Data 100, Spring 2022	April 8, 2022

Discussion #11

PCA

1. 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 the SVD on this data, we obtain $X = U\Sigma V^T$, where:

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

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

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

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

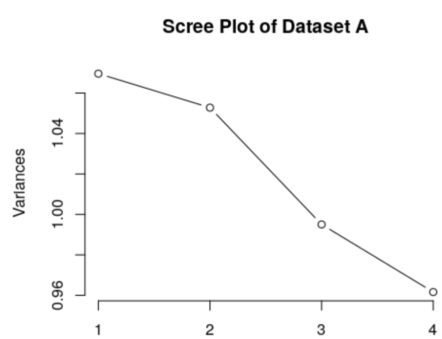
(a) Recall that XV contains the principal components of dataset X. and that we can alternatively calculate it given that $XV = U\Sigma$. Show that $XV = U\Sigma$.

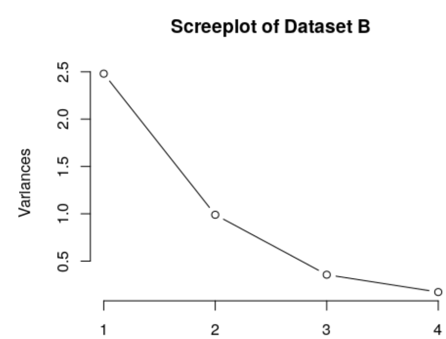
(b) Compute the first principal component (round to 2 decimals).

(c) Given the results of (a), how can we interpret the rows of V^T? What do the values in these rows represent?

(d) Show that the principal component vectors are orthogonal. In other words, for all vectors v_i, v_j that are principal components of dataset X for $i \neq j$, it is true that $v_i^T v_j = 0$.
To show this easily, we can demonstrate that given the principal component matrix P, $P^T P$ is a diagonal matrix. Show that $P^T P$ is diagonal and justify why this proves that the principal component vectors are orthogonal.
Hint: The fact that $(AB)^T = B^T A^T$ might help!

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. This means you can interpret the vertical axis as proportional to the fraction of variance captured by each principal component.





Application of PCA

3. Anirudhan wants to apply PCA to a dataset of rare rabbits to understand patterns in rabbit population per location as a function of the year. Provided is a Pandas DataFrame, `rabbit_pop` (shown below), which contains the rabbit population for every particular year and location. Note that not every year and location is shown here.

	2017	2018	2019	2020
Site A	8789	29372	49271	101822
Site B	18573	38317	102847	192742
Site C	402	3928	20212	80272
Site D	4392	28172	93172	203082

He needs to preprocess his current dataset in order to use PCA.

(a) Select all appropriate preprocessing steps used for PCA.

☐ A. Transform each row to have a magnitude of 1 (Normalization)

☐ B. Transform each column to have a mean of 0 (Centering)

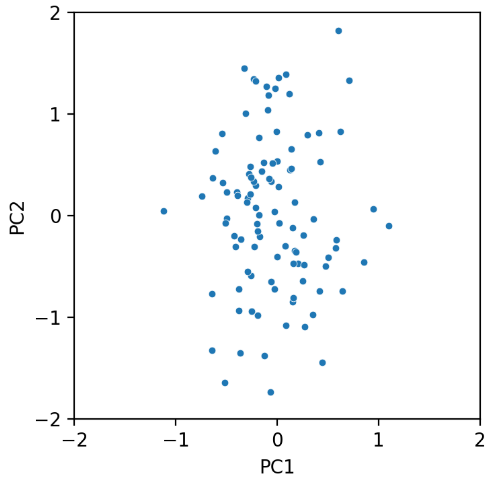
☐ C. Transform each column to have a mean of 0 and a standard deviation of 1 (Standardization)

☐ D. None of the above

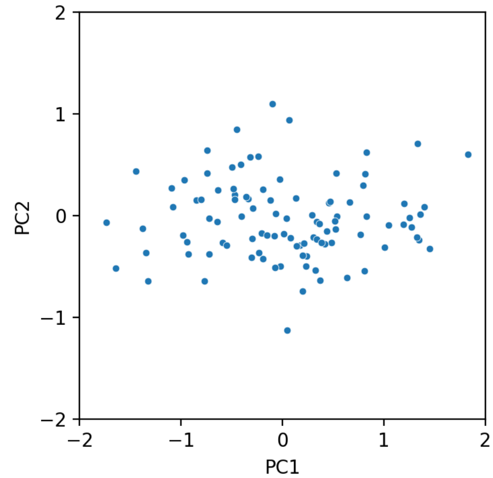
(b) Assume you have correctly preprocessed your data using the correct response in part (a). Write a line of code that returns the first 3 principal components assuming you have the correctly preprocessed DataFrame `rabbit_PCA` and the following variables returned by SVD.

```
u, s, vt = np.linalg.svd(rabbit_PCA, full_matrices = False)
first_3_pcs = ________________________________
```

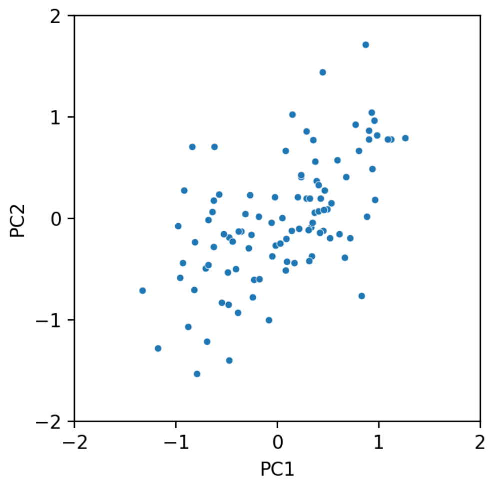
(c) We now wish to display the first two principal components in a scatterplot. Which of the following plots could potentially display the first two principal components given that the first principal component captures 60% of the variance and the second principal component captures 15% of the variance?



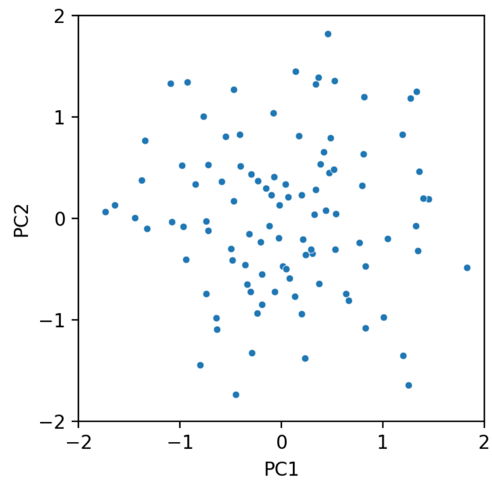
A.



B.



C.



D.