Data Science and Social Inquiry: HW2

Yu-Chang Chen

Due: 10/20

Question 1: Matrix operations

This question helps you get used to matrix algebra.

Consider $\boldsymbol{X}$, a 2×1 random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

be a matrix of real numbers,

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

be a vector of real numbers, and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$ be another random vector. Note that $\boldsymbol{Y}$ is a linear transformation of $\boldsymbol{X}$. Show the following.

(a) (1pt) $E[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$.

(b) (2pt) $Cov(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$.

Question 2: PCA with non-diagonal covariance matrix

In class, we went through PCA for the case when the covariance matrix is a diagonal matrix. But, of course, not every covariance matrix is diagonal. For general cases, we can apply the *Real Spectral Theorem* to diagonalize the covariance matrix. Although I have already briefly outlined how to apply the theorem for PCA in class, having hands-on experience of the process will help you learn, which is the purpose of this question.

By the Real Spectral Theorem, we can always "diagonalize" any covariance matrix since it is symmetric. [1] For example, suppose that our data set contains 5 variables, i.e.,

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix},$$

and the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'\right] = \boldsymbol{\Sigma} = \begin{pmatrix} 0.838 & 0.049 & 0.138 & -0.04 & -0.067 \\ 0.049 & 0.838 & 0.178 & -0.309 & 0.136 \\ 0.138 & 0.178 & 0.264 & 0.172 & 0.117 \\ -0.04 & -0.309 & 0.172 & 1.557 & -0.534 \\ -0.067 & 0.136 & 0.117 & -0.534 & 1.17 \end{pmatrix},$$

which can be decomposed as $\boldsymbol{\Sigma} = \boldsymbol{PDP^{-1}}$, where

$$\boldsymbol{P} = \begin{pmatrix} -0.135 & -0.763 & 0.606 & -0.182 & 0.006 \\ -0.27 & -0.477 & -0.745 & -0.271 & 0.268 \\ -0.405 & -0.191 & -0.062 & 0.892 & -0.015 \\ -0.539 & 0.095 & -0.065 & -0.242 & -0.798 \\ -0.674 & 0.381 & 0.266 & -0.197 & 0.539 \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} 0.855 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.942 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.738 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.109 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 2.024 \end{pmatrix},$$

and

$$\boldsymbol{P}^{-1} = \begin{pmatrix} -0.135 & -0.27 & -0.405 & -0.539 & -0.674 \\ -0.763 & -0.477 & -0.191 & 0.095 & 0.381 \\ 0.606 & -0.745 & -0.062 & -0.065 & 0.266 \\ -0.182 & -0.271 & 0.892 & -0.242 & -0.197 \\ 0.006 & 0.268 & -0.015 & -0.798 & 0.539 \end{pmatrix}.$$

[1] You do not need to know how to prove Real Spectral Theorem and how to diagonalize a matrix for this class. These topics are covered in linear algebra courses.

Answer the following questions.[2]

(*a*) (2 pts) What is the first principal component? Explain how you reach your answer carefully and write down its coefficients.

(*b*) (1 pt) Calculate the proportion of variance explained by each component and draw the scree plot.

Question 3: Image Compression with PCA

As discussed in class, one of the primary motivations behind dimension reduction is data compression. Moreover, image compression serves as an excellent example to elucidate the concept of PCA.

Download the Jupyter notebook titled `faces_question.ipynb` and the dataset `faces.csv` from NTU Cool. The initial two blocks in the notebook will guide you through importing the necessary packages and visualizing 16 randomly selected images from the dataset.

(*a*) (2pt) Find the principal components, referred to as the "principal faces". Plot both the mean face and the first three principal faces.

(*b*) (2pt) Focus on compressing the last face present in our dataset. Achieve this compression by computing inner products with the principal faces. Subsequently, plot the reconstructed faces using $q = 5, 20, 100, 200$ principal components.

(*c*) (1pt) When employing $q = 200$ principal components for data compression, what percentage of storage space do we save compared to the size of the original dataset?

Question 4: Prototyping CEO's behavior

Navigate to the website of the *Journal of Political Economy* and download the dataset provided by Bandiera et al., 2020.[3] The dataset in focus is `survey_response_data.csv`, containing records of activities undertaken by each CEO.

The objective is to investigate the nature of the outsiders that CEOs meet. The investigation will concentrate on seven types of outsiders: **clients, suppliers, banks, investors, lawyers, politicians, and government officials**.

Start by aggregating the activity data to count how many activities each CEO has conducted with each of the seven types of outsiders. The subsequent steps will be based on this aggregated data.

[2]Feel free to verify that $\boldsymbol{PP'} = \boldsymbol{I}$ (so $\boldsymbol{P}^{-1} = \boldsymbol{P}$) and $\boldsymbol{\Sigma} = \boldsymbol{PDP^{-1}}$.

[3]If you encounter any difficulty, refer to this link: https://www.journals.uchicago.edu/doi/full/10.1086/705331?af=R&mobileUi=0. Utilize NTU VPN for data access.

id	clients	suppliers	banks	investors	lawyers	politicians	govoff
2	0	12	0	0	0	0	0
3	0	8	0	0	0	0	0

Table 1: Example of Aggregated Data

(*a*) (1 pt) Initiate every data analysis by scrutinizing the raw data. Utilize a box plot to summarize the seven marginal distributions.

(*b*) (1 pt) Subsequent to the box plot, construct a pair plot for these seven variables. What observations can be made from the box plot and pair plot?

(*c*) (1 pt) Employ a heatmap to represent the correlations between the number of activities. Identify which type exhibits the highest correlation with **politicians**.

(*d*) (1 pt) Standardize the seven variables by centering around their mean and scaling by their standard deviations. Execute PCA to determine the first principal component.[4]

(*e*) (1 pt) Plot the scree plot. How many principal components are required to explain 70% of the variation.

(*f*) (1 pt) Position the first component on the x-axis and the second component on the y-axis, then plot the coefficients of each variable. Provide an interpretation for the first two principal components.

Question 5: Practicing the hierarchical clustering algorithm

Suppose that our data has 5 observations:

$$(X_i, Y_i) = (0, 4),\ (-3, 1),\ (3, 3),\ (3, 5),\ (-3, 3)$$

(*a*) (1 pt) Perform the hierarchical clustering with average linkage. Clearly indicate which observations are pooled in each step.

(*b*) (1 pt) Perform the hierarchical clustering with the complete linkage.

P.S. They are not programming questions. You should not use any software for this question.

References

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, *128*(4), 1325–1369.

[4]Implementation of PCA from scratch is not mandatory; Python modules are available.