DS 100: Principles and Techniques of Data Science **Date: February 22, 2019**

Discussion #5 Solutions

Name:

Dimensionality Reduction

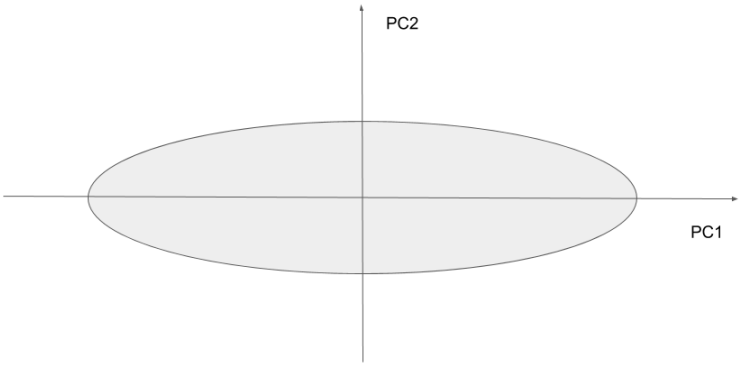
1. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques because it is relatively easy to compute and its output is interpretable. To get a better understanding of what PCA is doing to a dataset, let's imagine applying it to points contained within this surfboard.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

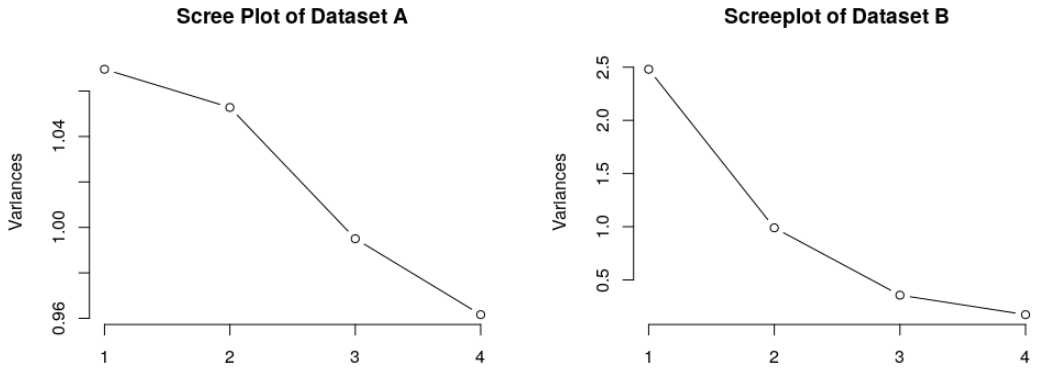
> **Solution:** Since the length of the board (nose to tail) is the longest dimension of the board (e.g. the dimension of the data with the most variation), the first PC would align with the length. The second PC would align with the width of the board, since the width is orthogonal to the length and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

Midterm Review

1. Probability and Sampling

3. A small town has 5 houses with the following people living in each house:

Suppose we take a **cluster sample** of 2 houses (without replacement), what is the chance that:

(a) Kim and Lars are in the sample

◯ 0 ◯ 1/20 ◯ 1/10 ◯ 1/6 ◯ 1/5 ◯ 2/5 ◯ 1

You may show your work in the following box for partial credit:

Solution: The chance that Kim and Lars are in the same sample is given by the chance of choosing their house. The chance of choosing the their house on the first draw is $\frac{1}{5}$. Because we are drawing without replacement. The chance of choosing their house on the second draw is given by the chance of not choosing their house on the first draw ($\frac{4}{5}$) times the chance of choosing their house on the second draw ($\frac{1}{4}$). Thus the total chance of choosing them in the first two draws is:

$$\frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = \frac{2}{5}$$

(b) Kim, Abe, and Ben are in the sample

◯ 0 ◯ 1/20 ◯ 1/10 ◯ 1/6 ◯ 1/5 ◯ 2/5 ◯ 1

You may show your work in the following box for partial credit:

Solution: To draw Kim, Abe, and Ben we would need to draw both of their houses. This can be done two ways (draw Abe and Ben's house first and then Kim's or vice versa). Each way has probability:

$$\frac{1}{5} \times \frac{1}{4}$$

Thus the total probability is:

$$2 \times \frac{1}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$$

(c) Kim and Dan are in the sample - **Select all that apply**

- ☐ The same as the chance Kim and Lars are in the sample
- ☐ The same as the chance Kim, Abe, and Ben are in the sample
- ☐ Neither of the above

Solution: Similarly, the probability that Kim and Dan are in the sample is $\frac{1}{\binom{5}{2}}$, which is the same as the chance Kim, Abe, and Ben are in the sample.

