Assignment 2

ISOM 674

Principle Components Decomposition and Dimension Reduction for Images

The file SelfieImageData.csv contains the grayscale image data for the usable images that students submitted as a task in an earlier assignment.

The images are 451x451 = 203,401 pixels. As you know from the discussion in class, these images have been carefully rotated and scaled so that each face's eyes are located in the same place in the image. The center of the eyes are on a horizontal line at pixel row 226, the eyes are 101 pixels apart, and the center between the eyes is at the image center which is at pixel (226,226).

The data is stored as a CSV file with the "|" character used as a delimiter. The first row contains the column labels. The first column contains the student's NetID taken from the image file name that was submitted. The 203,401 columns that follow the first contains the pixel values arranged in row precedence. The labels for these columns in the first row are of the form `Px(i,j)`, where `i` is the row number and `j` is the column number. Indexing starts at 1, so the range of values for `i` and `j` are 1 to 451. The values of the pixels range from 0 for pure black to 255 for pure white.

To aid you with this assignment, I have provided a template R script that reads in the data and creates a matrix called ImgData that contains the pixel values for each image as rows in the matrix. The `Px(i,j)` labels are the column names of this matrix and the NetIDs are the row names. The template R-script also plots out each of the images.

Some Hints

Your assignment will be to use principle components decomposition to reconstruct the images based on a lower dimensional representation. In doing this you will need to do things like rescale vectors so that they have length 1, subtract out column means, and so on. As a hint, the following two functions may be very useful:

```
apply(matrix,dim,FUN=function)

sweep(matrix,dim,statistics,function)
```

If you are not familiar with these functions, you should read the documentation. The function `apply` can be used to do things like calculate column means (in that case, dim=2 and function = mean). The function `sweep` can be used to subtract out means (FUN="-") or rescale a column to have length 1 (FUN="/").

Depending upon how you write your code, the drop=F argument (default is drop=T) to the subsetting function (represented by square brackets [ ]) may be useful. Suppose that X is an n×p

matrix. In R, X[,1] returns a vector, not a matrix. Thus, X[,1] is likely not to work in a matrix multiplication. However, X[,1,drop=F] returns an n×1 matrix instead of a vector.

One other hint that you may find useful. When an eigenvalue is numerically very close to 0, it usually does not make sense to use the corresponding eigenvector (it explains essentially no variance) and may cause numerical problems if you do use it.

Google Form Questions

The questions for this part of the assignment can be found on the Google Form at the link provided in the assignment module.

Unfortunately, Google Forms does not have the ability to use math notation in questions. So just to be clear, when I refer to the sample variance-covariance matrix of the images or to SigmaHat, in the Google Forms questions, I am referring to $\hat{\Sigma}$ as defined in the slides PrincipleComponentsForImages.pptx.

The Google Forms questions ask that you upload images in png format. Save the png images from the plot window in RStudio by first clicking on the Zoom button above the plot and right-clicking on the large image that appears and "Save Image As." The plots should be square and fill the window just as the images do in the code template I provided with the assignment that reads in the data and plots the images.

The Google Forms Questions ask for you to reduce the dimension and reconstruct the images in two ways. The

Please name the files EXACTLY as indicated substituting your Emory Network ID for NetID. Also, please title each plot with your NetID and what the plot is as indicated below. I need to know what I am looking at both when I look at the file names and when I have the file open. For clarity, I am summarizing this below.

1. For Q2, the Average Face:

   File name: `NetID-Q2-AveFace.png`
   Plot Title: NetID: Average Face

2. For Q5, the Scree Plot

   File name: `NetID-Q5-ScreePlot.png`
   Plot Title: NetID: Scree Plot

3. For Q8, Your (reconstructed) Face using 20 dimensions based on the original data

   File name: `NetID-Q8-MyFace20.png`
   Plot Title: NetID: My Face 20D

4. For Q10, the Eigenface for eigenvector 8

   File name: `NetID-Q10-Eigenface8.png`
   Plot Title: NetID: Eigenface 8

5. For Q14, Your (reconstructed) face using 20 dimensions based on the centered data. Note that you will have to add back in the mean face to correctly reconstruct the image.

File name:	`NetID-Q8-MyFace20-From-Residuals.png`
Plot Title:	NetID: My Face 20D

Remember, substitute your Emory Network ID for NetID in the above.