DS 100: Principles and Techniques of Data Science **Date: July 10, 2019**

Discussion #6 Worksheet

Name:

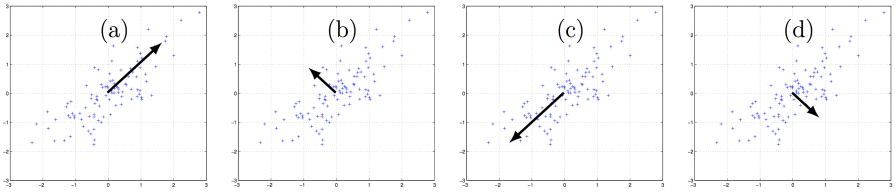
Dimensionality Reduction

1. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques because it is relatively easy to compute and its output is interpretable. To get a better understanding of what PCA is doing to a dataset, let's imagine applying it to points contained within this surfboard. The origin is in the center of the board, and each point within the board has three attributes: how far (in inches) along the board's length, width, and thickness the point is from the center. These three dimensions determine the spread of the data.

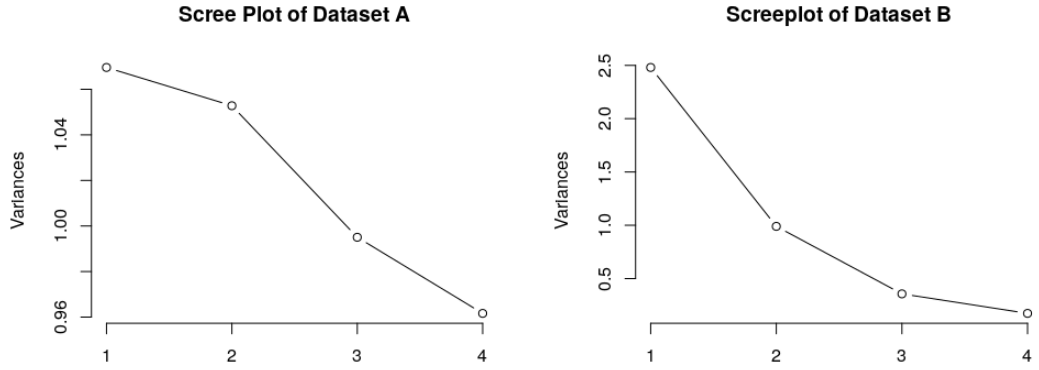
   (a) If we were to apply PCA to the surfboard, what would the first three principal components (PCs) represent? Feel free to draw and label these dimensions on the image of the surfboard.

   (b) Which of the three PCs should be used to create a 2D representation of the surfboard? How come? Make a sketch of the 2D projection below.

2. Which of the following figures correspond to possible values that PCA may return for the first eigenvector / first principal component?



3. Compare the scree plots produced by performing PCA on dataset A and on dataset B. For which of the datasets would PCA provide a scatter plot that describes the variability of the data without leaving out much information? Note that the columns of both datasets were centered to have means of 0 and scaled to have a variance of 1.



4. You perform principal component analysis on a data matrix D using the following Python code from lecture:

   n = D.shape[0]
   X = (D - np.mean(D, axis=0)) / np.sqrt(n)
   u, s, vt = np.linalg.svd(X, full$_matrices = False$)

   The resulting value of s is $np.array([3, 1, 0, 0, 0])$.

   a) To draw a histogram of the data's distribution along the first principal component of X, which of the following arrays would you visualize?

   ◯ `X @ u.T[:,0]` ◯ `(u * s)[:,0]`
   ◯ `X @ vt[0,:]` ◯ `(X @ vt.T)[:,0]`

   b) What proportion of the total variance in D is accounted for by the first principal component?

   c) What is the rank of X?