathbf{x})$  for the 5 new points? Why do you think this is?

**Problem 2. (65 Points)** We will continue working with the Yale Faces B data set from the last homework with the goal of representing the images using PCA. We will use four lighting conditions, P00A+000E+00, P00A+005E+10, P00A+005E-10, and P00A+010E+00, which are closest to straight on lighting. We will use the `pixmap` library to manipulate the data. Load this library and make sure that the folder `YaleCropped` is in your working directory.

- a. (10 Points) Load the views P00A+000E+00, P00A+005E+10, P00A+005E-10, and P00A+010E+00 for all subjects. Convert each photo to a *vector*; store the collection as a matrix where each row is a photo. What is the size of this matrix?
- b. (10 Points) Compute a “mean face,” which is the average for each pixel across all of the faces. Subtract this off each of the faces. Display the mean face as a photo in the original size and save a copy as `.png`. Include this in your write up.
- c. (10 Points) Use `prcomp()` to find the principal components of your image matrix. Plot the number of components on the x-axis against the proportion of the variance explained on the y-axis.
- d. (10 Points) Each principal component is a picture, which are called “eigenfaces.” Display the first 9 eigenfaces in a 3-by-3 grid. What image components does each describe? (Note: `pixmapGrey()` is fairly flexible and will automatically rescale data to have min 0 and max 1. You can do this manually or allow `pixmapGrey()` to do it.)
- e. (15 Points) Use the eigenfaces to reconstruct `yaleB01_P00A+010E+00.pgm`. Starting with the mean face, add in one eigenface (weighted principal component) at a time until you reach 24 eigenfaces. Save the results in a 5-by-5 grid. Again, starting with the mean face, add in five eigenfaces at a time until you reach 120 eigenfaces. Save the results in a 5-by-5 grid. Include both of these in your write up. How many faces do you feel like you need until you can recognize the person?
- f. (10 Points) Remove the pictures of subject 05 from your image matrix (there should be four pictures of him) and recenter the data. Rerun `prcomp()` to get new principal components. Use these to reconstruct `yaleB05_P00A+010E+00.pgm`. Do this by subtracting off the mean face and projecting the remaining image onto the principal components. Print the reconstructed image. Does it look like the original image? Why or why not?