2. Transformations and Smoothing

4. Which of the following are reasonable motivations for applying a power transformation? **Select all that apply**:

   ☐ To help visualize highly skewed distributions

   ☐ Bring data distribution closer to random sampling

   ☐ To help straighten relationships between pairs of variables.

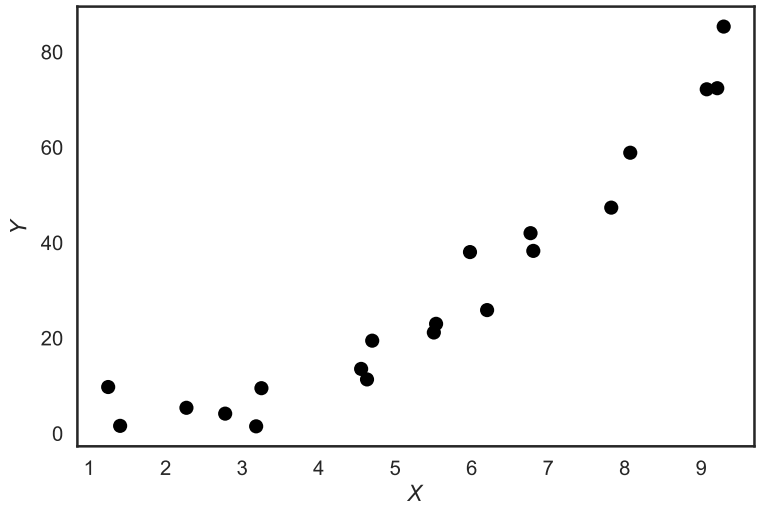
   ☐ Reduce the dimension of data

   ☐ Remove missing values

Solution:

√ To help visualize highly skewed distributions

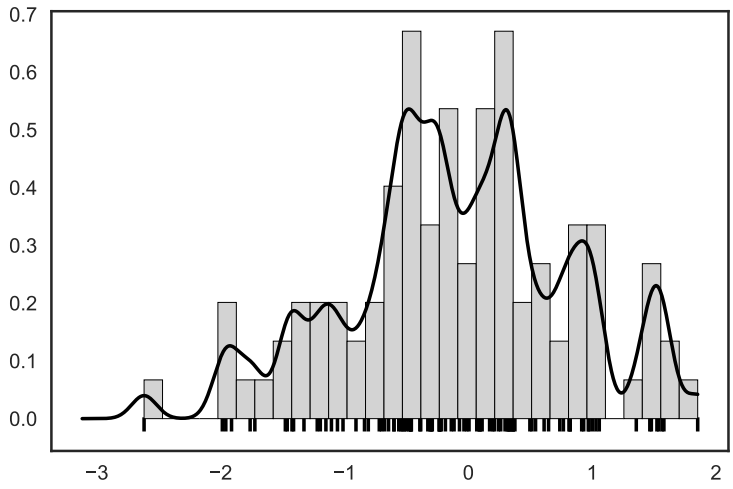
√ To help straighten relationships between pairs of variables.



5. Which of the following transformations could help make linear the relationship shown in the plot below? **Select all that apply**:

☐ $\log(y)$ ☐ x^2 ☐ $\sqrt{y}$ ☐ $\log(x)$ ☐ y^2 ☐ None of the above

Solution: y is proportional to the square of x. Plotting y vs x^2, $\sqrt{y}$ vs x, or $\log y$ vs x will linearize the relationship.



6. The above plot contains a histogram, rug plot, and Gaussian kernel density estimator. The Gaussian kernel is defined by:

$$K_\alpha(x, z) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-z)^2}{2\alpha^2}\right)$$

Judging from the shape of separate standing peaks, which of the following is the most likely value for the kernel parameter α.

◯ $\alpha = 0$ ◯ $\alpha = 0.1$ ◯ $\alpha = 10$ ◯ $\alpha = 100$

Solution: $\alpha = 0.1$. The value α determines the width of the Gaussian kernel. A large $\alpha = \{10, 100\}$ will over-smooth the density estimation and mask the structure of the data, while a small $\alpha = 0$ will yield a density estimation spiky.