DS 100/200: Principles and Techniques of Data Science **Date: May 1, 2020**

Discussion #13

Name:

PCA

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.

1 : length

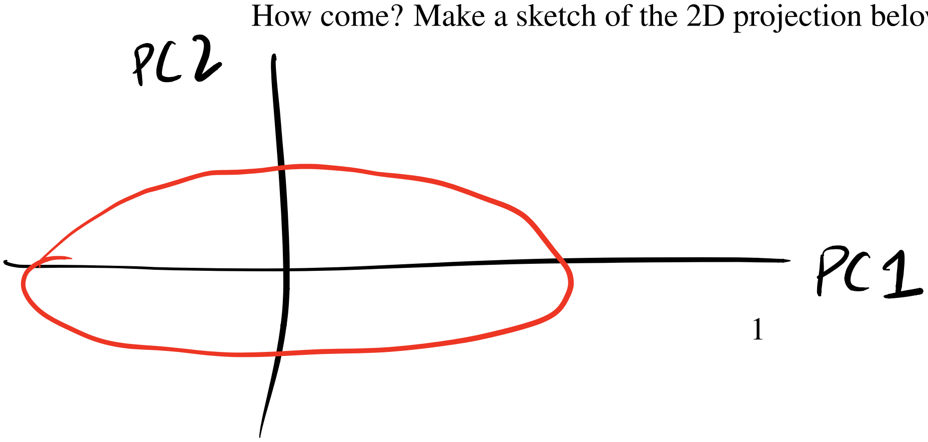
2 : width

3 : depth/thickness

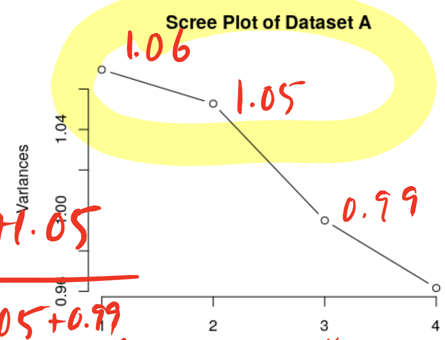
(b) Which of the three PCs should be used to create a 2D representation of the surfboard? How come? Make a sketch of the 2D projection below.

length, width

→ looks like flat surfboard

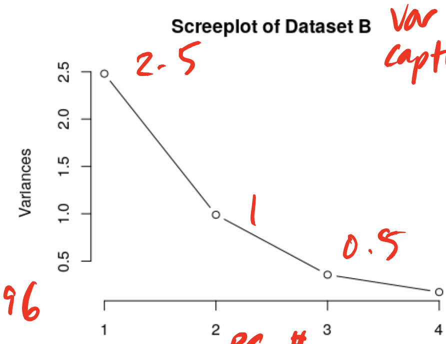


2. Compare the scree plots produced by performing PCA on dataset A and on dataset B. For which dataset would PCA provide the most informative scatter-plot (i.e. plotting PC1 and PC2)? Note that the columns of both datasets were centered to have means of 0 and scaled to have a variance of 1.



var captured $= \dfrac{1.06+1.05}{1.06+1.05+0.99+0.96}$

$\approx 50\%$



Var captured $= \dfrac{2.5+1}{2.5+1+0.5+0.1}$

$\boxed{\approx 80\%}$

3. Consider the following dataset X:

$PC1 = 1 \times V_1 + 0.02 \times V_2 + 0 \times V_3$

Observations	Variable 1	Variable 2	Variable 3
1	-3.59	7.39	-0.78
2	-8.37	-5.32	0.90
3	1.75	-0.61	-0.62
4	10.21	-1.46	0.50
Mean	0	0	0
Variance	63.42	28.47	0.68

After performing PCA on this data, we find that $X = U\Sigma V^\top$, where:

$X \cdot V[:,0] = \begin{bmatrix} -3.44 \\ -8.47 \\ 1.74 \\ 10.18 \end{bmatrix}$

$$U = \begin{bmatrix} -0.25 & 0.81 & 0.20 \\ -0.61 & -0.56 & 0.24 \\ 0.13 & -0.06 & -0.85 \\ 0.74 & -0.18 & 0.41 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 13.79 & 0 & 0 \\ 0 & 9.32 & 0 \\ 0 & 0 & 0.81 \end{bmatrix}$$

$$V = \begin{bmatrix} 1.00 & -0.02 & 0.00 \\ 0.02 & 0.99 & 0.13 \\ 0.00 & -0.13 & 0.99 \end{bmatrix}$$

first PC = first col of $U\Sigma$

Note: Values were rounded to 2 decimals. U and V are not perfectly orthonormal due to approximation error.

13.79 0 0

$U[:,0] \cdot \Sigma[:,0] = \begin{bmatrix} -0.25 & \cdots \\ -0.61 & \cdots \\ 0.13 & \\ 0.74 & \end{bmatrix} \begin{bmatrix} 13.79 & \cdots \\ 0 & \\ 0 & \end{bmatrix} = \begin{bmatrix} 13.79 \cdot (-0.25) \\ 13.79 \cdot (-0.61) \\ 13.79 \cdot (0.13) \\ 13.79 \cdot (0.74) \end{bmatrix}$

$$= \begin{bmatrix} -3.44 \\ -8.47 \\ 1.74 \\ 10.18 \end{bmatrix}$$

(a) The first principal component can be computed through two approaches:

1. Using the left-singular matrix and the diagonal matrix.
2. Using the right singular-matrix and the data matrix. **Hint:** Shuffle the terms of the SVD.

Compute the first principal component using both approaches (round to 2 decimals).

(b) Given the results of (a), how can we interpret the columns of V? What do the values in these columns represent?

Columns of V contain "instructions" of how to project onto PC

(c) Is there a relationship between the largest entries in the columns of V and the variances of X's variables? If so, what is it?

Clustering

4. (a) Describe the difference between clustering and classification.

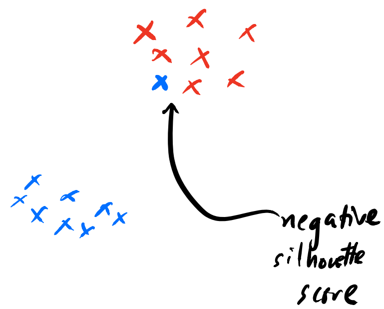
Classification is supervised (know the labels), clustering is unsupervised (no target)

(b) The process of fitting a K-means model outputs a set of K centers. We can compute the quality of the output by computing the distortion on the dataset. A Data 100 student suggests that distortion is not well-defined when evaluating the output of our agglomerative clustering algorithm because the algorithm doesn't return centers, but simply labels each point individually. Is the student correct?

student wrong → can calculate centroids ourselves and then compute distortion

(c) Describe qualitatively what it means for a data point to have a negative silhouette score.

$$S = \frac{B - A}{\max(A, B)} \qquad S < 0 \Rightarrow B < A$$


negative
silhouette
score