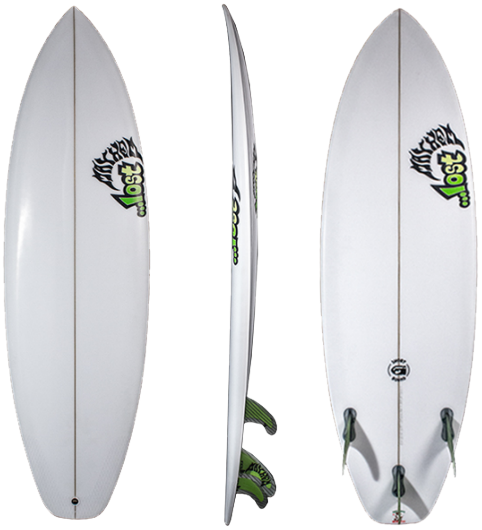
DS 100: Principles and Techniques of Data Science **Date: July 10, 2019**

Discussion #6 Solutions

Name:

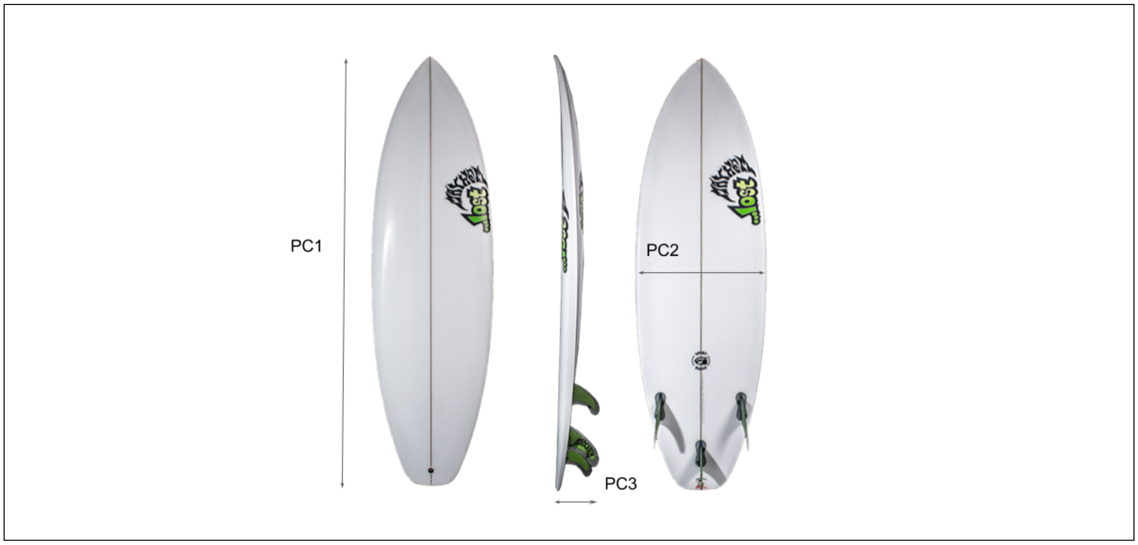
Dimensionality Reduction

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.



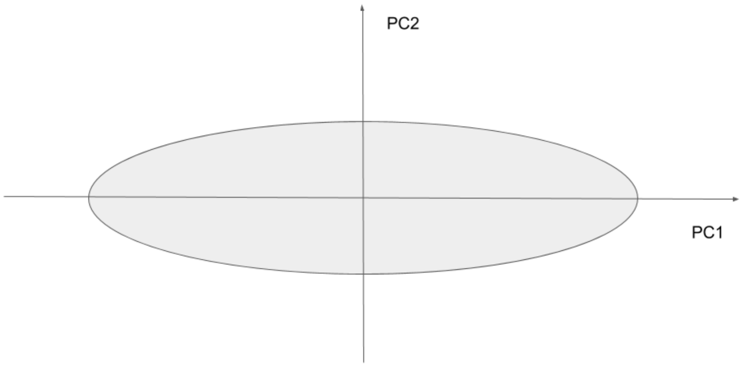
(a) If we were to apply PCA to the surfboard, what would the first three principal components (PCs) represent? Feel free to draw and label these dimensions on the image of the surfboard.

Solution: Since the length of the board (nose to tail) is the longest dimension of the board (e.g. the dimension of the data with the most variation), the first PC would align with the length. The second PC would align with the width of the board, since the width is orthogonal to the length and is more variable than the thickness. Finally, the third PC would be the thickness of the board, which is orthogonal to the first two.

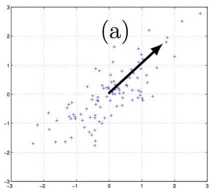


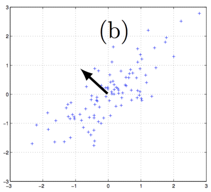
(b) Which of the three PCs should be used to create a 2D representation of the surfboard? How come? Make a sketch of the 2D projection below.

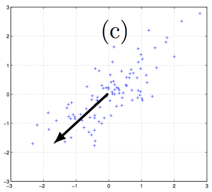
Solution: The first two PCs should be used for the 2D projection of the surfboard, since they are the pair of PCs which contain the most information about the original data.

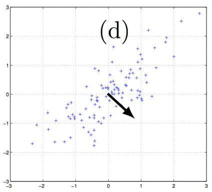


2. Which of the following figures correspond to possible values that PCA may return for the first eigenvector / first principal component?



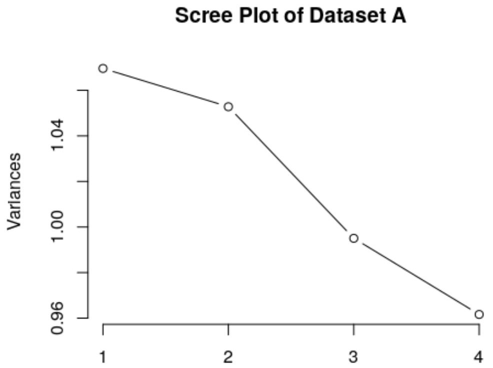


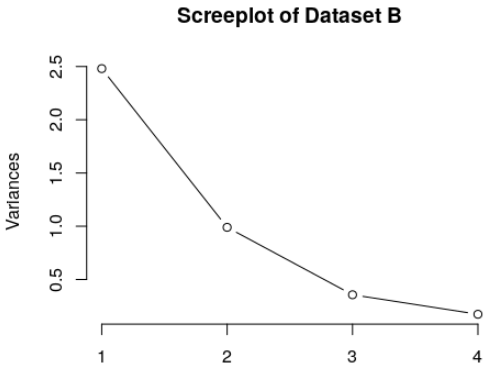




Solution: (a) The maximal variance is along the y = x line, so this option is correct. (b) The first principal component is aligned with the direction of maximal variance, but this is aligned with the direction of minimal variance. (c) The maximal variance is along the y = x line, so the negative vector along that line is correct for the first principal component. (d) see (b)

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.





Solution: PCA is a good choice for reducing dataset B to 2 dimensions, but not dataset A. Paying close attention to the y-axis of dataset A's screeplot, it is apparent that the four largest PCs have eigenvalues of roughly equal size. This signifies that a low-dimensional representation of this dataset using only two PCs would omit a substantial amount of the variability within the data. On the other hand, dataset B's scree plot clearly shows that the first two PCs account for a majority of the variability in the data. We can use these PCs to produce a two-dimensional representation of the data without losing much information.

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?

Solution: $\frac{9^2}{9^2+1^2+0^2+0^2+0^2} = \frac{9}{10}$

c) What is the rank of X?

Solution: 2