DS 100/200: Principles and Techniques of Data Science Date: October 9, 2019 Discussion #7 Name: Raguric Kunani

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. 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.



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 = | [[1 | 0.43 1.07 0.22 28 | $\begin{array}{c} 1.39 \\ -0.97 \\ -0.10 \\ -0.32 \end{array}$ | $\begin{array}{c} 0.34 \\ 0.41 \\ -1.47 \\ 0.71 \end{array}$ |
|----------|---------|--|--|--|
| \sum = | = | $\begin{bmatrix} 7.96\\0\\0 \end{bmatrix}$ | $\begin{array}{c} 0 \\ 5.38 \\ 0 \end{array}$ | $\begin{bmatrix} 0\\0\\0.47 \end{bmatrix}$ |
| V = | = | $1.00 \\ 0.02 \\ 0.00$ | -0.02 0.99 -0.13 | $\begin{array}{c} 0.00\\ 0.13\\ 0.99 \end{array}$ |

- (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?
- (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?

(a) 1. First principal component = first column of UZ

$$U = \begin{bmatrix} -0.43 & 1.39 & 0.34 \\ -1.07 & -0.97 & 0.41 \\ 0.22 & -0.10 & -1.47 \\ 1.28 & -0.32 & 0.71 \end{bmatrix} = \begin{bmatrix} -u_{1} & -u_{2} & -u_{3} &$$

2. First principal component = first column of XVwhy? We already know $X = UZV^T$. This also means $XV = UZV^TV$. Since V is orthogonal, $V^TV = I$, so $UZV^TV = UZ$. Thus, we have that XV = UZ.

(b) Each column of V holds the weights of a linear combination of the columns of X. For each column of V, the resulting linear combination is a principal component. (c) Remember that each column of V holds the coefficients of a linear combination of X's columns. Thus, a large entry in a column of V means that column of X contributes q lot to the corresponding principal component. For example, the 1.00 entry in the first column of V means that Variable I contributes most to PC1.