Data 100, Spring 2024

Discussion #13

Note: Your TA will probably not cover all the problems. This is fine; the discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, labs, and homework.

Terminology: The notation used for PCA this semester differs from previous semesters a bit.

1. Principal Component: The columns of V. These vectors specify the principal coordinate system and represent the directions along which the most variance in the data is captured.
2. Latent Vector Representation of X: The projection of our data matrix X onto the principal components, $Z = XV = US$ (as denoted in lecture). In previous semesters, the terminology was different and this was termed the principal components of X. In other classes, the term principal coordinate is also used. The i-th latent vector corresponds to the i-th column of V.
3. S (as in SVD): The diagonal matrix containing all the singular values of X.
4. Σ : The covariance matrix of X. Assuming X is centered, $\Sigma = X^T X$. In previous semesters, the singular value decomposition of X was written out as $X = U\Sigma V^T$. Note the difference between Σ in that context compared to this semester.

PCA Basics

1. Consider the following dataset, where X is the corresponding 4×3 design matrix. The mean and variance for each of the features are also provided.

Observations	Feature 1	Feature 2	Feature 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	47.56	21.35	0.51

Suppose we perform a singular value decomposition (SVD) on this data to obtain $X = USV^T$: (Note: U and V^T are not perfectly orthonormal due to rounding to 2 decimal places.)

$$U = \begin{bmatrix} -0.25 & 0.81 & 0.20 \\ -0.61 & -0.56 & 0.24 \\ 0.13 & -0.06 & -0.85 \\ 0.74 & -0.18 & 0.41 \end{bmatrix}, S = \begin{bmatrix} 13.79 & 0 & 0 \\ 0 & 9.32 & 0 \\ 0 & 0 & 0.81 \end{bmatrix}, V^T = \begin{bmatrix} 1.00 & 0.02 & 0.00 \\ -0.02 & 0.99 & -0.13 \\ 0.00 & 0.13 & 0.99 \end{bmatrix}$$

(a) Recall that we define the columns of V as the principal components. By projecting X onto the principal components (i.e. computing XV), we can construct the latent vector representation of X. Alternatively, you can also calculate the latent vector representation using US. Prove, using the definition from lecture, that $XV = US$.

(b) Compute the projection of X onto the first principal component (round to 2 decimal places). You can also view this as the best rank-1 approximation of X.

(c) What is the component score of the first principal component? In other words, how much variance does it capture of the original data X?

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

Applications of PCA

2. Lillian wants to apply PCA to `food_PCA`, a dataset of food nutrition information to understand the different food groups.

id	energy_kcal	protein_g	fat_g	carb_g	sugar_g	fiber_g	vita_mcg
1001	717.0	0.85	81.11	0.06	0.06	0.0	684.0
1002	717.0	0.85	81.11	0.06	0.06	0.0	684.0
1003	876.0	0.28	99.48	0.00	0.00	0.0	840.0
1004	353.0	21.40	28.74	2.34	0.50	0.0	198.0
1005	371.0	23.24	29.68	2.79	0.51	0.0	292.0

She needs to preprocess her current dataset in order to use PCA.

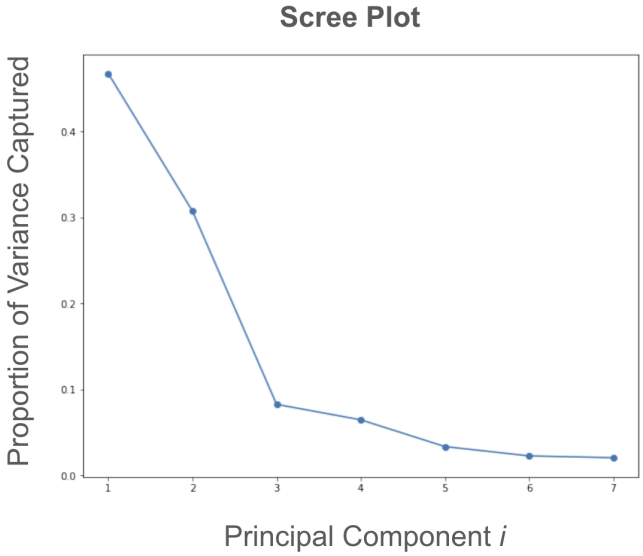
(a) What are the appropriate preprocessing steps when performing PCA on a dataset?

- ☐ 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 latent vectors assuming you have the correctly preprocessed `food_PCA` and the following variables returned by SVD.

```
u, s, vt = np.linalg.svd(food_PCA, full_matrices = False)
first_3_latent_vecs = _____
```

(c) The scree plot below depicts the proportion of variance captured by each principal component.



Which of the following lines of code could have created the plot above?

◯ A. `plt.plot(s**2/np.sum(s**2), u)`

◯ B. `plt.plot(food_PCA[:, :7]), s**2/np.sum(s))`

◯ C. `plt.plot(np.arange(1, food_PCA.shape[1]+1), s**2/np.sum(s**2))`

◯ D. `plt.plot(np.arange(1, food_PCA.shape[1]+1), s**2/np.sum(s))`

◯ E. `plt.plot(u@s, s**2/np.sum(s**2))`

(d) Using the elbow method, how many principal components should we choose to represent the data?

Interpreting PCA Plots

3. Yash has three datasets A, B, and $C \in \mathbb{R}^{100\times 2}$. That is, each dataset consists of 100 data points in two dimensions. He visualizes the datasets using scatterplots, labeled Plot A, Plot B, and Plot C, respectively:

(a) If he applies PCA to each of the above datasets and uses only the first principal component, which dataset(s) would have the lowest reconstruction error? Select all that apply.

☐ A. Dataset A
☐ B. Dataset B
☐ C. Dataset C
☐ D. Cannot determine with the given information

(b) If he applies PCA to each of the above datasets and uses the first two principal components, which dataset(s) would have the lowest reconstruction error? Select all that apply.

☐ A. Dataset A
☐ B. Dataset B
☐ C. Dataset C
☐ D. Cannot determine with the given information

(c) Suppose he decides to take the Singular Value Decomposition (SVD) of one of the three datasets, which we will call Dataset X. He runs the following piece of code:

```
X_bar = X - np.mean(X, axis=0)
U, S, V_T = np.linalg.svd(X_bar)
```

He gets the following output for `S`:

```
array([15.59204498, 3.85871854])
```

and the following output for `V_T`:

```
array([[0.89238775, -0.45126944], [0.45126944, 0.89238775]])
```

Based on the given plots and the SVD, which of the following datasets does Dataset X most closely resemble? Select one option.

◯ A. Dataset A
◯ B. Dataset B
◯ C. Dataset C