Data Science and Social Inquiry: HW2

Yu-Chang Chen and Ming-Jen Lin

September 25, 2022

Question 1: 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 HW1.

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

[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.

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

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

and

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

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) Find the corresponding variance of each principal component. Graph the scree plot for the five components. [3]

Question 2: Prototyping CEO's behavior

Go to the website of *Journal of Political Economy* and download the data set provided by Bandiera et al., 2020.[4] We will use the data set `survey_response_data.csv`, which is a data set of activities done by each CEO.

We are interested in investigating who are the outsiders CEOs have meetings with. Specifically, we will focus on the following seven types of outsiders: **clients, suppliers, banks, investors, lawyers, politicians, and government officials**.

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

[3] Our template will show you how to use the package *tikz* to make professional plots in latex.

[4] In case you can't find it, here is the link: https://www.journals.uchicago.edu/doi/full/10.1086/705331?af=R&mobileUi=0.

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: Demo of aggregated data

First, aggregate the activity data and count many activities each CEO has with each of the seven type. We will work with the aggregated data in this question.

(*c*) (1 pt) Every data analysis should start with examining the raw data. Use a box plot to summarize the seven marginal distributions.

(*d*) (1 pt) Use a heatmap to summarize the correlations between the number of activities. Which type correlates with type **politicians** most?

(*e*) (1 pt) Run PCA. What is the first principal component?[5]

(*f*) (1 pt) Make the scree plot. How many principal components are needed to explain 70% of the variation?

(*g*) (1 pt) Put the first component on the x-axis and the second component on the y-axis. Plot the coefficients of each variable. How would you interpret the first two components?

Question 3: Practicing the hierarchical clustering algorithm

Suppose that our data has 5 observations:

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

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

(*i*) (1 pt) Perform the hierarchical clustering with single linkage.

P.S. Questions (*h*) and (*i*) 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.

[5] You do not have to implement PCA by yourself. There are Python modules available.