DS 100/200: Principles and Techniques of Data Science | Date: October 9, 2019

# Discussion #7 Solutions

*Name:*

## Dimensionality Reduction

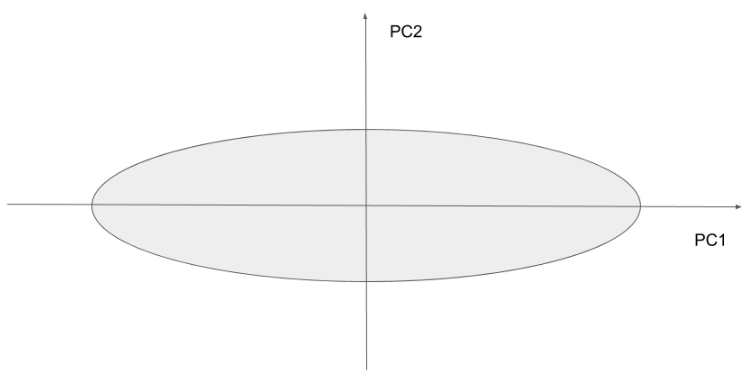
1. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques because it is relatively easy to compute and its output is interpretable. To get a better understanding of what PCA is doing to a dataset, let's imagine applying it to points contained within this surfboard. 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.

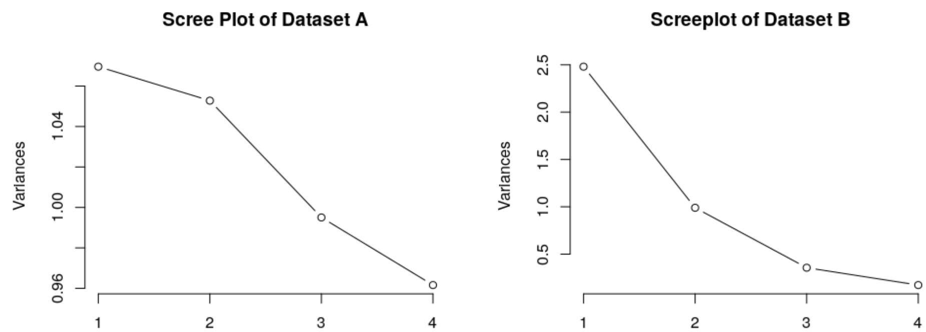
   **Solution:** Since the length of the board (nose to tail) is the longest dimension of the board (e.g. the dimension of the data with the most variation), the first PC would align with the length. The second PC would align with the width of the board, since the width is orthogonal to the length and is 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. 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.




**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.

3. Consider the following dataset $X$:

| 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:

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

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

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

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

(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).

> **Solution:** Computing the first principal component using the first approach is straightforward: it is the first column of $U\Sigma$., i.e. $U[:, 0:1]{*}13.79 \approx [-3.44\ -8.47\ 1.74\ 10.18]^\top$ (your values may differ slightly due to rounding).
>
> Computing principal components using the loadings matrix and the data matrix is also straightforward once you realize that $X = U\Sigma V^\top \iff XV = U\Sigma$. Therefore, we can compute the principal components of the data matrix by multiplying it by the transpose of the loadings matrix. We compute the first principal component by multiplying $X$ by the first column of $V$: $X(V[:, 0:1]) \approx [-3.44\ -8.47\ 1.74\ 10.18]^\top$ (your values may differ slightly due to rounding).

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

> **Solution:** Each principal component of $X$ is a linear combinations of $X$'s variables. The columns of $V$ correspond to the weights of each variable in the linear combinations that make up their respective principal components.

(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?

> **Solution:** Yes, there is! Since the entries in the columns of $V$ are the weights of the linear combinations in their respective principal components, the variables with the largest entries in the columns of $V$ are the most "important" for their corresponding principal components. Since the first principal component is the projection that maximizes the variance in $X$, it is natural that variable 1, the variable with largest variance, has the largest entry in the first column of $V$. Similarly, variables 2 and 3 have the largest entries in the second and third columns of $V$ since they are the most important for the second and third principal components of $X$.